American Journal of
EPIDEMIOLOGY

Volume 161

Number 2

January 15, 2005

Sponsored by the Society for Epidemiologic Research
Published by Oxford University Press

SPECIAL ARTICLE

Ethics and Sample Size

Peter Bacchetti[1], Leslie E. Wolf[2], Mark R. Segal[1], and Charles E. McCulloch[1]

[1] Division of Biostatistics, Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, CA.
[2] Program in Medical Ethics, Department of Medicine, School of Medicine, University of California, San Francisco, CA.

Received for publication October 2, 2003; accepted for publication June 9, 2004.

The belief is widespread that studies are unethical if their sample size is not large enough to ensure adequate power. The authors examine how sample size influences the balance that determines the ethical acceptability of a study: the balance between the burdens that participants accept and the clinical or scientific value that a study can be expected to produce. The average projected burden per participant remains constant as the sample size increases, but the projected study value does not increase as rapidly as the sample size if it is assumed to be proportional to power or inversely proportional to confidence interval width. This implies that the value per participant declines as the sample size increases and that smaller studies therefore have more favorable ratios of projected value to participant burden. The ethical treatment of study participants therefore does not require consideration of whether study power is less than the conventional goal of 80% or 90%. Lower power does not make a study unethical. The analysis addresses only ethical acceptability, not optimality; large studies may be desirable for other than ethical reasons.

ethics committees; ethics, research; sample size

Editor's note: *An invited commentary on this Special Article appears on page 111, and the authors' response appears on page 113.*

Many investigators stand accused of conducting unethical research because their studies were "underpowered" (1–4). This is based on the idea that the projected scientific or clinical value of a study will be unacceptably low if it has low power, that is, if it has less than an 80 percent chance of producing $p < 0.05$ under an assumed minimum important effect size. It is therefore unethical to ask participants to accept the risks and discomforts of participation. Critics of this argument have noted that studies may be valuable—and therefore ethical—even if their results do not reach $p < 0.05$, specifically by producing useful estimates and confidence intervals or by contributing to meta-analyses (5–7). Although these observations may have had some impact (8), many remain unconvinced (4, 9). Here, we examine a more serious flaw in the argument: Even assuming the controversial premise that a study's projected value is determined only by its power, with no value from estimates, confidence intervals, or potential meta-analyses, the balance between a study's value and the burdens accepted by its participants does not improve as the sample size increases. Thus, the argument for ethical condemnation of small studies fails even on its own terms.

Correspondence to Dr. Peter Bacchetti, Box 0560, University of California, San Francisco, CA 94143-0560 (e-mail: peter@biostat.ucsf.edu).

THE ETHICAL BALANCE

For a study to be ethical in its design, its projected value must outweigh the projected risks to participants (10, 11). In many cases, the risks, inconvenience, and discomforts to be borne by participants outweigh the benefits they may *personally* receive as a direct result of participation in a study, so there is a projected net burden. This is the situation where small studies have been characterized as unethical. If there is no projected net burden, then any sample size is ethical, and sample size can be determined entirely by other considerations. When there is a net burden, a study may nevertheless be ethical if the projected benefit to society—the projected clinical or scientific value—outweighs the projected participant burdens, the risks are minimized and reasonable, and the participants make an informed decision to accept the burdens, despite the lack of direct personal benefit, to help produce this value. (After the study has concluded, participants may benefit from the knowledge gained, just like others who did not participate. Because this is not contingent on their personal participation, we consider this part of the societal benefit, rather than part of the determination of net burden or benefit from participating.)

The balance point between burden and value cannot be precisely calculated in most situations, because both the projected participant burdens and the study's projected value are difficult to quantify, particularly on comparable scales (11). We therefore examine here what can be reliably deduced about the influence of sample size on the ethical balance *in a way that does not depend on specific calculations or on the specific way in which burden or value is measured*. For simplicity, we use standard approximations, focusing on equal sample sizes in two groups to be compared, down to those that produce 10 percent power. (Extension to unequal or smaller sample sizes is possible but more complicated to present.) Although we discuss the standard goal of 80 percent power, our results do not depend on this particular value. We do not address optimality, only ethical acceptability. Optimization would require finding the sample size that maximizes study value minus total participant burden, but ethical acceptability is determined only by whether value exceeds burden, without regard to the amount of the excess. This makes the influence of sample size on ethical acceptability much easier to study.

If changes in the planned sample size do not change the composition of the study population (e.g., if a 1:1 randomization or a fixed 1:2 case:control ratio is planned), then the total participant burden will increase exactly in proportion to sample size. For example, a study 10 times larger will have 10 times the total burden that must be balanced by projected scientific or clinical value. This implies that the average projected net burden per participant remains constant regardless of sample size. Now consider how sample size influences the projected value per participant. One framework for evaluating projected value is the classical Neyman-Pearson statistical hypothesis-testing paradigm. Under the assumption that an important departure from the null hypothesis exists, a study may be regarded as valuable if it rejects the null hypothesis and (arguably (5–7)) as worthless if it does not. This leads to statistical power, the probability of rejecting the null hypothesis, as the measure of a study's projected value. This is the measure assumed by arguments for ethical condemnation of small studies. (We note that this is somewhat slanted in favor of larger sample sizes, because we cannot be sure that an important effect is in fact present, and larger sample sizes do not reduce the risk of wrongly rejecting the null hypothesis, which is typically fixed by design at $p = 0.05$ regardless of sample size.) Larger studies have higher power, but they also impose the net burden of participation on more subjects. A given sample size will be ethical if the study's projected value, here assumed to be its power at the minimum important effect, exceeds the total burden to be accepted by the participants, which is the sample size times the projected net burden per participant. Equivalently, the sample size is ethical if the power per participant exceeds the projected net burden per participant.

Figure 1 illustrates the ethical balance for comparing two groups by an unpaired t test assuming equal variances, with an arbitrarily chosen participant burden and the minimum important difference assumed to correspond to a standardized effect size of 0.25. Because the power per participant decreases as the sample size increases, we see that smaller sample sizes have a more favorable ethical balance than do larger ones. In this case, sample sizes up to about 130 per group are ethical because the study's projected value is greater than the participants' total burden. For larger sample sizes, the projected value is not sufficient to justify the participant burden. The burden shown is relatively large, so the maximum ethical sample size produces a study with only 53 percent power, and the usual goal of 80 percent power would require too much participant burden to be ethical. Different levels of participant burden will produce different cutoffs for the maximum ethical sample size, but the general conclusion that smaller studies have a better ethical balance than larger ones holds for *any* supposed level of burden. Figure 2 shows the general relation between a study's power and the power per participant (see Appendix). The key aspect of this curve is that it is decreasing, which implies that smaller studies have a more favorable ethical balance. Note also that nothing special happens around the conventional goal of 80 percent power. In particular, there is no sudden surge in value that would make studies with 80 percent power more ethical.

In the above illustration, we have adopted the questionable premise that a study's projected value is determined only by its power, which implies an exclusive focus on p values. This was to show that the case for ethical condemnation of studies with low power does not hold up even under its own assumptions. Many epidemiologists and statisticians, however, consider estimates and their confidence intervals to be a better basis for inference than p values (12, 13). For study results with $p > 0.05$ in particular, the strength of any negative conclusion is much better addressed by confidence intervals than by p values and pre- or post-hoc power calculations (14, 15). We therefore consider the implications of using an alternative measure of a study's projected value based on predicted confidence intervals (14). Wider confidence intervals indicate more uncertainty about the issue being studied, while narrower intervals indicate more certainty and more precise estimates. If we therefore assume that a study's

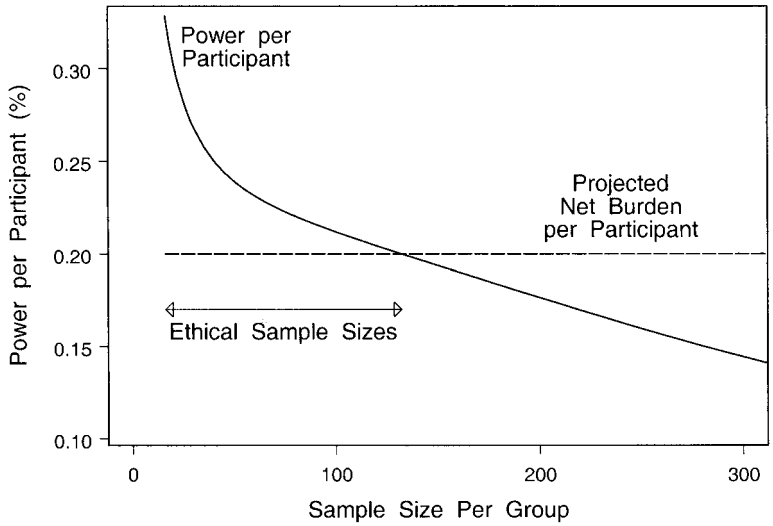

FIGURE 1. Power per participant versus sample size per group for an unpaired *t* test comparing two groups assuming equal variances, with a standardized effect size (difference in means divided by standard deviation) of 0.25. The curve is calculated using the standard formula based on a normal approximation (20) for sample sizes beginning at a minimum of 16 per group (total = 32), at which the study power is approximately 10%.

projected value is inversely proportional to the predicted width of the confidence interval, then we obtain the dashed curve shown in figure 2. Again, this relation shows decreasing value per participant as study power (or sample size) increases, so by this measure smaller studies again have a more favorable ethical balance. We show in the Appendix that two other definitions of projected study value based on different aspects of predicted confidence interval width also have decreasing value per participant as the sample size increases.

DISCUSSION

Despite the gravity of labeling research as unethical (16), the argument for condemning studies with less than the conventional goals of 80 percent or 90 percent power does not seem to have been subjected to detailed scrutiny. We show here that it is not valid under either its own assumption that power is the best estimate of a study's projected value or alternative methods based on the width of predicted confidence intervals. Indeed, we doubt that any reasonable

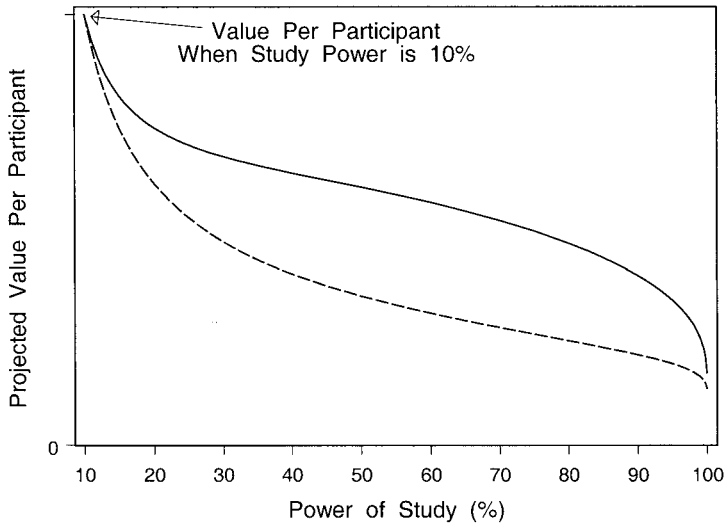

FIGURE 2. General relations between study power and projected value per participant, with a study's projected value assumed to be proportional to power (solid line) or inversely proportional to confidence interval width (dashed line). For visual clarity, curves are scaled to be equal when study power is 10%. The vertical scale differs for different particular situations, but the shapes shown remain constant. Because only the shapes are illustrated, intermediate tick marks with specific values for the vertical axis are not shown. See Appendix for derivation and further discussion.

measure of projected value can be constructed that would provide what the argument requires: the value *per participant* reaching a maximum at or above the sample size where 80 percent power is achieved. Mathematical statisticians are familiar with the central role that the *square root* of the sample size plays in determining widths of confidence intervals and other statistical properties. Because the square root exhibits diminishing marginal returns (an additional subject increases the square root less if the sample size is already large than if it is still small), we believe that any other sensible measure of projected study value will also fail to support the ethical condemnation of studies solely because they have less than 80 percent power. The argument presented here depends only on diminishing marginal returns, not the particular level of burden in a study or its importance, and this appears to be a reliable feature of any reasonable measure of a study's projected value (see Appendix).

Although this issue has been debated mainly in the context of randomized clinical trials, the concepts apply equally well to other research studies (including prevention trials, case-control studies, and cross-sectional studies) that impose some burden on participants, which most studies do. The measurement of risk factors and outcomes in epidemiologic studies typically requires from participants at least a commitment of time and some risk of loss of privacy, but it can also involve phlebotomy, radiologic procedures, insulin clamps, tissue biopsies, or many other burdensome procedures. If the potential direct benefit does not balance these burdens, there will be a net burden of participation. As in randomized clinical trials, this burden must be balanced by the study's projected scientific or clinical value.

Small studies may be more susceptible to two factors that could modify projected value: 1) the possibility that results will never be disseminated and 2) the possibility that results will be misinterpreted. These risks, however, are not inherent in the study itself and not inevitable; they can be prevented directly, by means other than increasing the sample size. If an investigator is firmly committed to publishing a study's results (or making them available by posting at a website or registry), then the study's projected value should not be discounted. Misinterpretation often results when large p values are thought to establish negative conclusions. This is not good reasoning for small or large studies and is readily prevented by a focus on confidence intervals rather than p values (12–15).

Two understandable—but indefensible—oversimplifications may have contributed to the widespread acceptance of the idea of unethically small studies. First is correctly noting that larger studies have more value than smaller ones, while failing to note that larger studies also impose the burdens of participation on more subjects, which makes the value *per participant* the more relevant quantity for ethical considerations. Second is equating 80 percent or greater power to a certainty of success and equating less than 80 percent power to a certainty of failure. Such dichotomization may be useful as a first approximation in some contexts, notably defining "statistical significance" as $p < 0.05$ and values "compatible" with observed data as those lying within a 95 percent confidence interval. Dichotomization, however, is not appropriate for addressing ethics and sample size (17), as is clear from the previous analysis of the ethical balance.

Although ethical treatment of participants does not require large studies, there are compelling reasons for conducting large studies in some cases. If there is no net burden or even a projected net benefit of participation, then ethical considerations may not constrain sample size. Another reason is when the issue to be studied may be very important, so that the scientific or clinical value is very large compared with the burden of participation. In such cases, large sample sizes may be desirable (18) and ethically acceptable, because the value per participant remains above the net burden even at large sample sizes. Finally, studies that are too small may have an unacceptably high cost per participant. In such cases, the high projected value per participant from a small study may nevertheless correspond to an unacceptably low projected value per dollar spent. (Note, however, that small studies performed with a low cost per participant are both cost efficient and ethical.)

Our analysis is limited to the ethical considerations in sample size selection when designing a study. We do not address the many other ethical considerations that investigators must attend to in designing and conducting a study, such as selection of the study population, recruitment, informed consent, and minimizing the risks and burdens of specific study interventions or procedures. In addition, we have not addressed the perspective of someone's deciding whether or not to participate as a research subject. Some have advocated that potential subjects be explicitly warned about the risk that a study with low power will miss an important effect (4). Such disclosures would be required for large studies, too, because even a study with 95 percent power has such a risk. Furthermore, potential subjects in very large studies would also need to be warned that, because they are just one among thousands, there is very little chance that their personal participation will make any difference in the study's outcome and consequent benefit to society. Given the complexity and controversial nature of this issue, we doubt that potential subjects would value such information when deciding whether to participate.

Finally, we reiterate that we have not addressed what sample size is best, only which qualitative ranges are ethically acceptable.

In conclusion, the analysis presented here suggests that the continuing conduct of "underpowered" studies is not the dreadful moral lapse lamented by some writers. In general, ethics committees and others concerned with the protection of research subjects need not consider whether a study is too small. In particular, we see no valid ethical argument against small, high-risk/high-payoff studies as have been recently advocated for rapidly fatal diseases (19). Indeed, a more legitimate ethical issue regarding sample size is whether it is too large.

REFERENCES

1. Newell DJ. Type II errors and ethics. (Letter). BMJ 1978;4:1789.

2. Altman DG. Statistics and ethics in medical research III: how large a sample? BMJ 1980;281:1336–8.
3. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA 2000;283:2701–11.
4. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288: 358–62.
5. Knapp TR. The overemphasis on power analysis. Nurs Res 1996;45:379–81.
6. Edwards SJL, Lilford RJ, Braunholtz D, et al. Why "underpowered" trials are not necessarily unethical. Lancet 1997;350:804–7.
7. Lilford R, Stevens AJ. Underpowered studies. Br J Surg 2002; 89:129–31.
8. Vail A. Experiences of a biostatistician on a U.K. research ethics committee. Stat Med 1998;17:2811–14.
9. Balasubramanian S. Underpowered studies. Br J Surg 2002;89: 811–12.
10. Department of Health and Human Services. Protection of human subjects (codified at 45 CFR §46.111(a)(2)). Washington, DC: Government Printing Office, 2002.
11. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: ethical principles and guidelines for the protection of human subjects of research. Washington, DC: Department of Health, Education, and Welfare, 1979. (DHEW publication no. (OS) 78-0012).
12. Rothman KJ. A show of confidence. N Engl J Med 1978;299: 1362–3.
13. Gardner MJ, Altman DG. Confidence intervals rather than *p*-values: estimation rather than hypothesis testing. BMJ 1986; 292:746–50.
14. 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;121:200–6.
15. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.
16. Hughes JR. The ethics of underpowered clinical trials. (Letter). JAMA 2002;288:2118.
17. Janosky JE. The ethics of underpowered clinical trials. (Letter). JAMA 2002;288:2118.
18. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23–40.
19. Horrobin DF. Are large clinical trials in rapidly lethal diseases usually unethical? Lancet 2003;361:685–7.
20. Hulley SB, Cummings SR, Browner WS, et al. Designing clinical research: an epidemiological approach. 2nd ed. Philadelphia, PA: Lippincott, Williams, and Wilkins, 2001:85.
21. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 1986;42:507–16.

APPENDIX

For comparison of two means, the sample size N needed to attain a specific type II error β and power $1 - \beta$ can be approximated by finding N to make the expected standard error of the estimated difference equal to $\Delta/(1.96 + Z_\beta)$, where Δ is the difference to be detected, Z_β is the point at which the standard normal cumulative distribution equals $1 - \beta$, and 1.96 is the Z value corresponding to a two-sided test at $p = 0.05$. The expected standard error is $2S/N^{1/2}$, where S is the standard deviation, so the needed N is approximately (20)

$$4S^2(1.96 + Z_\beta)^2/\Delta^2, \quad (1)$$

and the power per participant is therefore

$$0.25(1 - \beta)\ \Delta^2/(S^2(1.96 + Z_\beta)^2). \quad (2)$$

To study how this changes with changing β within a study, we divide the general expression 2 by its particular value for a referent level of power. For example, if power is 80 percent, then $\beta = 0.2$, and $Z_\beta = 0.84$. Substituting these in expression 2 produces $0.25\ (0.8)\ \Delta^2/(S^2(1.96 + 0.84)^2) = 0.025\ \Delta^2/S^2$. Dividing expression 2 by this particular value shows how the power per participant for general β compares with the power per participant when study power is 80 percent: $9.81(1 - \beta)/(1.96 + Z_\beta)^2$. This is illustrated by the solid curve in figure 2. Choosing any other referent level of power changes only the leading constant; the shape shown in figure 2 remains identical. Note that the values of S and Δ reflecting any particular situation have canceled out, leaving a general relation.

These general derivations also apply to other situations where the needed N can be approximated (14) by setting the standard error equal to $\Delta/(1.96 + Z_\beta)$, including comparison of rates by the chi-square test and comparison of survival data by the log-rank test (21). Precise calculations will differ from these general relations if the sample size is small enough that the t distribution should be used instead of normal approximations or if the standard error of the difference is not approximately equal under the null and alternative hypotheses. Nevertheless, we have verified for a variety of realistic situations that the more complex formulas needed in these cases still produce decreasing power per participant; we have not presented them in detail because they do not follow one common shape for all particular situations. We note that a study with only three per group cannot reach $p < 0.05$ by tests for comparing rates or by a nonparametric Mann-Whitney test for numerical data, so power is exactly zero. The analysis of power per participant given here therefore does not defend sample sizes below four per group.

The width of a confidence interval is proportional to the standard error, so a study value inversely proportional to confidence interval width is $N^{1/2}/S$, and the value per participant is then $N^{-1/2}/S$. Using expression 1 above for N, we obtain a value per participant proportional to $\Delta/(S^2(1.96 + Z_\beta))$. Dividing this by its value when power is 80 percent gives the expression $7.85/(1.96 + Z_\beta)$, which is shown by the dashed curve in figure 2 and is again a general relation. The marginal return from increasing from N to $N + 1$ under this definition is proportional to $N^{-1/2}$, which is a decreasing function in N and therefore implies diminishing marginal returns. Diminishing marginal returns imply a decreasing value per participant as the sample size increases. Alternative definitions could result from assuming that the marginal return is equal to the arithmetic or relative reduction in confidence interval width. Because the width is proportional to $N^{-1/2}$ and the reduction is equal to the opposite of the derivative of the width with respect to N, the arithmetic reduction is propor-

tional to $N^{-3/2}$. The relative reduction is the arithmetic reduction divided by the width and is therefore proportional to N^{-1}. Both $N^{-3/2}$ and N^{-1} are decreasing functions in N, so diminishing marginal returns remain under these alternative definitions.

A reviewer raised the possibility that some of the value of a study may lie in the potential identification of a rare but severe side effect. If the probability of this for any given subject is p, then the probability of seeing it one or more times in N subjects is $1 - (1 - p)^N$. Taking the derivative of this with respect to N, we see that the marginal increase in this probability from increasing from N to $N + 1$ subjects is $(1 - p)^N(-\log(1 - p))$, which is a decreasing function in N for any $0 < p < 1$. So this possible source of value also shows diminishing marginal returns.

The projected value of a study could perhaps be more realistically defined as a weighted sum of one or more of the above measures for different study outcomes or goals. Because all such components would exhibit decreasing value per participant as the sample size increases, such a sum would also.