The Continuing Unethical Conduct of Underpowered Clinical Trials

Scott D. Halpern, MSCE
Jason H. T. Karlawish, MD
Jesse A. Berlin, ScD

Despite long-standing critiques of the conduct of underpowered clinical trials, the practice not only remains widespread, but also has garnered increasing support. Patients and healthy volunteers continue to participate in research that may be of limited clinical value, and authors recently have offered 2 related arguments to support the validity and value of underpowered clinical trials: that meta-analysis may "save" small studies by providing a means to combine the results with those of other similar studies to enable estimates of an intervention's efficacy, and that although small studies may not provide a good basis for testing hypotheses, they may provide valuable estimates of treatment effects using confidence intervals. In this article, we examine these arguments in light of the distinctive moral issues associated with the conduct of underpowered trials, the disclosures that are owed to potential participants in underpowered trials so they may make autonomous enrollment decisions, and the circumstances in which the prospects for future meta-analyses may justify individually underpowered trials. We conclude that underpowered trials are ethical in only 2 situations: small trials of interventions for rare diseases in which investigators document explicit plans for including their results with those of similar trials in a prospective meta-analysis, and early-phase trials in the development of drugs or devices, provided they are adequately powered for defined purposes other than randomized treatment comparisons. In both cases, investigators must inform prospective subjects that their participation may only indirectly contribute to future health care benefits.

JAMA. 2002;288:358-362 www.jama.com

More than 20 years have passed since investigators first described the ethical problems of conducting randomized controlled trials (RCTs) with insufficient statistical power.[1,2] Because such studies may not adequately test the underlying hypotheses, they have been considered "scientifically useless"[2] and therefore unethical in their exposure of participants to the risks and burdens of human research.[2-4] Despite this long-standing challenge, many clinical investigators continue to conduct underpowered studies[5,6] and fail to calculate or report appropriate (a priori) power analyses.[6-10] Not only do these scientific and ethical errors persist in the general medical literature, but 3 recent reports[11-13] also highlight the alarming prevalence of these problems in more specialized fields.

Patients and healthy volunteers thus continue to participate in research that may be of limited clinical value.[4,14] Furthermore, authors[15-17] recently have offered 2 related arguments to support the validity and value of underpowered clinical trials. First, meta-analysis may "save" small studies by providing a means to combine the results with those of other similar studies to enable estimates of an intervention's efficacy. Second, although small studies may not provide a good basis for testing hypotheses, they may provide valuable estimates of treatment effects using confidence intervals. Based on these arguments, authors have suggested that institutional review boards (IRBs) drop the documentation of statistical power as a criterion for study approval.[16]

If meta-analysis and estimating treatment effects provided investigators and IRBs with a means to justify underpowered research, challenges to the ethics of underpowered studies would have to cease. In this article, we examine these arguments in light of the distinctive moral issues associated with the conduct of underpowered trials, the disclosures that are owed to potential participants in underpowered trials so they

See also p 363.

Author Affiliations: Center for Clinical Epidemiology and Biostatistics (Mr Halpern and Dr Berlin), Center for Bioethics (Mr Halpern and Dr Karlawish), Center for Education and Research on Therapeutics (Mr Halpern and Dr Berlin), Department of Medicine (Dr Karlawish), and Leonard Davis Institute of Health Economics (Dr Karlawish), University of Pennsylvania School of Medicine, Philadelphia.
Corresponding Author and Reprints: Scott D. Halpern, MSCE, Center for Clinical Epidemiology and Biostatistics, 108 Blockley Hall, 423 Guardian Dr, Philadelphia, PA 19104-6021 (e-mail: shalpern@mail.med.upenn.edu).

may make autonomous enrollment decisions, and the circumstances in which the prospects for future meta-analyses may justify individually underpowered trials.

We conclude that underpowered trials can be ethical in only 2 situations. First, small trials of interventions for rare diseases may be justified if investigators document explicit plans for including their results with those of similar trials in a prospective meta-analysis. Second, early-phase trials in the development of drugs, devices, or other interventions need not be powered to make randomized treatment comparisons provided they are adequately powered for other defined purposes and designed to guide the conduct of subsequent, comparative trials. In both cases, investigators must inform prospective participants that their participation may only indirectly contribute to future health care benefits.

STATISTICAL POWER AND THE PLANNING OF CLINICAL TRIALS

Investigators use power analysis to determine the probability that a given study will reject the null hypothesis when it is, in fact, false. In other words, power analysis determines the chance of detecting a true-positive result. By tradition, researchers consider a study to be adequately powered if it has at least an 80% chance of detecting a clinically significant effect when one exists. This exact value is arbitrary; higher power will always be preferable and should be set with consideration of the importance of limiting both false-negative conclusions (ie, type II errors) and false-positive conclusions (ie, type I errors).

To calculate a study's power to detect a given effect, investigators use a set of other variables, including the number of individuals to be enrolled, the expected variability of their outcomes, and the chosen probability of making a type I error. Reformulating these variables allows one to calculate the numbers of study participants needed to detect a clinically important effect size with acceptable power. Although consensus among reasonable clinicians will generally enable determinations of how small an effect would be clinically important to detect, disagreement about this value may occasionally emerge. In such cases, we advocate a 3-tiered, hierarchical approach for investigators to use in determining the effect size to be entered into sample size calculations.

First, when empirical definitions of clinically meaningful effects exist, such as in the percentage reduction of reported pain necessary to define analgesic efficacy,[18] these values ought to be used. Second, if there is neither clinical consensus nor empirical evidence to guide definitions of clinically important effects, but data from earlier trials or observational studies reliably indicate an intervention's plausible effect, this value may be used. Finally, if none of the foregoing criteria are met, then previously published definitions of moderate effect sizes, such as those described by Cohen,[19] should be used. Trials that cannot reliably detect effect sizes defined using this hierarchical approach may be defined as underpowered.

ARGUMENTS FOR ALLOWING UNDERPOWERED TRIALS

There are several practical barriers to conducting large RCTs, particularly for rare diseases. Because the results of smaller, underpowered trials may later be combined in meta-analyses, authors have argued that prohibiting underpowered trials would "thwart many independent investigations . . . [which] may seriously diminish the stock of the world's knowledge."[16] There are both practical and ethical problems with this argument.

The first practical problem stems from an overly optimistic view of the usefulness of the information that underpowered trials may provide. Acknowledging that hypothesis tests are inordinately likely to produce false-negative results when inadequately powered, proponents argue that quantifying the range of plausible effect sizes will still be possible by examining confidence intervals.[16] However, studies containing too few subjects to detect a positive effect (if one exists) via hypothesis testing will also yield unacceptably wide confidence intervals around the point estimate of this effect. Because such confidence intervals will often contain both the null and clinically important effect sizes, the approach provides ambiguous conclusions.

One might argue that if no trials were conducted, the confidence intervals around the (unknown) effect would remain infinitely wide. Thus, any well-designed trial, no matter how small, would at least reduce this uncertainty. However, the marginal value of narrowing confidence intervals to widths still compatible with both positive and negative results generally is insufficient to justify exposing individuals to the common risks and burdens of research. Although these risks and burdens may often be outweighed by the benefits of trial participation,[20] these beneficial effects are not uniform,[20] and their potential is insufficient to justify human research.[14]

The second practical problem with meta-analysis is that even if investigators conducted multiple underpowered trials, difficulties in synthesizing the results may prevent the calculation of valid treatment effects. Under ideal conditions, meta-analyses offer potential advantages over a single RCT in gauging a treatment effect. Meta-analyses may enhance generalizability by incorporating more heterogeneous populations and may overcome the risk that any single RCT, even a very large one, could be weakened by bias.

For meta-analyses to be useful, however, comparable research methods must have been used among the primary trials, and these trials must be selected for inclusion in an unbiased fashion. The infrequency with which these ideal conditions are met may help explain why 2 independent meta-analyses of the same literature sometimes arrive at different conclusions.[21,22]

As Bailar notes, "Such disagreement argues powerfully against any notion that meta-analysis offers an assured way to distill the 'truth' from a collection of research papers."[22]

Finally, because underpowered trials are more likely to produce negative results and consequently may not be published (the so-called publication bias[23,24]), underpowered trials may be less accessible for inclusion in meta-analyses.[25] This may fatally bias the approach. Thus, the ideal conditions for combining evidence may be particularly unlikely when the component trials are underpowered; therefore, even the most rigorously conducted meta-analyses will be unable to augment such trials' abilities to further medical knowledge. Only if widely accessible registries of RCTs[26,27] are expanded to include privately sponsored trials could the potential for publication bias in retrospective meta-analyses be eliminated.[28]

SAMPLE SIZE AND INFORMED CONSENT

In addition to the practical problems mentioned herein, underpowered studies will also be ethically deficient if investigators do not convey these studies' limited value to prospective participants. Failure to communicate a study's value (or lack thereof) limits the quality of the information on which individuals must base their enrollment decisions. Individuals commonly participate in research to fulfill altruistic motives, such as desires to advance medical science and thereby help others.[29-35] Therefore, to respect prospective participants' autonomy, investigators must inform them of the limited capacities of small trials to produce public benefit.

Investigators occasionally deprive participants of such information for 3 reasons. First, investigators may simply fail to conduct an a priori power analysis. Such investigators are acting negligently. In addition to risking the enrollment of too few participants to answer the research question, investigators who fail to conduct or improperly conduct a power analysis may enroll too many individuals. This outcome is also troubling because it exposes too many individuals to the risks of research and overconsumes limited societal resources.[2,36,37]

Second, investigators might conduct an appropriate power calculation but fail to recruit sufficient numbers of participants in a timely fashion.[38,39] Such cases may arise, for example, when prospective participants' clinicians have reservations about enrolling their patients[40] or when patients themselves are dissuaded by some feature of the trial, such as the existence of a placebo group.[34,41] Investigators should attempt to identify potential recruitment problems beforehand and modify their approaches accordingly.[42]

Perhaps most concerning is the third scenario, in which investigators conduct an appropriate power analysis, find they will be unlikely to recruit an adequate number of participants, and choose to proceed without conveying this information to participants in the informed consent process. This knowing failure of information disclosure entails deception. In addition to abrogating participants' rights, if such deception were publicized, it could undermine people's trust in science, further curtailing future enrollment.

Investigators may fear that disclosing information regarding power will itself reduce enrollment. Because study participants so often seek to fulfill altruistic motives,[29-35,43] it seems logical that they would rather participate in adequately powered trials. Nonetheless, this potential barrier to efficient recruitment does not justify enrolling individuals without full disclosure.

RARE DISEASES

For research on diseases with low prevalence or incidence, the numbers of afflicted (or newly afflicted) individuals at any one time may make it impossible to conduct even a multicenter RCT that could reliably distinguish between interventions. It has been argued that in such cases some evidence is better than none.[16,17] This view ignores the fact that only when the effect sizes are extremely large—indeed, larger than anticipated—will small trials be able serendipitously to document them. In all other cases, false-negative conclusions may be drawn and post hoc power analyses will be unable to elucidate the error.[44] Although investigators commonly relax inferential standards to avoid this result, doing so increases the risk of drawing false-positive conclusions.

Instead, if investigators explicitly plan to make the results of a small trial available for inclusion in a prospective meta-analysis,[21] excessive risks of both types of false conclusions may be averted. Prospectively designed meta-analyses are less susceptible to the problems with traditional, retrospective meta-analyses because the methods of the component studies may be synchronized in advance. This avoids the possibility that component studies may not be combinable if, for example, one study investigated a high-dose intervention among men and another study investigated a lower dose of the intervention among women.[45] Because dose and sex would be inextricably confounded between these studies, retrospective meta-analysis would be of little use.

Therefore, only prospectively designed meta-analyses can justify the risks to participants in individually underpowered trials because they provide sufficient assurance that a study's results will eventually contribute to valuable or important knowledge.[46] Although a multicenter trial could similarly contribute to generalizable knowledge and may provide more internally valid results, it requires that investigators have access to a sufficient number of patients during the trial's conduct. This may not be possible for very rare diseases, making prospective meta-analyses of single-center and multicenter trials necessary to obtain adequate power. Furthermore, prospectively designed meta-analyses retain the innovation possible in conducting several smaller studies, while providing the organizational frame-

work to ensure that their results can be synthesized.

EARLY-PHASE STUDIES OF EXPERIMENTAL INTERVENTIONS

Just as prospective meta-analyses may ensure the value of small single studies for rare diseases, plans for large, comparative trials of experimental interventions can justify the conduct of small studies in earlier phases of drug or device development. Thus, several smaller phase 1/2 trials may be justified as long as each is adequately powered for another aim, such as reliably determining whether a new therapy shows at least some promise of benefit, and is explicitly aimed at guiding a definitive phase 3 trial that will be adequately powered to make a reliable treatment comparison. Investigators conducting these studies must tell participants that their participation will not directly provide information of immediate clinical value, but rather will guide future studies that may do so.

CONCLUSION

Despite long-standing critiques of the conduct of underpowered clinical trials, the practice not only remains widespread, but also has garnered increasing support. We have provided 2 main arguments for why these trends cannot be ethically reconciled. First, failing to conduct a priori power analyses fails to respect participants' decision-making autonomy by limiting the information disclosed during the informed consent process. Second, proceeding with underpowered trials, in the absence of explicit plans for definitive studies in the future, shifts the risk-benefit calculus that helps justify research in an unfavorable direction.[14] Participants in such trials experience personal risks and benefits commensurate with those in adequately powered trials, but are denied the same opportunity to contribute to the improved care of future patients. Therefore, IRB members should carefully monitor the statements made in the consent forms regarding the potential benefits of participation to ensure that these statements accurately reflect the strength of the underlying study design.

Low statistical power is merely one manifestation of a much larger problem: that many clinical investigators are not properly trained in research methods. The consequences are not only that investigators fail to properly assess the required sample size. Poor training may also explain why investigators may improperly assess the state of knowledge before initiating new studies, fail to appreciate how new trials ought to be conducted to advance this knowledge, choose inappropriate end points, and poorly report the results of their work.

We have focused our discussion on power because it remains one prominent problem, both scientifically and ethically, for which a workable solution is possible. We recommend that investigators always conduct a priori power calculations and relay the results to potential study participants. This should not be an overwhelming task. Simplified statements regarding both the inherent uncertainty in all research and whether the relative level of uncertainty in the proposed study conforms to standards of clinical investigation should be understandable by potential participants.

After conveying information in this way, the research must still meet one of the following conditions: either enough patients will be enrolled to obtain at least 80% power to detect a clinically important effect or, if this is not possible, the researchers will be able to document a clear and practical plan to integrate the results of their trial with those of future trials. Absent one of these 2 circumstances, ethics review boards, research funding agencies, and medical journal editors should maintain strict requirements for adequate research methods, including appropriate statistical power, for any clinical trial to be approved, funded, or published, respectively.

Author Contributions: *Study concept and design:* Halpern, Karlawish, Berlin.
Drafting of the manuscript: Halpern.
Critical revision of the manuscript for important intellectual content: Halpern, Karlawish, Berlin.
Statistical expertise: Berlin.
Study supervision: Karlawish, Berlin.
Funding/Support: Mr Halpern is supported by a predoctoral fellowship from the American Heart Association, Dallas, Tex, and a National Research Service Award in Cardiopulmonary Epidemiology from the National Heart, Lung, and Blood Institute, Bethesda, Md. Dr Karlawish is supported by a Brookdale National Fellowship, a National Institute on Aging Mentored Clinical Scientist Development Award, and a Paul Beeson Fellowship.
Acknowledgment: We thank Jon F. Merz, JD, PhD, and Kathleen Joy Propert, ScD, for their insightful comments on an early version of the manuscript.

REFERENCES

1. Newell DJ. Type II errors and ethics. *BMJ.* 1978; 4:1789.
2. Altman DG. Statistics and ethics in medical research III: how large a sample? *BMJ.* 1980;281:1336-1338.
3. Rutstein DD. The ethical design of human experiments. In: Freund PA, ed. *Experimentation With Human Subjects*. New York, NY: George Braziller; 1970: 383-401.
4. Freedman B. Scientific value and validity as ethical requirements for research: a proposed explication. *IRB Rev Hum Subjects Res.* 1987;9:7-10.
5. 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;299:690-694.
6. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA.* 1994;272:122-124.
7. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med.* 1982;306:1332-1337.
8. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Engl J Med.* 1987;317:426-432.
9. Altman DG, Dore CJ. Randomization and baseline comparisons in clinical trials. *Lancet.* 1990;335: 149-153.
10. Schumm LP, Fisher JS, Thisted RA, Olak J. Clinical trials in general surgical journals: are methods better reported? *Surgery.* 1999;125:41-45.
11. Nichol MB, Venturini F, Sung JC. A critical evaluation of the methodology of the literature on medication compliance. *Ann Pharmacother.* 1999;33:531-535.
12. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. *J Bone Joint Surg Am.* 1999;81:1454-1460.
13. 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; 320:1308-1311.
14. Emmanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA.* 2000;283:2701-2711.
15. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. *Stat Methods Med Res.* 1993; 2:161-172.
16. Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. *Lancet.* 1997;350:804-807.
17. Knapp TR. The overemphasis on power analysis. *Nursing Res.* 1996;45:379-381.
18. Farrar JT, Portenoy RK, Berlin JA, Kinman J, Strom BL. Defining the clinically important difference in pain outcome measures. *Pain.* 2000;88:287-294.
19. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.

20. Braunholtz DA, Edwards SJL, Lilford RJ. Are randomized clinical trials good for us (in the short term)? evidence for a "trial effect." *J Clin Epidemiol.* 2001; 54:217-224.
21. Simes JR. Prospective meta-analysis of cholesterol-lowering studies: the Prospective Pravastatin Pooling (PPP) Project and the Cholesterol Treatment Trialists (CTT) Collaboration. *Am J Cardiol.* 1995;76:122C-126C.
22. Bailar JC. The promise and problems of meta-analysis. *N Engl J Med.* 1997;337:559-561.
23. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc A.* 1988; 151:419-463.
24. Egger M, Smith GD. Meta-analysis: bias in location and selection of studies. *BMJ.* 1998;316:61-66.
25. Matthews JNS. Small clinical trials: are they all bad? *Stat Med.* 1995;14:115-126.
26. Current Controlled Trials Ltd. Current Controlled Trials. Available at: http://www.controlledtrials.com. Accessibility verified October 29, 2001.
27. National Institutes of Health. ClinicalTrials.gov. Available at: http://clinicaltrials.gov. Accessibility verified October 29, 2001.
28. Horton R, Smith R. Time to register randomised trials. *BMJ.* 1999;319:865-866.
29. Cassileth BR, Lusk EJ, Miller DS, Hurwitz S. Attitudes toward clinical trials among patients and the public. *JAMA.* 1982;248:968-970.
30. Mattson ME, Curb JD, McArdle R, and the AMIS and BHAT Research Groups. Participation in a clinical trial: the patients' point of view. *Control Clin Trials.* 1985;6:156-167.
31. Schron EB, Wassertheil-Smoller S, Pressel S, for the SHEP Cooperative Research Group. Clinical trial participant satisfaction: survey of SHEP enrollees. *J Am Geriatr Soc.* 1997;45:934-938.
32. Koblin BA, Heagerty P, Sheon A, et al. Readiness of high-risk populations in the HIV Network for Prevention Trials to participate in HIV vaccine efficacy trials in the United States. *AIDS.* 1998;12: 785-793.
33. Sugarman J, Kass NE, Goodman SN, Perentesis P, Fernandes P, Faden RR. What patients say about medical research. *IRB Rev Hum Subjects Res.* 1998; 20:1-7.
34. Welton AJ, Vickers MR, Cooper JA, Meade TW, Marteau TM. Is recruitment more difficult with a placebo arm in randomised controlled trials? a quasirandomised, interview based study. *BMJ.* 1999;318: 1114-1117.
35. Karlawish JHT, Casarett D, Klocinski J, Sankar P. How do Alzheimer's disease patients and their caregivers decide whether to enroll in a clinical trial? *Neurology.* 2001;56:789-792.
36. Altman DG. The scandal of poor medical research. *BMJ.* 1994;308:283-284.
37. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. *J Clin Epidemiol.* 2001;54:109-110.
38. Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. *Control Clin Trials.* 1987;8(suppl 4):6S-30S.
39. Meinert CL. Patient recruitment and enrollment. In: *Clinical Trials: Design, Conduct, and Analysis*. New York, NY: Oxford University Press; 1986:149-158.
40. Taylor KM, Margolese RG, Soskolne CL. Physicians' reasons for not entering eligible patients in a randomized clinical trial of adjuvant surgery for breast cancer. *N Engl J Med.* 1984;310:1363-1367.
41. Feagan BG, Fedorak RN, Irvine EJ, et al. A comparison of methotrexate with placebo for the maintenance of remission in Crohn's disease. *N Engl J Med.* 2000;342:1627-1632.
42. Halpern SD, Metzger DS, Berlin JA, Ubel PA. Who will enroll? predicting participation in a phase II AIDS vaccine trial. *J Acquir Immune Defic Syndr.* 2001;27: 281-288.
43. Freedman B. Suspended judgement: AIDS and the ethics of clinical trials: learning the right lessons. *Control Clin Trials.* 1992;13:1-5.
44. 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;121:200-206.
45. Berlin JA, Colditz GA. The role of meta-analysis in the regulatory process for foods, drugs, and devices. *JAMA.* 1999;281:830-834.
46. Department of Health and Human Services. Common rule (45 CFR §46). Federal policy for the protection of human subjects; notices and rules. 1991:28003-28032.

> The scientific attitude implies . . . the postulate of objectivity—that is to say, the fundamental postulate that there is no plan; that there is no intention in the universe.
>
> —Jacques Monod (1910-1976)