Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Sample size of surgical randomized controlled trials: a lack of improvement over time

Usama Ahmed Ali, MD, MSc,[a,*,1] *Joren R. ten Hove, MD,*[a,1] *Beata M. Reiber, MD,*[b] *Pieter C. van der Sluis, MD,*[a] *and Marc G. Besselink, MD, MSc, PhD*[b]

[a] *Department of Surgery, University Medical Center Utrecht, Utrecht, The Netherlands*
[b] *Department of Surgery, Academic Medical Center, Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 20 April 2017
Received in revised form
6 November 2017
Accepted 13 February 2018
Available online 21 March 2018

Keywords:
Trials
Power
Sample size

ABSTRACT

Background: Interpretation of randomized controlled trials (RCTs) without a significant difference regarding the primary outcome (negative RCTs) is frequently challenging, due to concerns about sample size and thus sufficient statistical power. We aimed to assess the adequacy of sample size and corresponding power of surgical RCTs.

Methods: We previously identified all surgical RCTs available in PubMed in two distinct years a decade apart (1999 and 2009). For all "negative" trials, we estimated whether the sample size of the trial was appropriate to detect a difference in the primary outcome measure. The main outcome measure was a sufficient sample size to detect large, medium, and small treatment effects. We also performed a post hoc power analysis based on the actual observed effect difference.

Results: A total of 228 negative RCTs (74 in 1999 and 121 in 2009) were included. The median sample size was 76 (± 222) and 80 (± 163) in 1999 and 2009, respectively. Sample size calculation was increasingly reported from 40% in 1999 to 54% in 2009 ($P = 0.02$). The proportion of studies adequately powered to detect large (57% *versus* 68%), medium (26% *versus* 25%), or small (8% *versus* 7%) differences did not differ significantly between 1999 and 2009, respectively. To reach sufficient power, the required increases in sample size were 130%, 240%, and 1032% for large, medium, and small differences, respectively. Reporting a sample size calculation was the only independent predictor for adequate power.

Conclusions: Despite slight improvement in the reporting of a sample size calculation, about a third of surgical trials remains underpowered to demonstrate differences that are likely to be clinically significant. Increased attention of researchers, medical ethical boards, and journal editors is required to reduce potentially wasted resources on underpowered trials.

Source of support: no funding was obtained for this work.

* *Corresponding author*. Department of Surgery, Room G4-228, Universtiy Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands. Tel.: +31 624448308; fax: +31 302541944.
E-mail address: u.ahmedali@gmail.com (U. Ahmed Ali).

[1] Both authors contributed equally to this study.

0022-4804/$ – see front matter © 2018 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.jss.2018.02.014

Introduction

Randomized controlled trials (RCTs) are essential to improve clinical practice. However, only well-designed trials offer reliable results suitable for decision-making. In the case of an RCT that does not show a statistically significant difference between the treatment arms, one can only conclude that there is no difference between treatments if the study is sufficiently powered.[1] Underpowered RCTs are, therefore, not particularly helpful and in certain cases potentially harmful because the use of resources and added risk for patients are not outweighed by the usefulness of the study results.

Previous studies have shown that RCTs in several fields are frequently underpowered. Dimick *et al.*[2] analyzed 90 trials from three surgical journals between 1988 and 1989 and found that only 22 (24%) trials had a power greater than 80% to detect a 50% difference in therapeutic effect. Maggard *et al.*[3] analyzed 127 RCTs in surgical literature, and only half of these studies were appropriately powered to detect a 50% effect change. Similarly, Lochner *et al.*[4] analyzed 117 RCTs in the orthopedic trauma literature and concluded that the type-II error rate for primary outcomes was 91%. For nonsurgical specialties, this problem is also widely prevalent.[5-7]

An evaluation of the current situation of statistical power in surgical RCTs is lacking, with the most recent reviews published over a decade ago.[2,3] This study aimed to 1) assess the adequacy of the obtained sample sizes in negative surgical RCTs, and 2) to identify whether the proportion of adequately sized studies has changed over the last decade, and which factors were associated with adequate power.

Methods

Search strategy

We used a search strategy aimed at identifying all surgical RCTs published in PubMed in two distinct years (1999 and 2009), as reported previously.[8] We searched PubMed using the MeSH term "surgery" and various permutations combined with the Cochrane Highly Sensitive Search Strategy. Subsequently, we selected all retrieved hits according to relevance by two independent reviewers. The inclusion criteria were as follows: (1) an RCT (defined as any prospective study assessing the effect of health-care interventions in humans randomly allocated to study groups), (2) surgical trials, defined as any trial performed by a corresponding author from a general surgical department or examining a general surgical procedure. The exclusion criteria were as follows: (1) non-RCTs and (2) publications in other languages than English, French, German, or Dutch. For all included RCTs, we extracted geographical (i.e., region and number of countries), publishing (i.e., number of participating authors and centers and impact factor), clinical (i.e., specialty and type of intervention), and epidemiological characteristics (i.e., number of randomized patients and methodological quality). "Low risk of bias" trials were defined as trials that adequately reported all of the following four items: adequate generation of allocation, adequate concealment of allocation, intention-to-treat analysis, and handling of dropouts.[8]

Data extraction

Two-arms, parallel-group trials without a significant difference regarding the primary outcome were selected for further analysis. The following additional data were extracted:

- Calculation of sample size: presence and methods of sample size calculation.
- Outcome type: dichotomous or continuous.
- Trial objective: superiority, noninferiority, or equivalence.
- Hypothesized direction of treatment effect on outcome: increase outcome (e.g., intervention is supposed to increase cure) or decrease outcome (e.g., intervention is supposed to reduce harm).
- Notion or discussion of limitation of sample size or lack of power by authors.
- Summary statistics (mean and standard deviation [SD]) for the primary and (maximum of three) secondary outcomes. If not present, we estimated the mean and SD from other summary statistics as described in the Cochrane Handbook for Systematic Reviews (Section 16.1.3) and by Hozo *et al.*[9,10] All formulas used in these steps are presented in Appendix I.

Extraction of these data was conducted by two independent reviewers for 30 studies. The inter-reviewer agreement kappa was then tested (kappa 0.92). This was followed by a review round in which discrepancies between the two reviewers were discussed, and consensus on how to proceed was reached. Finally, a final verification round of yet another 30 studies was conducted. With satisfactory agreement (kappa 1.0), remaining studies were extracted by one reviewer each.

Study outcomes

Our primary endpoint was the presence of a sufficient sample size to detect large, medium, and small treatment effects. We defined these for continuous outcomes as a multiple of the SD as follows: large (0.8 SD), medium (0.5 SD), and small (0.2 SD). For dichotomous outcomes, these treatment effects were calculated as a relative change from the control group as follows: large (40% change), medium (20% change), and small (10% change). The primary outcome was chosen as follows: 1) the endpoint used for the sample size calculation; 2) if not present, a clearly stated primary outcome and 3) if not present, the most clinically relevant outcome.

Based on the actual observed estimate in the control arm and the hypothesized direction of the intervention (an increase or decrease in outcome), the appropriate difference was either added or subtracted. The result of this calculation was the hypothesized treatment effect in the intervention arm. For dichotomous outcomes, calculated values could never be lower than 0% or higher than 100%. Functions used for calculation are presented in Appendix I.[5,11] For all calculation, we assumed an alpha (α) value (i.e., risk of type I error)

of 0.05, beta (β) value (i.e., risk of type II error) of 0.2, and a two-arms parallel trial design with equal group sizes.

Our secondary endpoints were as follows:

- The actual power based on the number of randomized patients and observed effect difference.
- The number of patients needed to increase study power to an adequate level.
- Study characteristics associated with an improved study power.
- The presence of a sufficient sample size to detect large, medium, and small treatment effects regarding secondary outcomes. For this, we chose the first three encountered outcomes in the abstract.
- The number of trials discussing power as a potential limitation of the study.

Statistical analysis

Analyses were conducted for both study years (1999 and 2009) separately. Results from both years were compared. In addition, subgroups analysis for studies with dichotomous *versus* continuous endpoints were performed. Dichotomous outcomes are presented as proportion and were compared using Chi-square test. Continuous data are presented as means with SD or median and interquartile range and were compared by the Student's *t*-test or Mann–Whitney *U*-test according to normality. For all dichotomous outcomes, relative ratio with corresponding 95% confidence intervals (95% CI) are presented. In addition, a regression analysis was performed to identify study characteristics potentially associated with an improved study power. For this, a univariate analysis was performed on all baseline characteristics, and factors with potential association (*P*-value <0.2) were subsequently entered in a multivariate analysis (Backwards regression). A *P*-value of <0.05 was used as threshold for statistical significance. Analyses were performed using SPSS software, version 21 (Armonk, NY: IBM Corp).

Results

Search results

A previous search retrieved 300 and 450 surgical RCTS out of 12,780 and 25,711 PubMed hits for 1999 and 2009, respectively.[8] Of these, 241 (33%) RCTs were potentially eligible trials fulfilling our inclusion criteria (two-arm negative parallel trials). Eventually, 13 (5% of all eligible trials) RCTs did not provide sufficient data for calculation and were excluded, resulting in 228 RCTs for final analyses (Figure).

Study characteristics

Baseline characteristics of all included trials are presented in Table 1. Less than half (49%) of included studies reported a sample size calculation. The number of included studies reporting a sample size increased over the 10-year study period (40% in 1999 *versus* 54% in 2009, $P = 0.02$). The total amount of randomized patients per trial did not increase over time (151 *versus* 136, respectively, $P = 0.8$), while collaboration between countries was more frequent in 2009. Noticeably, the quality of RCTs did improve over time, with an increase in the proportion of the "low risk of bias" RCTs from 8% to 21% ($P < 0.01$), respectively.

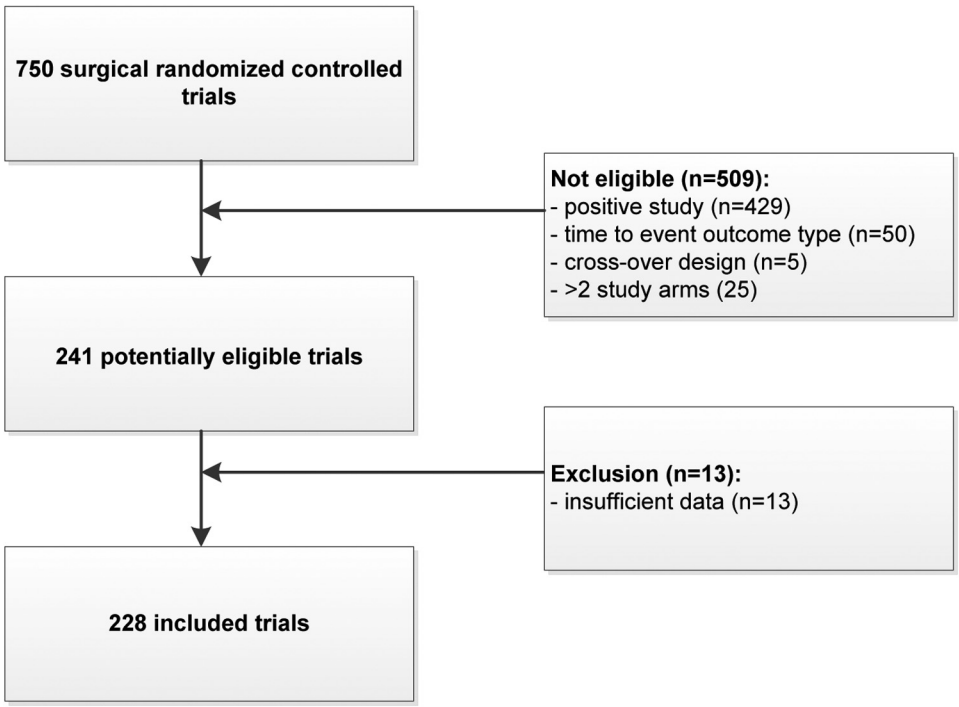

Fig – Flow diagram of the selection process.

Table 1 – Baseline characteristics of included trials.

Characteristics	1999 ($n = 90$) (%)	2009 ($n = 138$) (%)	P-value
Speciality			
General surgery	16 (18)	37 (27)	0.04
Gastrointestinal surgery	43 (48)	69 (50)	
Vascular surgery	16 (18)	12 (9)	
Other subspecialties	15 (17)	20 (14)	
Number of authors			
Median	5	6	0.07
Number of centers			
Single center	52 (58)	80 (58)	0.36
Two centers	18 (20)	19 (14)	
Multiple centers	20 (22)	39 (28)	
Impact factor (IF)*			
Median	2.2	2.4	0.32
Number of countries			
Single country (*versus* International)	89 (99)	121 (88)	<0.01
Region			
Europe	50 (56)	62 (45)	<0.01
North America	27 (30)	18 (13)	
Asia/Oceania	13 (14)	52 (38)	
Africa/South America	0 (0)	6 (4)	
Funding			
Industry funded	20 (22)	27 (20)	<0.01
Nonindustry/no external funding	15 (17)	52 (38)	
Not reported	55 (61)	59 (43)	
Type of intervention studied			
Surgical procedure	38 (42)	67 (49)	0.35
Other type of intervention	52 (58)	71 (51)	
Number of randomized patients—continuous			
Mean	151	136	0.80
Median (range)	76 (19-1560)	80 (17-1032)	
Number of randomized patients—categorized			
<50	26 (29)	32 (23)	0.49
50-100	26 (29)	49 (36)	
>100	38 (42)	57 (41)	
Sample size calculation performed			
Performed (*versus* not performed)	36 (40)	75 (54)	0.02
Methodological quality			
Low risk of bias	7 (8)	29 (21)	<0.01

*For studies published in 1999, the 1999 impact factor was used; for studies published in 2009, the 2009 impact factor was used.

Power of surgical trials

One hundred and ninety-five (86%) trials provided sufficient data on the primary outcome for analysis. For the remaining trials, only data on secondary outcomes were available. Adequacy of the sample size to provide sufficient power to identify large, medium, or small effect differences in the primary outcome is shown in Table 2. About two-thirds of trials (68%) in 2009 had a sufficient sample size to detect a large difference (0.8 of the SD or 40% relative change for continuous and dichotomous outcomes, respectively) with >80% power. This was slightly more than that in 1999 (57%). These proportions decreased substantially for medium (25% in 2009) and small (7% in 2009) differences. To reach sufficient power, the required increases in sample size were 130%, 240%, and 1032% for large, medium, and small differences, respectively. Mean study power and the required increments needed to reach an adequate sample size were not significantly different between 1999 and 2009.

In a subgroup analysis of continuous and dichotomous outcomes, results were consistent with the overall analysis (Table 3). Only a slightly larger proportion of trials with a

Table 2 – Primary outcome: adequacy of sample size by year of publication.

	1999 ($n = 74$)			2009 ($n = 121$)		
	Large effect	Medium effect	Small effect	Large effect	Medium effect	Small effect
Power (%)						
>80% (adequate)	42 (56.8)	19 (25.7)	6 (8.1)	82 (67.8)	30 (24.8)	8 (6.6)
51-80%	17 (23.0)	17 (23.0)	4 (5.4)	23 (19.0)	37 (30.6)	8 (6.6)
<50%	15 (20.3)	38 (51.4)	64 (86.5)	16 (13.2)	54 (44.6)	105 (86.8)
Required number of patients (as percentage of actual sample size)						
Mean (SD)	157 (239)	274 (266)	1080 (1002)	113 (139)	218 (179)	1002 (823)
Required increase in sample size (mean)	288.3	330.4	635.1	176.1	220.4	598.2

continuous outcome had an adequate sample size compared with those with dichotomous outcomes (71% *versus* 64%).

The regression analysis for factors associated with adequate sample size revealed the presence of an a priori power calculation as the only associated factor with adequate sample size (odds ratio [OR] 2.34, 95% CI 1.24-4.44, $P < 0.01$). In univariate analysis, the number of participating departments (OR 1.14, 95% CI 0.99-1.31, $P = 0.058$) and low risk of bias studies (OR 2.31, 95% CI 0.94-5.65, $P = 0.067$) showed a nonsignificant trend, which disappeared in the multivariable analysis. An additional analysis examining the effect of performing a sample size calculation showed that adequacy of the included sample size was significantly higher for all chosen effect sizes (large 85% *versus* 73%, $P = 0.001$; medium 65% *versus* 49%, $P = 0.000$; and small 31% *versus* 22%, $P = 0.008$).

When examining whether the sample size of the included RCTs was adequate in regard to secondary outcomes, consistent results were seen in comparison with the primary outcomes (Table 4). There were no significant differences between the proportion of studies with an adequate sample size for primary (57%) and secondary (60%) outcomes. Out of 228 included reports, 160 (70.2%) did not mention the sample size as a possible limitation to the study. For the remaining studies, 50 (21.9%) mentioned the sample size as a possible limitation, whereas 7 (3.1%) performed additional calculations pertaining to the study power.

Discussion

This study assessed the adequacy of the recruited sample size of published "negative" surgical RCTs. In both study years, about two-thirds of trials had a sufficient sample size to detect large differences. However, the proportion of trials able to detect medium effect differences with sufficient power, a more realistic reflection of real practice, did not exceed 25% in both years. Also, the number of randomized patients per trial did not increase between the 2 y (median of 70-80 patients). There was, however, an increase in the reporting of a sample size calculation which was the only study factor associated with adequate power. Other factors such as year of publication, specialty, methodological quality, and impact factor of journals were not associated with an adequate study power.

There is no doubt that negative RCTs are quite valuable for clinical practice because they can prevent the use of redundant or even harmful therapies and have the potential to reduce health costs. Yet, underpowered clinical trials bring certain risks, especially if not recognized as such.[12] First, funding and medical resources are spent on a study that does not produce valuable and reliable results. Second, it can be considered unethical because patients are needlessly exposed to risks relating to treatments and investigations. Moreover, when authors and clinicians draw incorrect conclusions from

Table 3 – Primary outcome: adequacy of sample size stratified per outcome type (year 2009).

	Continuous ($n = 66$)			Dichotomous ($n = 55$)		
	Large effect	Medium effect	Small effect	Large effect	Medium effect	Small effect
Power (%)						
>80% (adequate)	47 (71.2)	12 (18.2)	0 (0.0)	35 (63.6)	18 (32.7)	8 (14.5)
51-80%	16 (24.2)	22 (33.3)	1 (1.5)	7 (12.7)	15 (27.3)	7 (12.7)
<50%	3 (4.5)	32 (48.5)	65 (98.5)	13 (23.6)	22 (40.0)	40 (72.7)
Required number of patients (as percentage of actual sample size)						
Mean (SD)	82 (47)	210 (121)	1312 (753)	150 (193)	228 (231)	1003 (823)
Required increase in sample size (mean)	13.6	89.6	748.6	330.6	344.7	455.4

Table 4 – Adequacy of sample size for secondary outcomes by year of publication.

	1999 ($n = 52$)			2009 ($n = 89$)		
	Large effect	Medium effect	Small effect	Large effect	Medium effect	Small effect
Power (%)						
>80% (adequate)	31 (59.6)	13 (25.0)	3 (5.8)	53 (59.6)	21 (23.6)	4 (4.5)
51-80%	14 (26.9)	14 (26.9)	4 (7.7)	22 (24.7)	30 (33.7)	4 (4.5)
<50%	7 (13.5)	25 (48.1)	45 (86.5)	14 (15.7)	38 (42.7)	81 (91.0)
Required number of patients (as percentage of actual sample size)						
Mean (SD)	136 (169)	245 (192)	1173 (955)	123 (134)	233 (187)	1044 (854)
Required increase in sample size (mean)	75.8	122.8	560.6	127.6	169.8	523.9

underpowered negative trials, this may result in an incorrect body of evidence and ultimately suboptimal patient care.[12]

Previous studies that analyzed the adequacy of sample size in published RCTs have shown worrying results as well. Bedard *et al.*[5] performed a large study including 423 negative oncologic trials, 45 (10.6%), 138 (32.6%), and 233 (55.1%) had adequate sample size to detect small, medium, and large effect sizes, respectively. Similar studies have shown strikingly low power for trials in general medicine,[13] critical care,[14] urology,[15] plastic surgery[16], and orthopedics.[4,17]

What sets this study apart is the fact that reports were retrieved from all available surgical journals, whereas most of the previous studies on this subject were only selected from a limited number of (high impact factor) journals. Moreover, we did not select studies according to outcome type (dichotomous or continuous). In doing this, we have attempted to provide a more complete overview of the body of evidence found in surgical literature. Our study was also able to compare two different time periods, in between which the revised Consolidated Standards of Reporting Trials (CONSORT) statement has been broadly implemented.[18] This allowed us to show that there was no clear improvement over time in this important aspect of study conductance. Yet, when looking specifically at the detection of large differences, reaching adequate sample sizes does not seem out of reach. Most studies required an increase that was less than two times the original number of patients.

In addition, we performed a regression analysis, identifying the performance of a sample size calculation as an important independent predictor of adequate power. This is an important finding because this is a modifiable factor that can be easily acted on by researcher and can be enforced by both medical ethical boards examining RCTs proposal as well as editor of journal publishing either protocol or final reports of RCTs. Our review also showed that there is ample room for improvement because less than half of published RCTs reported such a calculation.

A number of studies have shown that a variety of medical specialties are insufficient in reporting sample size calculations.[19-23] Moher *et al.*[13] were one of the first to report that less than half of the studies published in high-impact journals reported an a priori sample size calculation. This seems to indicate that this problem is quite widespread. The lack of an association between impact factor and publication in a top 10 journal and sufficient sample size in our study further confirms this conclusion.

Finally, underpowered trials can be useful when meta-analysis combining several of such trials is conducted. However, the power limitation should be discussed clearly (preferable in the abstract) by the authors of such RCTs to prevent drawing of incorrect conclusions. This practice is currently infrequent, with 21.9% of RCTs including such discussion in their report. In addition, results should be published in a form that allows for meta-analysis, that is, mean and SD, to facilitate this process and prevent reporting bias.

There are certain limitations to the methodology used in this study. First, there are downsides to the use of fixed prespecified effect sizes. There is no easy way around this limitation because the clinically relevant effect size is different for each intervention. By using three different prespecified effect sizes (small, medium, and large), we aimed to provide an overview of the full spectrum of effect size to maximally inform the reader. Second, the search has been conducted about 6 y ago. The delay is in great part because of the magnitude of the study.

Conclusion

In conclusion, the results of this study demonstrate that inadequate study power among negative surgical RCTs is common and persist over time. Reporting an adequate sample size calculation is the only predictor for improved power. The overwhelming majority of inadequately powered studies do not discuss this as a limitation in their publication. The persistence of this important problem calls for improvement from investigators, medical ethical boards, and journal editors. Ways in which these can be achieved include consultation of an epidemiologist, adherence to the CONSORT statement, and awareness among peer-reviewers after reports have been submitted. The performance of a formal sample size calculation seems to be the best way to ensure a proper power when conducting a study.

Disclosure

The authors reported no proprietary or commercial interest in any product mentioned or concept discussed in this article.

REFERENCES

1. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
2. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? *Arch Surg*. 2001;136:796–800.
3. Maggard MA, O'Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? *Surgery*. 2003;134:275–279.
4. Lochner HV, Bhandari M, Tornetta P. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am*. 2001;83-A:1650–1655.
5. Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. *J Clin Oncol*. 2007;25:3482–3487.
6. Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. *J Rheumatol*. 2005;32:2083–2088.
7. Yuen SY, Pope JE. Learning from past mistakes: assessing trial quality, power and eligibility in non-renal systemic lupus erythematosus randomized controlled trials. *Rheumatology (Oxford)*. 2008;47:1367–1372.
8. Ahmed Ali U, van der Sluis PC, Issa Y, et al. Trends in worldwide volume and methodological quality of surgical randomized controlled trials. *Ann Surg*. 2013;258:199–207.
9. [updated March 2011]. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version* 5.1.0. The Cochrane Collaboration; 2011. Available at: http://training.cochrane.org/handbook. Accessed July 2014.
10. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol*. 2005;20:5–13.
11. Devore JL. *Probability and statistics for engineering and sciences*. 2nd edition. Belmont, CA: Brooks/Cole; 1987.
12. Brody BA, Ashton CM, Liu D, et al. Are surgical trials with negative results being interpreted correctly? *J Am Coll Surg*. 2013;216:158–166.
13. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994;272:122–124.
14. Harhay MO, Wagner J, Ratcliffe SJ, et al. Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med*. 2014;189:1469–1478.
15. Breau RH, Carnat TA, Gaboury I. Inadequate statistical power of negative clinical trials in urological literature. *J Urol*. 2006;176:263–266.
16. Chung KC, Kalliainen LK, Spilson SV, Walters MR, Kim HM. The prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. *Plast Reconstr Surg*. 2002;109:1–6. discussion 7-8.
17. 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.
18. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001;357:1191–1194.
19. Hebert RS, Wright SM, Dittus RS, Elasy TA. Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. *J Negat Results Biomed*. 2002;1:1.
20. Weaver CS, Leonardi-Bee J, Bath-Hextall FJ, Bath PMW. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke*. 2004;35:1216–1224.
21. Alam M, Rauf M, Ali S, Nodzenski M, Minkis K. A systematic review of reporting in randomized controlled trials in Dermatologic Surgery: Jadad scores, power analysis, and sample size determination. *Dermatol Surg*. 2014;40:1299–1305.
22. Ayeni O, Dickson L, Ignacy TA, Thoma A. A systematic review of power and sample size reporting in randomized controlled trials within plastic surgery. *Plast Reconstr Surg*. 2012;130:78e–86e.
23. Sexton SA, Ferguson N, Pearce C, Ricketts DM. The misuse of "no significant difference" in British orthopaedic literature. *Ann R Coll Surg Engl*. 2008;90:58–61.