Journal of Negative Results in BioMedicine

BioMed Central

Research

Prominent medical journals often provide insufficient information to assess the validity of studies with negative results

Randy S Hebert*[1], Scott M Wright[2], Robert S Dittus[3] and Tom A Elasy[3]

Address: [1]Division of General Internal Medicine, University of Pittsburgh, 933W MUH, 200 Lothrop Street, Pittsburgh, PA 15213, USA, [2]Division of General Internal Medicine, Johns Hopkins Bayview Medical Center, Johns Hopkins University School of Medicine, A6W, 4940 Eastern AvenueBaltimore, MD 21224, USA and [3]Division of General Internal Medicine, Vanderbilt University Medical Center, S-1121 MCN 2587, Nashville, TN 37232, USA

E-mail: Randy S Hebert* - hebertrs@msx.upmc.edu; Scott M Wright - swright@jhmi.edu; Robert S Dittus - robert.dittus@Vanderbilt.Edu; Tom A Elasy - tom.elasy@Vanderbilt.Edu

*Corresponding author

Published: 30 September 2002

Journal of Negative Results in BioMedicine 2002, 1:1

Received: 19 July 2002
Accepted: 30 September 2002

This article is available from: http://www.jnrbm.com/content/1/1/1

Abstract

Background: Physicians reading the medical literature attempt to determine whether research studies are valid. However, articles with negative results may not provide sufficient information to allow physicians to properly assess validity.

Methods: We analyzed all original research articles with negative results published in 1997 in the weekly journals BMJ, JAMA, Lancet, and New England Journal of Medicine as well as those published in the 1997 and 1998 issues of the bimonthly Annals of Internal Medicine (N = 234). Our primary objective was to quantify the proportion of studies with negative results that comment on power and present confidence intervals. Secondary outcomes were to quantify the proportion of these studies with a specified effect size and a defined primary outcome. Stratified analyses by study design were also performed.

Results: Only 30% of the articles with negative results comment on power. The reporting of power (range: 15%-52%) and confidence intervals (range: 55–81%) varied significantly among journals. Observational studies of etiology/risk factors addressed power less frequently (15%, 95% CI, 8–21%) than did clinical trials (56%, 95% CI, 46–67%, $p < 0.001$). While 87% of articles with power calculations specified an effect size the authors sought to detect, a minority gave a rationale for the effect size. Only half of the studies with negative results clearly defined a primary outcome.

Conclusion: Prominent medical journals often provide insufficient information to assess the validity of studies with negative results.

Background

Physicians are faced with the challenge of assessing whether the conclusions of research studies are valid. Power, the probability that a study will detect an effect of a specified size, is analogous to the sensitivity of a diagnostic test. [1] Just as a negative result does not rule out disease when the test applied has low sensitivity, a negative study with inadequate power cannot disprove a research hypothesis. Power/sample size calculations play an important role in study planning, give readers an idea of

the adequacy of the investigation, and help readers assess the validity of studies with negative results. [2–4] Effect size (delta) is a critical component of power calculations. Investigators choose from a wide range of possible deltas when calculating sample size. Clinicians and investigators also often struggle to determine what effect size is reasonable to expect.[2,5–8] Consequently, it is important for investigators to report the effect size they wish to detect. However, this is often neglected.[8]

Sample size calculations alone are insufficient for the interpretation of studies with negative results; power and confidence intervals compliment each other and should both be reported.[6,9] Confidence intervals take into account the data actually collected, define the upper and lower range consistent with a study's data, provide an estimate of precision, and can give readers some indication of the clinical significance of the results. [10–13]

Our work adds to the literature in several ways. Several authors have found that many randomized controlled trials were underpowered, or had an unacceptable risk of missing an important effect due to inadequate sample size. [14–21] Because power calculations are often complicated,[21] many readers are unlikely to have the statistical sophistication necessary to perform a power analysis. Therefore, we were interested in whether articles provided information necessary for readers to assess the validity of studies with negative results. We looked for evidence of power/sample size calculations and effect size. In addition, unlike prior work, we examined studies for documentation of confidence intervals.[22] Finally, because the calculation of sample size is applicable to all comparative studies, we did not limit our study to randomized controlled trials.[23]

Our primary objective was to quantify the proportion of studies with negative results within prominent general medical journals[24] that comment on power and present confidence intervals. Secondary outcomes were to quantify the proportion of these studies with a specified delta and a defined primary outcome.

Methods

All articles from the 1997 issues of the British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), Lancet, and the New England Journal of Medicine (NEJM) were reviewed. Because the Annals of Internal Medicine (Annals) is published bimonthly, all articles from 1997 and 1998 were reviewed so as to include a comparable number of articles. One investigator (RSH) manually searched the journals and reviewed all articles for eligibility. Review articles, meta-analyses, modeling studies, decision and cost-effective analyses, case reports, editorials, letters, and studies without inferential statistics (i.e. descriptive studies) were excluded. Equivalence trials (studies designed to show equivalent efficacy of treatments) were included because power analysis, confidence intervals, and delta are particularly important to their design. Methodological issues involved in the design and analysis of these studies have been described elsewhere.[25,26]

Articles were classified as having negative results if 1) the primary outcome(s) was not statistically significant (i.e. the article had an explicit statement that the comparison between two groups did not reach statistical significance) or 2) in those articles with no primary outcome(s), any of the first three outcomes were not statistically significant. Other outcomes were not evaluated. A second author (TAE) reviewed the full text of a simple random sample of 50 articles and the kappa statistic was calculated to assess the intraobserver variability for our classification scheme.

We examined articles to see if the authors named a primary outcome variable. We employed a decision rule, modified from Moher and colleagues, to define the primary outcome in those articles where none was specified.[19] If an article reported a sample size calculation, this was assumed to be the primary outcome.[27] If calculations were not performed, a total of three outcomes, if present, were examined. In those articles with multiple outcomes and none defined as primary, the three outcomes evaluated were the first three listed in the abstract (or result section if less than three outcomes were listed in the abstract).

The full text of included articles was systemically reviewed. Data was abstracted by a single author (RSH) and recorded in standardized fashion. Information was recorded on whether the article had a primary outcome(s), commented on power, sample size calculations, and confidence intervals pertaining to the outcomes evaluated, a projected delta, and a reason for this delta. A paper was given credit for addressing power if sample size calculations or comments on power/sample size were present. Power, sample size calculations, and confidence intervals could pertain to any one of the three outcomes evaluated and was not necessary for all outcomes.

Comparisons were made across journals by Chi-square analysis. We also assessed articles for comment on power and/or presentation of confidence intervals while stratifying by study design (clinical trials, observational studies of etiology/risk factors, screening/diagnosis, prognosis, and other). Responses were summarized as proportions and 95% confidence intervals. All data was analyzed using STATA 6.0 (Stata Corp., College Station, TX).

Table 1: Negative articles addressing power/sample size and confidence intervals

Journal	Power/Sample Size*	Confidence Intervals†	Power/Sample Size and Confidence Intervals*
		n (%, 95% CI)	
Annals	6/41 (15, 3–26)	33/41 (80, 68–93)	5/41 (12, 2–23)
BMJ	11/57 (19, 9–30)	46/57 (81, 70–91)	10/57 (18, 7–28)
JAMA	10/44 (23, 10–36)	24/44 (55, 39–70)	3/44 (7, 0–15)
Lancet	24/46 (52, 37–67)	34/46 (74, 61–87)	20/46 (43, 29–58)
NEJM	19/46 (41, 27–56)	34/46 (74, 61–87)	13/46 (28, 15–42)
Total	70/234 (30, 24–36)	171/234 (73, 67–79)	51/234 (22, 16–27)

* $P < 0.001$ † $P = 0.038$

Results

One thousand thirty eight articles were eligible for analysis. Two hundred thirty four (23%) were classified as negative. There was good agreement between observers in the classification of articles (k = 0.74). The percent of negative articles per journal was: Annals 20% (41/203), BMJ 22% (57/256), JAMA 23% (44/191), Lancet 22% (46/205), and NEJM 25% (46/183) (p = 0.857).

Thirty percent (70/234) of studies with negative results had comments on power and/or sample size calculations. Seventy three percent (171/234) included confidence intervals. The reporting of power (range: 15%-52%) and confidence intervals (range: 55–81%) varied significantly among journals. Twenty two percent of the studies included both power/sample size calculations and confidence intervals. There existed significant variation between journals in the reporting of power/sample size calculations and confidence intervals (Table 1). Because clinical trials (n = 87) and observational studies of etiology/risk factors (n = 109) were the predominant study designs (84% of the negative studies), articles with other study designs were not examined further. Fifty six percent (95% CI, 46–67%) of negative clinical trials and 15% (95% CI, 8–21%) of negative observational risk factor/etiology studies addressed power/sample size ($p < 0.001$). For reporting confidence intervals, the corresponding percentages were 79% (95% CI, 71–87%) and 75% (95% CI, 65–84%), respectively (p = 0.489).

Of the negative articles including information about sample size, 87% (61/70) specified a delta or the effect size that the authors sought to detect. A minority, 43% (26/61), explained the rationale behind the delta chosen. Of these, 77% (20/26) cited references or pilot studies to support their rationale.

Only 52% (122/234) of articles with negative results had a clearly defined primary outcome(s).

Discussion

Many articles underreport power/sample size calculations and confidence intervals. Significant variation exists among journals. Our work demonstrates that power was reported more often in clinical trials than in observational studies of etiology/risk factors. Investigators involved in randomized clinical trials may be more familiar with the importance of power and sample size calculation.[28] Also, investigators conducting observational studies often do not have the ability to determine sample size prior to beginning their work. Most articles with sample size calculations reported a projected effect size, but only a minority shared the rationale behind this delta and even less provided empiric evidence to support the rationale.

While this manuscript describes an analysis of a large body of studies with negative results, several limitations must be considered. First, although most negative studies did not list power/sample size calculation, we cannot be certain this had not been performed a priori. It is also possible that, for the sake of brevity, authors and/or editors omitted power/sample size calculations from the final text when preparing manuscripts for submission. While it is possible these calculations were done but not reported, this may not be the case.[29] Second, our definition of a negative study may seem unduly broad. We examined three outcomes in order to classify articles because articles frequently report several outcomes, often with none defined as primary. [30–33] Previous authors, limiting their work to randomized controlled trials, who have encountered multiple outcomes have defined the primary outcome as "the most clinically important"[19] or the outcome that was the "primary focus of the article".[20] These outcomes are often not possible to discern in observational studies. Nonetheless, our results may represent a best-case scenario given the publication bias against articles with negative results and the fact that we examined the more prominent general medical journals.[34]

Conclusions

In summary, this study demonstrates that prominent medical journals often provide insufficient information to assess the validity of studies with negative results. Authors and journal editors need to include this information so readers can be informed consumers of the medical literature.

Competing interests

1. The research was not supported by external funds.

2. There are no competing interests including financial, stocks, honoraria, speaker's fees, and any competing academic, religious, moral, or personal interests for all authors.

3. We have no financial interest in the material contained in the manuscript.

4. The manuscript is neither under review by another publisher nor previously published.

5. All authors have participated in the design, analysis, and writing of the accompanying manuscript.

6. All authors have approved the final manuscript and have taken care to ensure the integrity of the work.

Authors' contributions

1. Randy S Hebert MD MPH

- Conception and design
- Acquisition of data
- Analysis and interpretation of data
- Drafted and revised the article
- Gives final approval of the version for publication

2. Scott M Wright MD

- Analysis and interpretation of data
- Revised the article for important intellectual content
- Gives final approval of the version for publication

3. Robert S Dittus MD MPH

- Analysis and interpretation of data
- Revised the article for important intellectual content
- Gives final approval of the version for publication

4. Tom A Elasy MD MPH

- Conception and design
- Analysis and interpretation of data
- Drafted and revised the article
- Gives final approval of the version for publication

References

1. Browner WS, Newman TB: **Are all significant P values created equal? The analogy between diagnostic tests and clinical research.** *JAMA* 1987, **257(18):**2459-2463
2. Goodman SN, Berlin JA: **The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results.** *Ann Intern Med* 1994, **121(3):**200-206
3. Gardner MJ, Machin D, Campbell MJ: **Use of checklists in assessing the statistical content of medical studies.** *In: Statistics with Confidence. (Edited by: Gardner MJ, Altman DG) London: British Medical Journal* 1989, 101-108
4. Halpern SD, Karlawish JH, Berlin JA: **The continuing unethical conduct of underpowered clinical trials.** *JAMA* 2002, **288(3):**358-362
5. Raju TN, Langenberg P, Sen A, Aldana O: **How much 'better' is good enough? The magnitude of treatment effect in clinical trials.** *Am J Dis Child* 1992, **146(4):**407-411
6. Smith AH, Bates MN: **Confidence limits vs. power calculations.** *Epidemiology* 1994, **5(2):**268-269
7. Hanley JA: **Confidence limits vs. power calculations.** *Epidemiology* 1994, **5(2):**264-266
8. Lipsey MW: **Design Sensitivity. Statistical Power for Experimental Research.** *Newbury Park: Sage Publications* 1990
9. Ware JH, Mosteller F, Delgado F, *et al*: **P Values.** *In: Medical Uses of Statistics. (Edited by: Bailar III J, Mosteller F) Boston, MA: NEJM Books* 1992, 181-200
10. Braitman LE: **Confidence intervals assess both clinical significance and statistical significance.** *Ann Intern Med* 1991, **114(6):**515-517
11. Braitman LE: **Confidence intervals extract clinically useful information from data.** *Ann Intern Med* 1988, **108(2):**296-298
12. Simon R: **Confidence intervals for reporting results of clinical trials.** *Ann Intern Med* 1986, **105(3):**429-435
13. Smith AH, Bates MN: **Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies.** *Epidemiology* 1992, **3(5):**449-542
14. Freiman JA, Chalmers TC, Smith H, Kuebler RR Jr: **The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials.** *N Engl J Med* 1978, **299(13):**690-694
15. Edlund MJ, Overall JE, Rhoades HM: **Beta, or type II error in psychiatric controlled clinical trials.** *J Psychiatr Res* 1985, **19(4):**563-567
16. Hall JC: **The other side of statistical significance: a review of Type II errors in the Australian Medical Literature.** *Aust NZ J Med* 1982, **12(1):**7-9
17. Brown CG, Kelen GD, Ashton JJ, Werman HA: **The beta error and sample size determination in clinical trials in emergency medicine.** *Ann Emerg Med* 1987, **16(2):**183-187
18. Williams HC, Seed P: **Inadequate size of 'negative' clinical trials in dermatology.** *Br J Dermatol* 1993, **128(3):**317-326
19. Moher D, Dulberg CS, Wells GA: **Statistical power, sample size, and their reporting in randomized controlled trials.** *JAMA* 1994, **272(2):**122-124
20. Dimick JB, Diener-West M, Lipsett PA: **Negative results of randomized clinical trials published in the surgical literature.** *Arch Surg* 2001, **136(7):**796-800

21. Freedman KB, Back S, Bernstein J: **Sample size and statistical power of randomised, controlled trials in orthopaedics.** *J Bone Joint Surg Br* 2001, **83(3):**397-402
22. Rhoads GG: **Reporting of power and sample size in randomized controlled trials.** *JAMA* 1995, **273(1):**22-23
23. Altman DG: **Practical Statistics for Medical Research.** *London: Chapman & Hall/CRC* 1991
24. ISI Thompson Scientific: **ISI Journal Citation Reports.** *Institute for Scientific Information. Accessed 8/01.* 2001 [http://jcrweb.com/]
25. Jarvis JB, Lewis JA, Ebbutt AF: **Trials to assess equivalence: the importance of rigorous methods.** *BMJ* 1996, **313(7048):**36-39
26. Jarvis JB, Lewis JA, Ebbutt AF: **Claims of equivalence in randomized controlled trials to the treatment of bacterial meningitis in children.** *Pediatr Infect Dis J* 2002, **21(8):**753-757
27. Feinstein AR: **Clinical biostatistics. XXXIV. The other side of 'statistical significance': alpha, beta, delta, and the calculation of sample size.** *Clin Pharmacol Ther* 1975, **18(4):**491-505
28. Moher D, Jones A, Lepage L, CONSORT Group: **Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation.** *JAMA* 2001, **285(15):**1992-5
29. Liberati A, Himel HN, Chalmers TC: **A quality assessment of randomized control trials of primary treatment of breast cancer.** *J Clin Oncol* 1986, **4(6):**942-951
30. Gotzsche PC: **Methodology and overt and hidden bias in reports of 196 double-blind trials of non-steroidal anti-inflammatory drugs in rheumatoid arthritis.** *Control Clin Trials* 1989, **10(1):**31-56
31. Smith DG, Clemens J, Crede W, *et al*: **Impact of multiple comparisons in randomized clinical trials.** *Am J Med* 1987, **83(3):**545-550
32. 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(7):**426-432
33. Pocock SJ: **Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation.** *Control Clin Trials* 1997, **18(6):**530-545
34. Callaham ML, Wears RL, Weber EJ, *et al*: **Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting.** *JAMA* 1998, **280(3):**254-257