RESEARCH DESIGN

Breaking Free of Sample Size Dogma to Perform Innovative Translational Research

Peter Bacchetti,[1]* Steven G. Deeks,[2] Joseph M. McCune[2]

Innovative clinical and translational research is often delayed or prevented by reviewers' expectations that any study performed in humans must be shown in advance to have high statistical power. This supposed requirement is not justifiable and is contradicted by the reality that increasing sample size produces diminishing marginal returns. Studies of new ideas often must start small (sometimes even with an *n* of 1) because of cost and feasibility concerns, and recent statistical work shows that small sample sizes for such research can produce more projected scientific value per dollar spent than larger sample sizes. Renouncing false dogma about sample size would remove a serious barrier to innovation and translation.

THE PROBLEM

Early studies of new ideas that have undergone little or no previous research, such as the first investigation in humans or nonhuman primates, is required to bring basic discoveries from the laboratory to the clinic. These studies may lack any preliminary data and, for practical reasons, are usually small. Unfortunately, grant reviewers and regulatory committees often downgrade or reject these proposals because they may be "underpowered" or have "inadequate" sample size. Such criticisms reflect the "threshold myth" (*1*)—an incorrect presumption that there is a sample size below which a study is doomed. In reality, small sample sizes can have scientific merit even if they do not meet conventional requirements for statistical power, and valid sample size choices can be made for cost or feasibility reasons alone (*2*).

Conventional power calculations provide precise sample sizes—but only by using precise assumptions. Accurately specifying such precise assumptions is challenging for any study and almost always impossible for early clinical studies. Because small differences in the assumptions produce large differences in the resulting calculated sample size (*1*), uncertainty about assumed values makes it difficult for the conventional approach to guide sample size choices. Furthermore, investigators cannot ignore cost and feasibility, conditions that are often the real determinants of sample size. Early translational studies tend to have a high per-subject cost (sometimes tens of thousands of dollars) and may put human subjects at considerable risk or require sacrifice of nonhuman primates, conditions that make large studies impractical and potentially unethical (*3*). Lastly, it is not possible to predict what will happen in a first-time intervention, so focus on any one specific hypothesis—as required for conventional power calculations—might not adequately reflect the potential value of the study.

BAD SOLUTIONS

Those entrusted with reviewing, approving, and funding early studies often expect standard power-based sample size justifications. Investigators and statisticians resort to a variety of strategies to meet this expectation for innovative studies, often manipulating calculations to justify a sample size that was really chosen for other reasons (*4*–*9*). Because every sample size produces 80% power under some set of assumptions, it is easy to provide a fabricated rationale for a sample size. For example, if costs and practical reasons dictate group sizes of no more than 10, a proposal might state:

> *In order to detect a difference in means of 3.2, assuming a standard deviation of 2.4, we calculated that 10 subjects per group would be needed.*

To be more honest, one might provide a calculation without actually saying how the sample size was chosen:

> *Assuming a standard deviation of 2.4, our proposed sample size of 10 per group will provide 80% power to detect a difference in means of 3.2 or greater.*

This type of statement, however, may still give the misleading impression that the calculation motivated the choice of sample size. Also, each of the above statements must provide specific values for the standard deviation (2.4) and difference in means (3.2), which are often completely unknown and just chosen to produce the desired result. Because such numbers can easily be challenged by reviewers, often with the fatal consequence that the study is not funded, generic "standardized effect sizes" may be invoked instead:

> *Our sample size of 10 per group will provide 80% power as long as the ratio of the difference in means to the standard deviation is 1.3 or greater.*

Such an abstract mathematical fact has no real connection to the particular study to be performed and is therefore no more relevant than would be the ritual recitation of any other abstract fact, such as "2 + 2 = 4."

Omitting any mention of how sample size was chosen or justified avoids dishonesty but invites reviewers to question whether a sample size is large enough. Labeling the study a "pilot" or "exploratory" may or may not deflect such criticism. Reviewers might ignore explicit (and accurate) assertions that uncertainty prevents any meaningful power calculation. Justifying a study as providing the information needed for future sample size calculations is sometimes reassuring to reviewers, but this justification is a misconception (*10*, *11*): If an initial study leaves substantial uncertainty about the primary issues to be investigated, then it will leave even more uncertainty about the assumptions needed for conventional sample size calculations. This uncertainty about assumptions is an inherent flaw in the conventional approach (*1*).

By enabling the conduct of innovative studies, the above tactics have served a useful purpose. We point out their flaws with some trepidation because we do not wish to make justifying sample size any harder than it already is. Nevertheless, trying to navigate an innovative study through the sample size minefield is currently haphazard, unpleasant, and often unsuccessful. The end result is a biomedical research enterprise that makes it difficult to translate novel findings from the laboratory into the clinic.

STARTING SMALL MAKES SENSE

There is little theoretical or empirical support for the conventional requirement that every study must have at least 80% power.

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143, USA. [2]Department of Medicine, University of California, San Francisco, CA 94143, USA.
*Corresponding author. E-mail: peter@biostat.ucsf.edu.

Cohen proposed using 80% power for sample size calculations in 1965 (*12*), offering this as "a conventional value … when no other basis is available" and adding, "Like all conventions, this value is arbitrary." More recently, even papers that explicitly condemn the conduct of trials with <80% power have acknowledged that this requirement is merely traditional (*13*, *14*), and challenges to the requirement (*1*–*3*) have not elicited any cogent responses.

Setting a goal that must be met regardless of cost or other practical constraints ignores the reality of diminishing marginal returns: each additional human or nonhuman subject adds less value than the previous one (*1*–*3*). The most cost-efficient study might therefore be one with a small sample size, producing more projected scientific value relative to cost than larger studies would. Although the projected value of a scientific study is difficult to precisely define, diminishing marginal returns are present for any reasonable definition, including statistical power (*2*). Diminishing returns might seem unsurprising and in accordance with "common sense," but the conventional approach contradicts this: It assumes that there is a substantial marginal increase in value upon reaching an "adequate" size, which is the threshold myth we noted previously (*1*).

A previous paper (*2*) showed with detailed mathematical derivations that diminishing marginal returns are especially pronounced for innovative early clinical studies and that this justifies the use of smaller sample sizes for such studies than for later, confirmatory studies. This justification can be understood less formally by considering three distinct possible outcomes for a study: a breakthrough, a complete failure with no reason to perform a follow-up study, and an indeterminate outcome (Fig. 1). When performing early studies, scientists always hope that the idea will be a breakthrough, producing new treatments or promising research directions. Although infrequent in history, breakthroughs—such as development of insulin therapy (*15*, *16*), development of the smallpox vaccine (*17*), and the first case of an HIV cure (*18*)—have helped make game-changing advances in the prevention and treatment of human disease. In these cases, initial studies were carried out in single subjects or families, and a larger n would not have been much more persuasive. When initial results turn out to be so promising, the resources that might have been consumed by a large initial study are often better directed toward follow-up studies that further develop the idea, such as fashioning or refining a practical intervention that can be widely used or examining a more heterogeneous population. Also, independent validation of promising initial results is essential. For the more common situation, in which the idea unfortunately turns out to be off-target, a larger sample size would also be wasteful. Although evidence that an idea is not viable has value, this information can often be ascertained from a smaller study.

In intermediate cases that are neither a breakthrough nor off-target, a larger initial study would produce added value. Nevertheless, a small initial study produces value by informing follow-up studies. A recent example of a small yet successful study is a report of intentional infection of a single human patient with an intestinal parasite as a treatment for ulcerative colitis (*19*). Although the sample size precluded any finding with a P value less than 0.05, this study was still interesting enough for publication in *Science Translational Medicine*. For a truly novel, first-time study, it is impossible to predict which of the three outcomes in Fig. 1 will emerge; accordingly, to avoid a sample size that could be wastefully excessive, a small initial sample size would seem to be a good place to start.

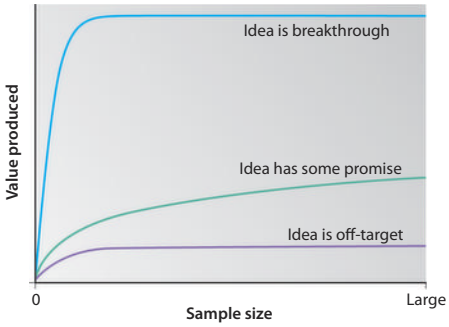

Fig. 1. How sample size influences the scientific or practical value that a study can be expected to produce, under three different scenarios. Numeric labels on the axes are absent because specific values will differ for different studies, but the qualitative relationship illustrated applies to all studies of new ideas. When the idea is a fundamental breakthrough, this will usually be apparent from a relatively small study. A much larger sample size does not produce much additional value. Similarly, when the idea is completely off-target, this will usually be apparent from a relatively small study. In intermediate cases, larger sample sizes might add value. However, increasing sample size still produces diminishing marginal returns.

BETTER SOLUTIONS

We recommend alternatives to the power-based approach for choosing and justifying sample sizes for early studies of new ideas. An established alternative in the statistical literature is the "value of information" approach (*4*, *20*, *21*), but this has rarely been used in practice. It calculates both the cost and value that can be expected over a range of sample sizes, choosing the one that maximizes value minus cost. The methods, however, may be too complex for frequent use with early studies.

A more recent proposal is to examine nine possibilities for the estimated effect and confidence interval that a study may produce (*1*). These possibilities result from three different possible effect sizes (the one that is expected or hoped for, no effect, and a possibility in between) and three levels of background uncertainty (best guess, high uncertainty, and low uncertainty). A proposal then discusses how valuable the study would be under each possibility.

Another recently developed approach focuses on costs and diminishing returns to select sample sizes that cannot be validly criticized as "inadequate" because they produce more projected value per dollar spent than any larger sample size (*2*). Although the projected value of a study is difficult to quantify, specific projections are not needed for these methods: They rely only on the fact that projected value has diminishing marginal returns, which holds true regardless of how projected value is quantified (*2*). For early studies, statistical arguments show that such a sample size is a choice called n_{root}. This is determined by examining the projected total cost of the study at different possible sample sizes and then choosing the sample size that minimizes the ratio of total cost to the square root of the sample size. A spreadsheet performing the calculations is available as a supplemental file for reference (*1*). The calculations are particularly simple if the total study cost is the sum of fixed costs, which do not depend on sample size, plus a set cost per subject. In such a linear cost case, n_{root} is equal to the total fixed study costs divided by the incremental cost per subject.

Using n_{root} to plan a sample size can appropriately adapt to the cost concerns of innovative studies. Higher incremental costs per subject reduce n_{root}, reflecting the practical reality that very high per-subject costs make smaller sample sizes more desirable. When costs are extreme, n_{root} may be

very small (even $n = 1$), but in such cases larger sample sizes will often be impractical. Reviewers and funders must then evaluate whether a small and expensive study is worthwhile.

Although we recommend using n_{root} for innovative studies, sample sizes larger than n_{root} might also be justifiable, especially when the idea is promising enough that the money spent to increase sample size above n_{root} produces added projected value that exceeds the added expense (*22*). Large studies may be practical for some innovative ideas, and we advocate more tolerance regarding sample size rather than a new form of knee-jerk intolerance directed at highly innovative studies with large sample sizes.

The preceding approaches are valid and worthy of consideration; nevertheless, investigators may simply choose a sample size that has worked well for similar past studies. This approach lacks a theoretical justification but appeals to common sense. Such choices will generally seem reasonable and be affordable, making them a suitable solution in many cases. These commonsense sample size determinations are familiar to most who carry out these types of studies, but they are difficult to justify to reviewers and regulators looking for an approach that seems more rigorous and objective.

PROPOSAL REVIEWING

Sample size has a continuous, gradual impact on a study's projected scientific value (Fig. 1), so we recommend that reviewers not focus on whether a proposed sample size is "adequate" or "valid." Criticism of sample size will rarely improve a study, unless the investigators have overlooked a method for efficiently increasing it. In all but extreme cases (such as study of an outcome that is so rare that none are likely to be observed), reviewers should instead assume that the sample size choice is acceptable.

As a hypothetical example, suppose promising results in an animal model with $n = 5$ lead to a proposal to investigate a novel method to reduce latent HIV in $n = 8$ human subjects. The investigators explain that this sample size is equal to n_{root} and that it represents a practical and cost-efficient choice: Exactly this many participants are willing, suitable, and already enrolled in an existing study that has funded infrastructure to offset most of the costs. In order to appease potential reviewers, the investigators might calculate—using the standard deviation (SD) from the animal study—that $n = 8$ will provide 80% power if the true rate of HIV reduction is at least 30% per year. (We believe that this calculation is essentially meaningless, because the n of 5 in the animal study leaves huge uncertainty about the SD, which in turn has a huge impact on the calculated sample size.)

Upon review of this proposal, a referee rejects it, arguing that it must be carried out with at least $n = 16$ in order to provide 80% power if the true rate of decrease is 22%. Because the referee feels that 22% would be large enough to be important and $n = 8$ only provides 50% power in this case, the sample size is "inadequate," and the study is fatally flawed. This type of criticism is based on the threshold myth and can prevent important research from happening. In reality, the fact of diminishing marginal returns implies mathematically that a study with $n = 8$ has more than half as much projected value as one with $n = 16$; at much less than half the cost, this is therefore an acceptable sample size. To justify $n = 8$ less mathematically, note that if the treatment is very successful (for example, >50% reduction per year, as shown in the top line of Fig. 1), this will be about as clear from studying $n = 8$ subjects as it would be from studying many more. Eight will also seem sufficient if the treatment does not work at all (as shown in the bottom line of Fig. 1).

ANALYSIS, INTERPRETATION, AND PUBLICATION OF EARLY STUDIES

Conventions for analysis and interpretation have grown out of a formal statistical hypothesis-testing paradigm that emphasizes distrust of any unplanned analyses or pursuit of unexpected issues, along with primary or exclusive focus on whether $P < 0.05$. This general approach has been criticized as unreliable, wasteful of information, harmful to scientific progress, and contrary to the original statistical theories that supposedly support it (*23*–*28*), but it remains engrained in research culture (*29*, *30*). As in the case of sample size, existing conventions are especially inappropriate for early, highly innovative studies, in which the potential for unexpected findings is high and where interpretation should focus on estimated effects and the uncertainty around them (as shown by confidence intervals) and whether the ensemble of results have a coherent and plausible biological explanation. We therefore recommend that investigators pursue all issues of potential interest regardless of whether they were prespecified, avoid formal adjustments to P values for "multiple comparisons" (*31*, *32*), and adjust methods and designs in response to early results when necessary. These actions all have some potential to increase the risk of a spurious finding with $P < 0.05$, adding to the need for cautious interpretation (*33*), along with full reporting of which results were unexpected, analyses that were undertaken but judged to be uninteresting, and any interim analyses and resulting adaptations made during the study. In addition, independent confirmation is particularly important for results of unplanned analyses, when many analyses were performed, or when interim adaptations were needed.

To prevent distortion of the scientific record, investigators should always disseminate results, even when they are disappointing or ambiguous. This is easier to overlook when a study is small and speculative, but it remains crucial. Selective dissemination will increase the apparent proportion of small, innovative studies in which striking results later prove to have been random flukes, creating an illusion that small sample size causes "false positive" results. The policy of some journals to accept any methodologically sound study can facilitate publication of small "negative" studies, as long as being small is not itself judged to be a methodological flaw. Also, the study registration system at ClinicalTrials.gov facilitates reporting of basic results (see http://clinicaltrials.gov/ct2/info/results). We believe that determined researchers can find a way to publish or otherwise disseminate results, but stronger incentives with less expensive and easier ways of doing this might be needed before investigators will approach the ideal of consistent dissemination of all studies' results.

CONCLUSIONS

As we have argued in this Perspective, the demand for conventional sample size calculations raises substantial barriers to the conduct and completion of innovative, bench-to-bedside translational research. Enforcement of the requirement for at least 80% power has no valid justification and is especially inappropriate for early translational studies investigating new ideas. Practical considerations, including the costs of larger versus smaller sample sizes, inevitably drive sample size choices, and the reality of diminishing marginal returns implies that this is scientifically valid. Indeed, the use of n_{root} determined only by costs is reasonable for innovative studies.

Acceptance of this and other alternatives to current conventions would remove a formidable barrier to the conduct of innovative translational research.

REFERENCES AND NOTES

1. P. Bacchetti, Current sample size conventions: Flaws, harms, and alternatives. *BMC Med.* **8**, 17 (2010).
2. P. Bacchetti, C. E. McCulloch, M. R. Segal, Simple, defensible sample sizes based on cost efficiency. *Biometrics* **64**, 577–585, discussion 586–594 (2008).
3. P. Bacchetti, L. E. Wolf, M. R. Segal, C. E. McCulloch, Ethics and sample size. *Am. J. Epidemiol.* **161**, 105–110 (2005).
4. A. S. Detsky, Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Stat. Med.* **9**, 173–184 (1990).
5. S. N. Goodman, J. A. Berlin, The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* **121**, 200–206 (1994).
6. G. H. Guyatt, E. J. Mills, D. Elbourne, In the era of systematic reviews, does the size of an individual trial still matter. *PLoS Med.* **5**, e4 (2008).
7. G. R. Norman, D. L. Streiner, *PDQ statistics* (B.C. Decker, Hamilton, Ontario, 2003).
8. K. F. Schulz, D. A. Grimes, Sample size calculations in randomised trials: Mandatory and mystical. *Lancet* **365**, 1348–1353 (2005).
9. S. Senn, *Statistical issues in drug development* (John Wiley & Sons, Chichester, England; Hoboken, NJ, 2007).
10. H. C. Kraemer, J. Mintz, A. Noda, J. Tinklenberg, J. A. Yesavage, Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch. Gen. Psychiatry* **63**, 484–489 (2006).
11. J. N. S. Matthews, Small clinical trials: Are they all bad? *Stat. Med.* **14**, 115–126 (1995).
12. J. Cohen, in *Handbook of Clinical Psychology*, B. B. Wolman, Ed. (McGraw-Hill, New York, 1965), pp. 95–121.
13. S. D. Halpern, Adding nails to the coffin of underpowered trials. *J. Rheumatol.* **32**, 2065–2066 (2005).
14. S. D. Halpern, J. H. T. Karlawish, J. A. Berlin, The continuing unethical conduct of underpowered clinical trials. *JAMA* **288**, 358–362 (2002).
15. F. G. Banting, C. H. Best, J. B. Collip, W. R. Campbell, A. A. Fletcher, Pancreatic extracts in the treatment of diabetes mellitus. *Can. Med. Assoc. J.* **12**, 141–146 (1922).
16. R. Madeb, L. G. Koniaris, S. I. Schwartz, The discovery of insulin: The Rochester, New York, connection. *Ann. Intern. Med.* **143**, 907–912 (2005).
17. C. P. Gross, K. A. Sepkowitz, The myth of the medical breakthrough: Smallpox, vaccination, and Jenner reconsidered. *Int. J. Infect. Dis.* **3**, 54–60 (1998).
18. K. Allers, G. Hütter, J. Hofmann, C. Loddenkemper, K. Rieger, E. Thiel, T. Schneider, Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation. *Blood* **117**, 2791–2799 (2011).
19. M. J. Broadhurst, J. M. Leung, V. Kashyap, J. M. McCune, U. Mahadevan, J. H. McKerrow, P. Loke, IL-22$^+$ CD4$^+$ T cells are associated with therapeutic *Trichuris trichiura* infection in an ulcerative colitis patient. *Sci. Transl. Med.* **2**, 60ra88 (2010).
20. A. R. Willan, Optimal sample size determinations from an industry perspective based on the expected value of information. *Clin. Trials* **5**, 587–594 (2008).
21. A. R. Willan, E. M. Pinto, The value of information and optimal clinical trial design. *Stat. Med.* **24**, 1791–1806 (2005).
22. P. Bacchetti, C. E. McCulloch, M. R. Segal, Simple, defensible sample sizes based on cost efficiency - Rejoinder. *Biometrics* **64**, 592–594 (2008).
23. J. S. Armstrong, Significance tests harm progress in forecasting. *Int. J. Forecast.* **23**, 321–327 (2007).
24. J. Cohen, The Earth is round (p <. 05). *Am. Psychol.* **49**, 997–1003 (1994).
25. M. J. Gardner, D. G. Altman, Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. *BMJ* **292**, 746–750 (1986).
26. S. N. Goodman, Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann. Intern. Med.* **130**, 995–1004 (1999).
27. G. Gigerenzer, Mindless statistics. *J. Socio-Economics* **33**, 587–606 (2004).
28. B. Lecoutre, J. Poitevineau, 2010. The significance test controversy and the Bayesian alternative. In StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies, http://statprob.com/encyclopedia/SignificanceTestControversyAndTheBayesianAlternative.html, accessed January 4, 2011.
29. L. C. Silva-Ayçaguer, P. Suárez-Gil, A. Fernández-Somoano, The null hypothesis significance test in health sciences research (1995-2006): Statistical analysis and interpretation. *BMC Med. Res. Methodol.* **10**, 44 (2010).
30. B. Lecoutre, M. P. Lecoutre, J. Poitevineau, Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *Int. Stat. Rev.* **69**, 399–417 (2001).
31. P. Bacchetti, Peer review of statistics in medical research: The other problem. *BMJ* **324**, 1271–1273 (2002).
32. K. J. Rothman, No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43–46 (1990).
33. J. P. A. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

34. **Funding:** Supported by NIH grant UL1 RR024131 and by the Harvey V. Berneking Living Trust. J.M.M. is a recipient of the NIH Director's Pioneer Award Program, part of the NIH Roadmap for Medical Research, through grant DPI OD00329. **Competing interests:** Contents are solely the responsibility of the authors and do not necessarily represent the official views of the funders. The authors declare no competing interests.

10.1126/scitranslmed.3001628

Citation: P. Bacchetti, S. G. Deeks, J. M. McCune, Breaking free of sample size dogma to perform innovative translational research. *Sci. Transl. Med.* **3**, 87ps24 (2011).

Science Translational Medicine

Breaking Free of Sample Size Dogma to Perform Innovative Translational Research

Peter Bacchetti, Steven G. Deeks and Joseph M. McCune

Sci Transl Med **3**, 87ps2487ps24.
DOI: 10.1126/scitranslmed.3001628