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 xxx

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)