Caution Regarding the Use of Pilot Studies to Guide Power Calculations for Study Proposals

Helena Chmura Kraemer, PhD; Jim Mintz, PhD; Art Noda, MS; Jared Tinklenberg, MD; Jerome A. Yesavage, MD

Clinical researchers often propose (or review committees demand) pilot studies to determine whether a study is worth performing and to guide power calculations. The most likely outcomes are that (1) studies worth performing are aborted and (2) studies that are not aborted are underpowered. There are many excellent reasons for performing pilot studies. The argument herein is not meant to discourage clinical researchers from performing pilot studies (or review committees from requiring them) but simply to caution against their use for the objective of guiding power calculations. *Arch Gen Psychiatry. 2006;63:484-489*

In all areas of medical research, null hypothesis significance testing (NHST) has long been the basis of drawing inferences from a sample to a population. Although much attention has recently been paid to the misuse of NHST,[1-7] efforts to bypass NHST have generally been rejected in favor of more careful attention to the proper use of NHST.[8]

Even when NHST is appropriately used, studies based on these methods are often underpowered (ie, the results are statistically nonsignificant not because the hypothesis being tested is untrue or is clinically nonsignificant but simply because the sample size is too small). As a result, terms such as *borderline significant*, *marginally significant*, and *trend toward significance* have come into common use, indicating that researchers were unable to reject the null hypothesis at their preset significance level (usually 5%) but that they still believe that their hypothesis is true. When manuscripts with nonsignificant results are submitted for peer review, reviewers and editors often ask for post hoc power calculations (ie, power calculations based on obtained results rather than on an a priori hypothesis). According to Tukey, power calculations, always "of vital importance before the experiment, are essentially meaningless once the experiment has been done."[9(p281)] Nevertheless, despite nonsignificant results, reviewers and editors may seek some indication whether a hypothesis still may be true.[10-13]

The appropriate use of NHST methods requires that an a priori hypothesis be tested based on the rationale and justification from theory, clinical experience, and evidence from previous studies. The null hypothesis is the denial of this hypothesis. A study is proposed (ie, the design, treatment, and measurement protocols), and an analytic plan is developed to test the null hypothesis. For a valid α-level test of significance, the analytic plan must guarantee that, whenever the null hypothesis is true, the probability of rejecting the null hypothesis (thus providing support to the hypothesis of interest) is less than α. By general consensus, $\alpha = .05$. In recent years, study proposals are also required to show that the study plan has at least a certain power (by growing consensus, 80%) to reject the null hypothesis if the hypothesis is not only true but also true to a degree that would imply clinical or practical significance.

The problem lies in the phrase "true to a degree that would imply clinical or practical significance." Biostatisticians have long faced the difficulty of trying to elicit from clinical researchers what the threshold of clinical significance is in a particu-

Author Affiliations: Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford (Drs Kraemer, Tinklenberg, and Yesavage and Mr Noda), Department of Psychiatry, University of California, Los Angeles (Dr Mintz), and Veterans Affairs Palo Alto Health Care System, Palo Alto (Drs Tinklenberg and Yesavage), Calif.

lar context for the objective of power calculations. In part, the difficulty is that the effect sizes that biostatisticians use to perform power calculations are designed to satisfy computational needs and are often unsuited for clinicians to interpret in terms of clinical effect in a specific situation.

Yet, the same clinicians who find difficulty in specifying a threshold of clinical significance a priori often have no difficulty in doing so after the fact, examining data from a study and declaring a post hoc result to be of clinical significance or not. Clinicians sometimes consider a statistically significant result from a study and comment that it is too small to be of clinical importance. More often, they express continued interest in a statistically nonsignificant result because the result still seems to be of possible clinical importance. Because of this impasse between a priori and post hoc perceptions of clinical significance, 3 nonideal approaches to power calculations have evolved. The first strategy is seldom recommended but is common; the second strategy is not ideal but is acceptable; and the third strategy presents the serious problem that is the subject herein. For simplicity, the discussion that follows is set in the context of a randomized clinical trial (RCT) with 2 groups of equal size. The issues, however, are generalizable to more complex NHST.

In the first strategy, a convenience sample size is set, and whatever power results is simply accepted. This strategy often leads to underpowered studies, and review committees often appropriately question this approach.

In the second strategy, biostatisticians set effect size levels for the power calculations based on their experience in various RCTs, sometimes without consideration of the specific clinical context of the proposed RCT. For example, in RCTs with normally distributed outcome measures with equal variance in the treatment groups, biostatisticians need the effect size expressed in terms of the standardized mean difference between the treatment and control groups (Cohen *d*). Cohen[14] proposed that *d* values of 0.2, 0.5, and 0.8 be considered small, medium, and large, respectively. However, after a study is completed, clinicians often find that an effect size that biostatisticians label as small may in certain circumstances be clinically significant and an effect size that they label as large may in other circumstances be clinically trivial. Indeed, Cohen warned of this,[14(p12)] pointing out that the interpretation of the effect sizes depends on the substantive context, cautioning against reification of his conventions, and inviting researchers "not to employ them if possible."[14(p534)] This strategy is far from the ideal solution to the problem.

In the third strategy, clinical researchers propose (or review committees require) that a small pilot study be performed to estimate the effect size. The researchers or the reviewers then examine the pilot study effect size and make a post hoc decision whether that observed effect size promises to be clinically significant. If it does not, the study would not be proposed at all or, if proposed, would not be funded (hence aborted). On the other hand, if the pilot study results suggest clinical significance, power calculations for the main study are then based on that obtained pilot study effect size. Because the pilot study effect size is an inaccurate estimate of the true effect size, there are 2 results of this strategy: (1) Using the inaccurate pilot study effect size, the true effect size will often be understated and the main study aborted, even when the true effect size is clinically significant. (2) The inaccurate pilot study effect size that justifies the main study may overestimate the true effect size, underestimating the sample size for the main study and underpowering the study. As a result, using the third strategy, even clinically significant effects are likely to be found statistically nonsignificant.

Some researchers understand these results intuitively. However, the fact that review committees continue to require such pilot studies before considering funding a proposal indicates that many do not.

To convince doubters and, most important, to provide some indication of the magnitude of the errors, our discussion centers on one example that will easily illustrate our thesis. The full mathematical formulas on which these results are based can be obtained from the authors and can be used to expand on the illustration. The results are analogous in other situations, although the computations may be much more complex.

RCT ILLUSTRATION: THE IDEAL APPROACH

The illustration is an RCT in which the hypothesis is that a new treatment (*T*) is more effective in treating a disorder than a control or comparison treatment (*C*). In the proposed study, subjects (N) are assigned with equal probability to the *T* and *C* groups (N/2 per group); for ease of computation, the outcome measure is assumed to be normally distributed in both groups with equal known variances.

The effect size for which biostatisticians need to compute power is δ, the difference between the treatment and control group means divided by their common standard deviation. The null hypothesis here is that δ is less than 0 (a 1-tailed test). What is needed is a critical value of δ (eg, $\delta^* > 0$) that clinicians would accept as the threshold of clinical significance. Therefore, if the true value of δ is between 0 and δ^*, a large enough sample size might well obtain statistical significance, but clinicians would not consider it to be of clinical significance. If the true value of δ is above δ^*, they would consider it to be of clinical significance, with the larger the δ, the greater the clinical significance. If the power is accurately set, the probability that a clinically significant result is also statistically significant would never be below 80%.

For example, suppose a treatment for patients with Alzheimer disease is to be compared with placebo treatment, and the outcome measure is the change in the Mini-Mental State Examination (MMSE)[15] score at the end of 1 year of treatment. The MMSE has integer scores ranging from 0 to 30, with lower values indicating loss of cognitive ability. A mean difference between *T* and *C* of 0.01 point on the MMSE would not be a convincing result for clinicians and consumers, particularly if the treatment is costly and has adverse effects. Therefore, a value of δ corresponding to a change of 0.01 point would be well below the critical level of δ^*. At the other extreme, suppose the mean difference between *T* and *C* was 10 points on the MMSE scale. Because the mean decrease in MMSE

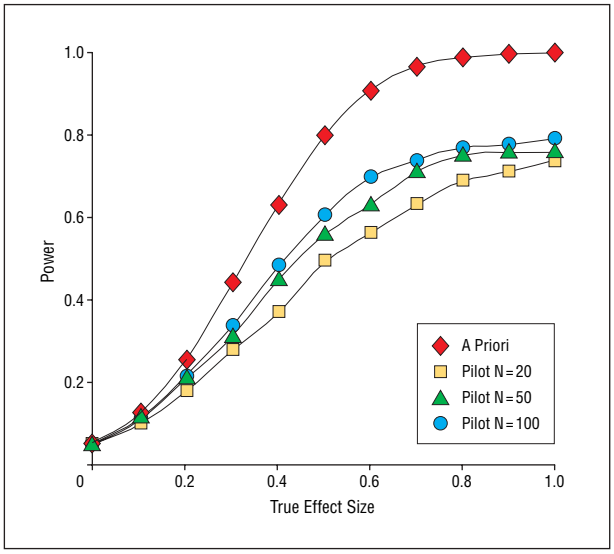

Figure 1. The power (probability of finding a statistically significant result) using a 1-tailed 5% test when the desired power to detect the effect size defining the threshold of clinical significance (δ*) is 0.5. The upper curve shows results using a priori power calculations. The 3 lower curves show results when the power calculation is based on effect sizes from pilot studies with sample sizes of 20, 50, and 100.

per year in patients with Alzheimer disease is only about 3 points,[16] such a difference would be large, and δ corresponding to a change of 10 points would be well above δ*. Between 0.01 and 10 points on the MMSE, there is a value of δ equaling δ* at which about half the clinicians and consumers would consider the result important and the other half would not, or the individual clinician or consumer might switch from considering the effect too small to considering it large enough to prefer *T* to *C*. That value is δ*.

The 5% level test would require that we compute the *z* statistic as follows: $z = N^{1/2}(M_T - M_C)/(2\sigma)$, where M_T and M_C are the 2 sample means and σ^2 is the known variance. We propose to reject the null hypothesis (ie, report the results as being statistically significant and support the effectiveness of *T* over *C*) if *z* exceeds 1.645, the 5% 1-tailed critical point of the standard normal distribution. If we knew δ* a priori, it would be easy to compute the N* that is required to ensure that the power to detect an effect size above the threshold of clinical significance (δ*) is greater than 80%, using the following equation: $N^* = 4(1.645 + 0.842)^2/\delta^{*2}$.[14,17] For example, for a 5% 1-tailed significance level and 80% power to detect a δ* threshold of clinical significance of 0.5, one would need 50 subjects per treatment group, for a total of 100 subjects.

In **Figure 1**, the upper (a priori) curve demonstrates the results of the performance of this test. The probability of finding a statistically significant result is indicated on the vertical axis, while the true effect size is indicated on the horizontal axis. Whenever the effect size is 0 or less (ie, the null hypothesis is true), the probability of finding a statistically significant result is always less than 5%, as required. When the effect size is greater than δ* (0.5 in our illustration), the probability of finding a statistically significant result is always greater than 80%, as required. Between the values of 0 and δ* is a range of true effect sizes that are too small to be clinically significant but where the probability of a significant result rapidly increases from the set 5% significance level to the 80% stipulated power. Therefore, it is possible that a clinically nonsignificant result might be found to be statistically significant, but it is unlikely that a clinically significant result would be found to be statistically nonsignificant. However, that depends on the ability to set δ* a priori.

RCT ILLUSTRATION: USING A PILOT STUDY

Ignoring for the moment the a priori method to select δ*, suppose for example that a review committee required a pilot study with N subjects to estimate the effect size to be used in power calculations for the main study. In the absence of funding, the sample size for the pilot study would typically be small (eg, N = 20). For the purpose of example, sample sizes of 50 and 100 will also be used. In the ideal situation, 100 would have been the sample size for the main study. The true effect size δ would be estimated by *d*, the pilot study effect size, where *d* has a normal distribution with a mean of δ and an SE of $2/N^{1/2}$.

The standard error of the estimated effect size (*d*) is the crux of the problem. For example, for a sample size of 20, the SE of *d* would be 0.45; for a sample size of 50, the SE would be 0.28; and even for a sample size of 100, the SE would still be 0.20. Unless the sample size of the pilot study approaches what the sample size of the main study should be (N* = 100) based on the a priori knowledge of δ*, these standard errors are much too large to render the estimated effect size of any practical use.

Nevertheless, the data from the pilot study are to be examined by the researchers and/or the review committee, and a post hoc decision is to be made as to whether the true effect size is greater than the elusive δ*. If the pilot study effect size *d* was less than 0 or the δ* threshold of clinical significance (eg, because the resulting sample size to be proposed seemed unreasonable or unfeasible in the specific context), the proposal would not be submitted, or, if submitted, the review committee would not recommend it for funding. The main study would then be aborted. Yet, given the error of *d* in estimating δ, this may happen even for a δ value that is well above the clinical threshold.

The probability that the pilot study effect size *d* is below the δ* threshold may be calculated using the following equation: $P(d < \delta^*) = \Phi(N^{1/2}[\delta^* - \delta]/2)$, where $\Phi(x)$ is the cumulative standard normal distribution (normdist[*x*,0,1,true] in Excel [Microsoft, Redmond, Wash]). This is shown in **Figure 2** (δ* = 0.5 in our illustration) using pilot study sample sizes of 20, 50, and 100. When δ equals δ* (whatever the N), the chance of performing (ie, not aborting) the study is only 50-50. When δ is less than δ*, the probability of aborting the study is greater than 50% but (using a small pilot study sample size) is nowhere near 100%. When δ is greater than δ*, the probability of aborting the study is less than 50% but (using a small pilot study sample size) is nowhere near 0%. Therefore, because of estimation error in the pilot study effect size, there is a good chance that the study will go forward when the true effect size is be-

low the threshold of clinical significance, and there is a good chance that the study will be aborted when the true effect size is well above that threshold.

Worse yet, the only instance in which the pilot study effect size would actually be used to compute the power would be when the study is not aborted. Therefore, the pilot study effect sizes that underlie the power calculation will be biased upward. For the small sample sizes used in pilot studies, the expected value of d always exceeds the true effect sizes δ and δ^*. The curves in **Figure 3** (in which $\delta^*=0.5$) suggest how serious this bias might be for sample sizes of 20, 50, and 100.

Because the effect size is overestimated, the sample size N** for the main study (which was computed to achieve 80% power to detect the effect size estimated from the pilot study) will be underestimated and the resulting power attenuated. Figure 1 compares the results of this procedure using pilot study sample sizes of 20, 50, and 100 when δ^* is set a priori. As the pilot study sample size N becomes very large (which never happens in a pilot study), the power approaches not 100%, as one might expect, but 80%. The significance level of the resulting main study is unaffected, which remains at 5%. However, the power to detect any positive effect is attenuated, particularly for the effect sizes above the threshold of clinical significance.

Consequently, 2 likely results of using pilot study data as the basis for power computation are the following: (1) There is a possibility that the study proposal will be aborted even when the actual effect is clinically significant. (2) If the study proposal is not aborted, the sample size estimated on the basis of the pilot study effect size will be too small and will result in a study that is underpowered to detect the effect sizes of clinical significance. Studies based on this strategy are likely to end in failed RCTs, wasting research time and money.

COMMENT

The crucial issue herein is to exercise caution in basing sample size selection for an RCT on the results of a pilot study. Although a simplified example has been used to present quantitative results, the principles generalize to all statistical tests comparing T and C in RCTs (eg, t tests, 2×2 χ^2 tests, and Mann-Whitney tests). In all cases, the power and the evaluation of clinical significance depend on an effect size. In the typical small pilot study, the standard error of that effect size is very large. Consequently, there is a substantial probability of underestimation of the effect size, which could lead to inappropriately aborting the study proposal for an RCT. If the study is not aborted, there is a substantial probability of serious overestimation of the effect size, which would lead to an underpowered study and a failed RCT.

In this article, power was calculated using the threshold of clinical significance rather than the true effect size. However, caution should be exercised in interpreting results no matter what critical value is used in the power computation. Power is calculated using the threshold of clinical significance rather than the true effect size not for statistical reasons but for ethical ones. For the conduct of an ethical RCT, Freedman[18(p144)] states the following:

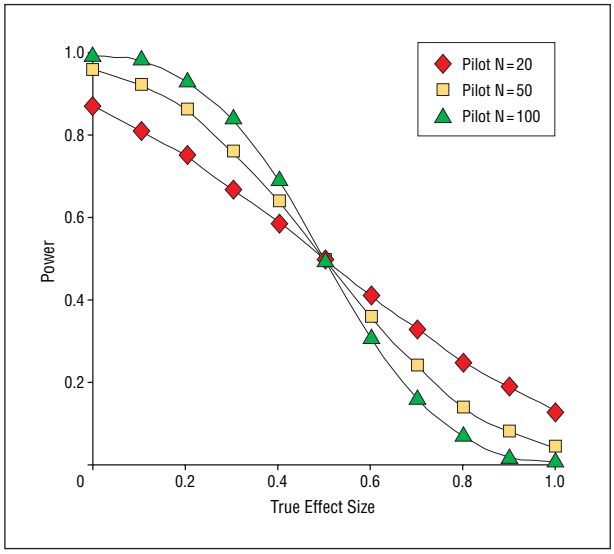

Figure 2. The probability of aborting a proposal using pilot studies with sample sizes of 20, 50, and 100 when the threshold effect size (δ^*) is 0.5.

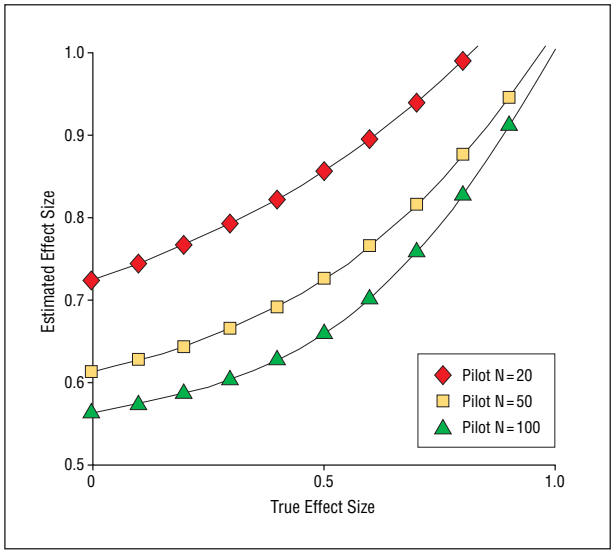

Figure 3. The estimated effect size if the study is not aborted, relative to the true effect size using pilot studies with sample sizes of 20, 50, and 100.

> [A]t the start of the trial there must be a state of clinical equipoise regarding the merits of the regimen to be tested, and the trial must be designed in such a way as to make it reasonable to expect that, if it is successfully concluded, clinical equipoise will be disturbed . . . [where] . . . clinical equipoise means a genuine uncertainty on the part of the expert medical community about the comparative therapeutic merits of each arm of a clinical trial.

At the time of the design of an RCT, the true effect size is unknown and cannot be used to define the critical value at which power is computed. Indeed, if the true effect size is known a priori with enough confidence to be used to calculate the necessary sample size, conducting an RCT is clinically unethical. What can be known with reasonable accuracy is the effect size that might "disturb" equipoise in the specific medical context (what we

call herein *the threshold of clinical significance*). This is a judgment call by the experts who are proposing the RCT, which is evaluated by the experts who are reviewing the RCT for funding as to what effect size in the specific context of the study might affect clinical decision making. These results depend on conflating the issues of aborting the study vs determining the sample size of an RCT that is proposed, funded, and executed.

Ignoring the possibility of aborting the study, the median sample size of 100 based on the pilot study effect size is approximately correct (50 subjects per group). However, the expected sample size is infinite because, using a 1-tailed test, if d is less than 0 (consistent with the null hypothesis), there is no finite sample size large enough to achieve 80% power. With a sample size of 20, the probability that d is less than 0 is 13%; with a sample size of 50, it is 4%; and only with a sample size of 100 is it less than 1%. It is impossible to propose a study if the necessary sample size to achieve adequate power is infinite!

Moreover, if one discounts the possibility of aborting the study, the probability of a gross overestimation of sample size (even if finite) is very large. In this case in which the a priori sample size was 100, with a pilot study sample size of 20 subjects, 22% of the time the estimated sample size will exceed 1000 subjects, 27% of the time it will exceed 500 subjects, and 37% of the time it will exceed 200 subjects. In short, there is a good chance that the proposed sample size will be orders of magnitude larger than the necessary sample size.

One of the ways in which review committees implicitly judge the threshold of clinical significance is by how feasible and reasonable the proposed sample size is, given the specific research question in the RCT. If researchers proposed a sample size of 10 000 in an RCT to test the effectiveness of a low-cost and safe vaccine to protect against Alzheimer disease, reviewers might well find such a proposal feasible and justified. However, if researchers proposed a sample size of 10 000 in an RCT to test a new antidepressant drug against placebo for mildly depressed patients, that proposal is unlikely to be seen as feasible or justifiable. What differentiates the 2 situations is an implicit awareness of different thresholds of clinical significance in those medical contexts. Consequently, the decision whether to submit a proposal and decisions related to review and funding of a proposal are not independent of the sample size, and the decision to abort a study is related to the decision about how large the sample size would be if the RCT were executed.

The full discussion of how to explicitly determine the threshold of clinical significance for a particular RCT is beyond the scope of this discussion, but there has been recent progress on this issue, particularly in defining clinically interpretable effect sizes.[19-25] Increasingly, the number needed to treat (NNT) (defined as the number of patients one would need to treat with T to expect 1 success more than if the same number had been treated with C[20,24,26]) appears to be an effect size that clinicians can interpret. As to setting thresholds, progress has been slower. For example, suppose that 40% of patients would fail with C but this could be reduced to 38% with T (NNT = 50 [1/{0.40–0.38}]) or to 35% (NNT = 20), 30% (NNT = 10), 20% (NNT = 5), or 0% (NNT = 2.5). How many patients should the ethical clinician be willing to treat with T to achieve 1 success more than would have been achieved with C? If C were a placebo, the answer might be different than if C were reasonably effective usual care or the standard of care. If T were high cost or high risk, the answer might be different than if T were low cost and low risk. If the population were particularly vulnerable (eg, infants or young children), the answer might be different than if the population were a more robust one. If failure meant death, the answer might be different than if failure meant a smaller reduction in symptoms or a slight delay in remission time. How to evaluate clinical significance under these various conditions is a question that statisticians cannot and should not answer. It is a question on which clinicians, consumers, and clinical researchers should attempt to reach consensus, not only to facilitate power computations in the planning of studies and to reduce the prevalence of failed RCTs but also to enhance interpretation of the results for clinical decision making after the research studies are completed.

Until that happens, the critical effect size at which the power is computed to be 80% should not be the effect size one aspires to obtain or expects to see; rather, it is the threshold below which clinicians are unlikely to be interested in the effect. The effect size one aspires to obtain or expects to see is usually well above this threshold value. The true effect size is unknown at the time an RCT is planned and may even be consistent with the null hypothesis. An effect size reported in a previously published study may serve as a discussion point about setting a threshold or defending it as clinically reasonable, but (given publication bias) the study results may have overestimated the threshold. Furthermore, because new studies differ in meaningful ways from published studies, prior data must be viewed cautiously. Most important, the effect sizes from a small pilot study should not be used to determine the sample size in the main study.

Research guidelines by statisticians may have unwittingly contributed to the present confusion about which effect size to use in power computations because most guidelines emphasize that the rationale and justification for a proposed hypothesis-testing study (including power calculations) are based on results "from previous research."[7] Some may even interpret this recommendation as endorsing a small pilot study to serve as previous research. That is not what is meant. The results of well-performed, adequately powered, previously published research studies serve as the basis of discussion for any proposed study. If existing research had already yielded a reliable estimate of the effect size sought in a study under consideration, there would be no rationale or justification for a new proposed research project with the objective of obtaining a reliable estimate of the same effect size. A small (typically inadequately powered and inadequately funded) pilot study does not fall under the rubric of previous research for the objective of power considerations.

Pilot studies are important in the preparation of proposals for hypothesis-testing studies. They serve to check on the availability of eligible and willing subjects using the recruitment methods proposed, to test the feasibil-

ity of the treatment and measurement protocols, to train researchers in study tasks, and to set up data collection, checking, storage, and retrieval capabilities. Glitches in the research design are often found and corrected during pilot testing, leading to a better-designed main study. However, every such correction casts doubt on whether any effect size estimate derived from a pilot study represents the true effect size in the main study.

Consequently, pilot studies cannot estimate the effect size with sufficient accuracy to serve as a basis of decision making as to whether a subsequent study should or should not be funded or as a basis of power computation for that study. The argument herein is not meant to discourage clinical researchers from performing pilot studies (or review committees from requiring them) but simply to caution against using them for the objective of power calculations.

Submitted for Publication: July 20, 2005; final revision received October 3, 2005; accepted October 5, 2005.
Correspondence: Helena Chmura Kraemer, PhD, Department of Psychiatry and Behavioral Sciences, Stanford University, 401 Quarry Rd, MC 5717, Stanford, CA 94305 (hck@stanford.edu).
Funding/Support: This study was supported by grant P30 AG17824 from the National Institutes of Health, by the Medical Research Service of the Veterans Affairs Palo Alto Health Care System, and by the Department of Veterans Affairs Sierra-Pacific Mental Illness Research, Education, and Clinical Center.

REFERENCES

1. Dar R, Serlin RC, Omer H. Misuse of statistical test in three decades of psychotherapy research. *J Consult Clin Psychol*. 1994;62:75-82.
2. Hunter JE. Needed: a ban on the significance test. *Psychol Sci*. 1997;8:3-7.
3. Krantz DH. The null hypothesis testing controversy in psychology. *J Am Stat Assoc*. 1999;44:1372-1381.
4. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods*. 2000;5:241-301.
5. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol Methods*. 1996;1:115-129.
6. Shrout PE. Should significance tests be banned? introduction to a special section exploring the pros and cons. *Psychol Sci*. 1997;8:1-2.
7. Wilkinson L; Task Force on Statistical Inference. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol*. 1999;54:594-604.
8. Wainer H. One cheer for null hypothesis significance testing. *Psychol Methods*. 1999;4:212-213.
9. Tukey JW. Tightening the clinical trial. *Control Clin Trials*. 1993;14:266-285.
10. Hoenig J, Heisey D. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55:19-24.
11. Korn EL. Projection from previous studies: a caution. *Control Clin Trials*. 1990;11:67-69.
12. Korn EL. Projecting power from a previous study: maximum likelihood estimation. *Am Stat*. 1990;44:290-292.
13. Levine M, Ensome M. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy*. 2001;21:405-409.
14. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence A Erlbaum Associates; 1988.
15. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12:189-198.
16. O'Hara R, Thompson JM, Kraemer HC, Fenn C, Taylor JL, Ross L, Yesavage JA, Bailey AM, Tinklenberg JR. Which Alzheimer's patients are at risk for rapid cognitive decline? *J Geriatr Psychiatry Neurol*. 2002;15:233-238.
17. Kraemer HC, Thiemann S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, Calif: Sage Publications; 1987.
18. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med*. 1987;317:141-145.
19. Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive nonparametric approach to measuring the size of treatment effects. *Stat Med*. 2006;25:591-602.
20. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310:452-454.
21. Grissom RJ. Probability of the superior outcome of one treatment over another. *J Appl Psychol*. 1994;79:314-316.
22. Grissom RJ, Kim JJ. *Effect Sizes for Research*. Mahwah, NJ: Lawrence A Erlbaum Associates; 2005.
23. Kraemer HC, Morgan GA, Leech NL, Gilner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. *J Am Acad Child Adolesc Psychiatry*. 2003;42:1524-1529.
24. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*. In press.
25. McGraw KO, Wong SP. A common language effect size statistic. *Psychol Bull*. 1992;111:361-365.
26. Altman DG, Andersen K. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ*. 1999;319:1492-1495.