Statistical Power, Sample Size, and Their Reporting in Randomized Controlled Trials

David Moher, MSc; Corinne S. Dulberg, PhD, MPH; George A. Wells, PhD

Objective.—To describe the pattern over time in the level of statistical power and the reporting of sample size calculations in published randomized controlled trials (RCTs) with negative results.

Design.—Our study was a descriptive survey. Power to detect 25% and 50% relative differences was calculated for the subset of trials with negative results in which a simple two-group parallel design was used. Criteria were developed both to classify trial results as positive or negative and to identify the primary outcomes. Power calculations were based on results from the primary outcomes reported in the trials.

Population.—We reviewed all 383 RCTs published in *JAMA, Lancet,* and the *New England Journal of Medicine* in 1975, 1980, 1985, and 1990.

Results.—Twenty-seven percent of the 383 RCTs (n=102) were classified as having negative results. The number of published RCTs more than doubled from 1975 to 1990, with the proportion of trials with negative results remaining fairly stable. Of the simple two-group parallel design trials having negative results with dichotomous or continuous primary outcomes (n=70), only 16% and 36% had sufficient statistical power (80%) to detect a 25% or 50% relative difference, respectively. These percentages did not consistently increase over time. Overall, only 32% of the trials with negative results reported sample size calculations, but the percentage doing so has improved over time from 0% in 1975 to 43% in 1990. Only 20 of the 102 reports made any statement related to the clinical significance of the observed differences.

Conclusions.—Most trials with negative results did not have large enough sample sizes to detect a 25% or a 50% relative difference. This result has not changed over time. Few trials discussed whether the observed differences were clinically important. There are important reasons to change this practice. The reporting of statistical power and sample size also needs to be improved.

(*JAMA.* 1994;272:122-124)

THE EFFICACY of new interventions are most readily accepted if the results are from randomized controlled trials (RCTs).[1] Essential to the planning of an RCT is estimation of the required sample size. The investigator should ensure that there is sufficient power to detect, as statistically significant, a treatment effect of an a priori specified size. The opposite perspective of conceiving the problem is that the investigator should ensure that there is a low β value, or probability of making a type II error, or concluding that the result is not statistically significant, when the observed effect is clinically meaningful.

The relationship between negative findings (ie, when statistical significance was not reached) and statistical power has been well illustrated in Freiman and colleagues'[2] review of 71 RCTs with negative results published during 1960 to 1977. These RCTs were drawn from a collection of 300 simple two-group parallel design trials with dichotomous primary outcomes published in 20 journals. Freiman and colleagues were interested in assessing whether RCTs with negative results had sufficient statistical power to detect a 25% and a 50% relative difference between treatment interventions. Their review indicated that most of the trials had low power to detect these effects: only 7% (5/71) had at least 80% power to detect a 25% relative change between treatment groups and that 31% (22/71) had a 50% relative change, as statistically significant (α=.05, one tailed).

Since its publication, the report of Freiman and colleagues[2] has been cited more than 700 times, possibly indicating the seriousness with which investigators have taken the findings. Given this citation record, one might expect an increase over time in the awareness of the consequences of low power in published RCTs and, hence, an increase in the reporting of sample size calculations made before clinical trials are conducted.

Our objective was to describe the pattern, over time, in the level of power and in the reporting of sample size calculations in published RCTs with negative results since the publication of Freiman and colleagues'[2] report. We did this by extending Frieman and colleagues' objectives during the period from 1975 to 1990.

METHODS

Study Sample

We reviewed RCTs published in *JAMA, Lancet,* and the *New England Journal of Medicine.* These were the three of the 20 journals from which more than half (36/71) of the trials with negative results reported by Frieman and colleagues were drawn.[2] To capture the denominator, each volume published in 1975, 1980, 1985, and 1990 was handsearched to extract the RCTs. To be considered an RCT, the study being assessed had to contain an explicit statement about randomization. The identified RCTs were consecutively coded and divided into three equal groups for review by members of the study team, with use of a structured data collection form. The data collected included information on whether the trial results were positive or negative, the study design, whether a priori or post hoc sample size calculations were performed, and the elements necessary for us to calculate power (eg, observed proportions or means and SDs of the primary outcome obtained in each group).

From the Clinical Epidemiology Unit, Loeb Medical Research Institute (Mr Moher and Dr Wells), and the Faculties of Medicine (Mr Moher and Dr Wells) and Health Sciences (Dr Dulberg), University of Ottawa, Ottawa, Ontario.

Presented in part at the Second International Congress on Peer Review in Biomedical Publication, Chicago, Ill, September 10, 1993.

Reprint requests to the Clinical Epidemiology Unit, Loeb Medical Research Institute, Ottawa Civic Hospital, 1053 Carling Ave, Ottawa, Ontario, Canada K1Y 4E9 (Mr Moher).

Downloaded From: by a The Bodleian Libraries of the University of Oxford User on 08/16/2018

Published Reports of Simple Two-Group Parallel Design Randomized Controlled Trials With Negative Results With Dichotomous (n=52) or Continuous (n=18) Primary Outcomes That Had at Least 80% Power to Detect Two Effect Sizes Between Control and Experimental Treatment Groups*

	Effect Size: Relative Difference	
Year	25%	50%
1975	12 (16)	25 (16)
1980	13 (15)	47 (15)
1985	7 (15)	27 (15)
1990	25 (24)	42 (24)
Overall	16 (70)	36 (70)

*Sample size calculations were based on a two-tailed α value of .05 with use of either a z test or a t test, as appropriate for the scale of measurement of the primary outcome measure. Values are percentages (number).

After closely following Frieman and colleagues'[2] selection criteria, our power calculations were performed on the subset of the trials with negative results in which a simple two-group parallel design was used. However, rather than calculating power only for trials with dichotomous outcomes, we also calculated power for trials with continuous primary outcomes. We calculated each trial's power to detect a 25% and a 50% relative change, with an α value of .05, using a z test or a t test, as appropriate for the scale of measurement of the primary outcome. Our calculations differed from those of Frieman and colleagues in that we employed a two-tailed rather than a one-tailed α. A standard program[3] was used for our power calculations. We also calculated the percentages over time of trials reporting sample size calculations.

Selection of Trials and Identification of Primary Outcome

For an RCT to be classified as having negative results, there had to be an explicit statement in the text that negative results had been obtained. When an explicit statement was missing, classification of a trial as having negative results required identification of the primary outcome measure. As Pocock et al[4] observed in 1987, primary outcome measures are not usually clearly specified. Encountering the same problem, we specified a series of decision rules to select the primary outcome. If an article reported a sample size calculation, the outcome used in the calculation was taken as the primary outcome. Published descriptive statistics on this variable were then used in our power calculations. If sample size calculations were not reported and multiple outcomes were evaluated, at least 50% of the results of statistical tests had to be nonsignificant for the RCT to be classified as having negative results. Among the multiple outcomes, the most serious was identified as the primary outcome. For example, if outcomes included both disease-free survival and overall mortality, mortality was considered to be the primary outcome.

RESULTS

Description of Study Sample

A total of 393 RCTs were published in *JAMA*, *Lancet*, and the *New England Journal of Medicine* during the 4 years of our review. Ten trials were excluded from the analysis for the following reasons: results based on invalid statistical analyses precluded classification of the trial (n=5), randomization was not explicit or not all patients were randomized (n=2), and no statistical analysis was performed (n=3).

Twenty-seven percent (n=102) of the remaining 383 trials had negative results. Although the number of RCTs published has more than doubled between 1975 and 1990 (67 vs 148), the percentage that had negative results has remained fairly stable over time: 33%, 27%, 25%, and 25% in 1975, 1980, 1985, and 1990, respectively.

Statistical Power

We calculated power for 70 of the 102 RCTs with negative results that employed a simple two-group parallel design with dichotomous (n=52) and continuous (n=18) primary outcomes. The Table presents the distribution, over time, in the percentage of trials that had at least 80% power (α=.05, two tailed) to detect two effect sizes: a relative difference between treatment groups of 25% and of 50%. Overall, only 16% and 36% of the trials had at least 80% power to detect a 25% or a 50% relative difference, respectively. These figures have not consistently improved over time.

Sample Size Calculations

Among the 102 trials with negative findings, only 32% (n=33) reported a sample size calculation. While this number is small, the situation has improved over time since 1980. None of the 22 trials with negative results published in 1975 was found to have included a sample size calculation, 32% (7/22) did so in 1980, 48% (10/21) did so in 1985, and 43% (16/37) did so in 1990. Only 20 (20%) of the 102 trials with negative results made any kind of statement about the clinical significance of the results with respect to the observed statistical differences between the treatment groups.

An examination of the 33 trials with negative results with sample size calculations revealed serious deficiencies in the reporting of the variables essential for these calculations. No trial indicated the statistical test on which the calculation was based. Only 45% reported the control group event rate, but 79% specified the power level. Slightly more than half (58%) reported the α level, but few (18%) indicated whether the α value was one tailed or two tailed. In fact, in only 30% of these trials was there sufficient detail to enable us to replicate the reported calculated sample size.

COMMENT

If a trial with negative results has a sufficient sample size to detect a clinically important effect, then the negative results are interpretable—the treatment did not have an effect at least as large as the effect considered to be clinically relevant. If a trial with negative results has insufficient power, a clinically important but statistically nonsignificant effect is usually ignored or, worse, is taken to mean that the treatment under study made no difference. Thus, there are important scientific reasons to report sample size and/or power calculations.

There are also ethical reasons to estimate sample size when planning a trial. Altman[5] noted that ethics committees may not want to approve the rare oversized trial because of the unnecessary costs and involvement of additional patients. More commonly, ethics committees may not want to approve trials that are too small to observe clinically important differences, because, as Altman put it, such a trial may "be scientifically useless, and hence unethical in its use of subjects and other resources."

Our results indicate that most trials with negative results had too few patients to detect a relative difference of 25% or 50% with sufficient statistical power and that this has not changed over time. Despite our unique set of decision rules to identify trials with negative results and primary outcomes, the results are similar to those originally reported by Freiman and colleagues[2] 16 years ago.

Our observation that most trials do not report sample size calculations is consistent with other descriptive surveys of RCTs that did not focus solely on trials with negative results. DerSimonian and colleagues[6] and Pocock and colleagues[4] evaluated general methodologic and statistical problems of clinical trials published in 1979 and 1985, respectively. Both reports found that statistical power was discussed in only about 12% of the published RCTs selected for review. More recently, Altman and Doré[7] reported that 39% of a convenience sample of 80 RCTs published in 1987 reported calculating sample size.

Downloaded From: by a The Bodleian Libraries of the University of Oxford User on 08/16/2018

It is possible that investigators do plan required sample sizes but that this information is not included in the published reports, but this does not appear to be the case. After personally contacting principal authors, Liberati and colleagues[8] discovered that only a very small percentage had conducted sample size calculations but not included this information in the published report.

Another explanation for the absence of sample size calculations is a lack of understanding of the concept of effect size. Indeed, one of the most challenging aspects of sample size planning is determining a clinically important effect. The 25% and 50% relative differences on which our power calculations and those of Frieman and colleagues[2] were based may, in fact, represent very large differences. More modest but clinically important treatment effects would necessitate trials with substantially larger sample sizes. Reviewing the cardiovascular literature to evaluate the magnitude of treatment effects, Yusuf and colleagues[9] found that relatively small treatment effects should be expected.

A third possibility for most trials failing to report sample size calculations is that this may reflect the view that planning sample size is unnecessary because RCTs, whatever their outcome, are invaluable to systematic reviews (meta-analyses).[10] Even if one holds this view, it does not preclude the value of publishing post hoc calculation of the power of a study with negative results to detect a clinically important difference. This calculation would enable the reader to make a more informed judgment as to the clinical relevance of the observed absence of statistical significance. Furthermore, evidence indicates that because of publication bias,[11] studies with positive results are more likely to be published than are those with negative results. As a consequence, unpublished trials with negative results might not enter into a systematic review.

In a recent commentary, Cohen[12] found the absence of power calculations in published psychological research to be inexplicable. He suggested that this problem could exemplify the "slow movement of methodological advance" or could reflect difficulties researchers face in performing the appropriate calculations. He also commented that the "passive acceptance of this state of affairs by editors and reviewers is even more of a mystery."

Our results indicate that even when sample size calculations were published, the details necessary to replicate the calculations were missing in most cases. These deficiencies are consistent with a general concern about the quality of reporting of trials.[13,14] Reports of RCTs should provide readers with detailed information about the design, execution, analysis, and interpretation of the trial and its findings. A minimum set of required information, ie, "structured reporting," would help readers to evaluate the validity of a trial. A similar approach, more informative abstracts, has had a positive impact on how the results of abstracts are communicated.[15]

We propose that authors should report sample size calculations and that the following information should be contained in all published reports of RCTs: (1) The primary dependent measure(s) should be clearly identified. (2) A clinically important treatment effect should be specified. (3) The treatment effect should be clearly indicated as being an absolute or a relative difference. (4) The statistical test, directionality, α level, and statistical power used to estimate sample size should be reported. If sample size calculations were not conducted a priori, a published report of an RCT with negative results should provide post hoc statistical power calculations to detect a clinically important difference. The benefits of including this information are clearly worth the extra space required in the publications.

References

1. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1992;102 (suppl):305S-311S.
2. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of 71 'negative' trials. *N Engl J Med.* 1978;299:690-694.
3. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. *Controlled Clin Trials.* 1990;11:116-128.
4. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *N Engl J Med.* 1987;317: 426-432.
5. Altman DG. Statistics and ethics in medical research, III: how large a sample? *BMJ.* 1980;281: 1336-1338.
6. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med.* 1982;306:1332-1337.
7. Altman DG, Doré CJ. Randomization baseline comparisons in clinical trials. *Lancet.* 1990;335:149-153.
8. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol.* 1986;4: 942-951.
9. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med.* 1984;3:409-420.
10. Chalmers TC. Clinical trial quality needs to be improved to facilitate meta-analyses. *Online J Curr Clin Trials.* September 11, 1993; Doc No. 89. Serial online.
11. Dickersin K, Min YI. NIH clinical trials and publication bias. *Online J Curr Clin Trials.* April 28, 1993; Doc No. 50. Serial online.
12. Cohen J. A power primer. *Psychol Bull.* 1992; 112:155-159.
13. Grant A. Reporting controlled trials. *Br J Obstet Gynecol.* 1989;96:397-400.
14. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials. *Controlled Clin Trials.* 1980;1:37-58.
15. Haynes RB, Mulrow CD, Huth EJ, Mtman DG, Gardner MJ. More informative abstracts revisited. *Ann Intern Med.* 1990;113:69-76.

Downloaded From: by a The Bodleian Libraries of the University of Oxford User on 08/16/2018