Editorial

Adding Nails to the Coffin of Underpowered Trials

More than 50 years have passed since the British Medical Research Council published what is widely regarded as the first modern randomized clinical trial (RCT)[1]. In the interim, RCT have increased not only in frequency, but also in breadth. Studies that were once published almost exclusively in select, high-readership, general medical journals are now increasingly found among the pages of journals with all readerships and impact ratings. Regrettably, there are reasons to suspect that the increased use of this powerful tool has not been accompanied by comparable improvements in the methodologic conduct of RCT. As a result, the value of RCT has not been fully realized.

In this issue of *The Journal*, Keen and colleagues[2] describe how one contributor to trial value — statistical power — is incompletely reported, and often inadequate, among RCT published in the rheumatology literature. The authors report that among RCT published in the rheumatology literature in 2001–2002, only 42% reported a power calculation, and only 50% of negative trials (and 79% of trials overall) had adequate power to reliably detect a true treatment effect. These findings should come as no surprise: other investigators have reported similar problems in the general medical[3-8], emergency medical[9], general surgical[10,11], plastic surgical[12], orthopedic[13], psychiatric[14], dermatologic[15], head injury[16], medication compliance[17], and family practice[18] literatures.

Although several factors contribute to an RCT's power, and definitions of acceptable power are more traditional than empiric[19], these studies have generally defined underpowered trials as those that enrol too few participants to identify differences between interventions at least 80% of the time that such differences truly exist. Underpowered RCT are therefore overly prone to making false-negative conclusions, or committing what epidemiologists call type II errors. Such trials are widely[19-22], but not universally[23,24], considered to be unethical, primarily because they expose participants to the risks and burdens of research without providing commensurate opportunity for their participation to contribute to generalizable knowledge.

Given that the linked epidemics of underpowered trials and trials with inadequate reporting are already well described, some readers may question why the study by Keen, *et al*[2] was necessary. I believe there are several benefits to repeatedly documenting these problems. First, by broadly casting the net of underpowered trials across various clinical disciplines, investigators like Keen, *et al* limit the possibility that future investigators might feel immune to the deficient conduct and reporting of RCT in other fields.

Second, despite focusing on a common theme, each report reveals novel insights, including at least 2 in the study by Keen and colleagues. First, the authors found that positive trials were no more likely than negative trials to report a power calculation. This is a concern because when authors fail to openly report the details of a trial's methods, including sample size calculations, readers are unable to determine whether significant findings are real or potential artefacts of a poorly designed study. Second, Keen, *et al* found that even among high-impact journals, nearly half of negative RCT were underpowered. This suggests that casual readers cannot assume that just because they are reading a highly respected journal, all reports found within will be well conducted.

An important caveat in interpreting the study by Keen, *et al* is that there are several reasons to believe that the actual prevalence of underpowered rheumatology trials is probably much greater than the 21% these authors report. First, the authors investigated only trials published in journals indexed in Medline. Because underpowered trials are more likely to yield negative findings, and negative trials are less likely to be published, this approach may selectively fail to identify underpowered trials, and thereby underestimate the proportion of RCT that are underpowered.

Second, the authors calculated power for all negative or indeterminate RCT, but assumed that positive trials were, by definition, adequately powered. This assumption reflects a common misconception; in fact, positive RCT might also be underpowered, and simply get lucky by finding an unexpectedly large treatment difference, thereby yielding statis-

See Prevalence of underpowered RCT in rheumatology, *page 2083*

tically significant results. Such a fortuitous outcome does not make these trials adequately powered because these same trials may have failed to detect smaller outcomes that were nonetheless clinically important[19]. Because power is an *a priori* phenomenon, finding a statistically significant result after the fact does not save an otherwise underpowered trial.

Third, as the authors acknowledge, they arbitrarily required that studies be powered to detect an effect size of 20% across all disease areas. However, it is not clear whether this effect size refers to a relative or an absolute difference. For example, using a dichotomous outcome measure in which patients are considered either responders or non-responders to therapy, an absolute difference of 20% might mean that 50% of patients responded to the inferior treatment, and 70% responded to the superior treatment. By contrast, a relative difference of 20% might mean that 50% of patients responded to the inferior treatment, but 60% (20% more than 50%) responded to the superior treatment. The distinction between relative and absolute differences not only influences clinical interpretation of individual studies, but also influences estimates of the proportion of trials considered underpowered.

In addition, regardless of whether Keen, *et al* were referring to absolute or relative differences, the choice of 20% may be substantially larger than differences deemed by many practicing rheumatologists to be clinically significant. Keen, *et al* properly note that 20% has been identified as a clinically important difference in studies of osteoarthritis and rheumatoid arthritis, but identifying differences as small as 5% or 10% may be important to physicians caring for patients with more devastating diseases such systemic lupus erythematosus, systemic sclerosis, or many of the vasculitides. If so, then the trials of these syndromes that were counted as adequately powered by Keen, *et al* because they were able to detect a 20% difference, might have been inadequately powered to detect smaller, yet clinically important, differences.

For each of these reasons, the prevalence of underpowered RCT in rheumatology reported by Keen, *et al* should be viewed as a minimal estimate. The fact that even this best-case scenario is unfavorable suggests continued need for exploration into, and public notification of, the problem of underpowered clinical trials. Because physician and investigator behaviors change slowly, only with repeated documentation can we improve the overall quality of reporting, and ensure that patients can participate in more potentially valuable studies. This issue may appear to be a dead horse, but I believe it is one still worth beating.

SCOTT D. HALPERN, MD, PhD, M Bioethics,
Center for Clinical Epidemiology and Biostatistics,
University of Pennsylvania School of Medicine,
115 Blockley Hall, 423 Guardian Drive,
Philadelphia, PA, 19104-6021.
E-mail: shalpern@cceb.med.upenn.edu

Address reprint requests to Dr. Halpern.

REFERENCES

1. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948;2:769-82.
2. Keen HI, Pile K, Hill CL. The prevalence of under-powered randomized clinical trials in rheumatology. J Rheumatol 2005;32:2083-8.
3. Freiman JA, Chalmers TC, Smith H, Jr, 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;299:690-4.
4. Geiman BJ, Donohoe MT. Statistical power and reporting of sample size calculations in randomized controlled trials. J Gen Intern Med 1999;14 Suppl 2:98.
5. Hall JC. The other side of statistical significance: a review of Type II errors in the Australian medical literature. Aust N Z J Med 1982;12:7-9.
6. Hebert R, Wright S, Dittus R, Elasy T. Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. J Negat Results Biomed 2002;1:1.
7. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-4.
8. Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. J Clin Epidemiol 2003;56:717-20.
9. 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;16:183-7.
10. 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.
11. Schumm LP, Fisher JS, Thisted RA, Olak J. Clinical trials in general surgical journals: Are methods better reported? Surgery 1999;125:41-5.
12. Chung KC, Kaliainen 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.
13. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg 1999;81:1454-60.
14. Edlund MJ, Overall JE, Rhoades HM. Beta, or type II error in psychiatric controlled clinical trials. J Psychiatr Res 1985;19:563-7.
15. Williams HC, Seed P. Inadequate size of 'negative' clinical trials in dermatology. Br J Dermatol 1993;128:317-26.
16. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I. Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 2000;320:1308-11.
17. Nichol M, Venturini F, Sung J. A critical evaluation of the methodology of the literature on medication compliance. Ann Pharmacother 1999;33:531-40.
18. Mengel MB, Davis AB. The statistical power of family practice research. Fam Pract Res J 1993;13:105-11.
19. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358-62.
20. Newell DJ. Type II errors and ethics. BMJ 1978;iv:1789.
21. Altman DG. The scandal of poor medical research. BMJ 1994;308:283-4.
22. Prentice R. Ethics and sample size - another view. Am J Epidemiol 2005;161:111-2.
23. Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. Lancet 1997;350:804-7.
24. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. Am J Epidemiol 2005;161:105-10.