Published in final edited form as:

Biometrics. 2008 June ; 64(2): 577–594. doi:10.1111/j.1541-0420.2008.01004_1.x.

Simple, Defensible Sample Sizes Based on Cost Efficiency

Peter Bacchetti, Charles E. McCulloch, and Mark R. Segal

Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560 USA

Summary

The conventional approach of choosing sample size to provide 80% or greater power ignores the cost implications of different sample size choices. Costs, however, are often impossible for investigators and funders to ignore in actual practice. Here, we propose and justify a new approach for choosing sample size based on cost efficiency, the ratio of a study's projected scientific and/or practical value to its total cost. By showing that a study's projected value exhibits diminishing marginal returns as a function of increasing sample size for a wide variety of definitions of study value, we are able to develop two simple choices that can be defended as more cost efficient than any larger sample size. The first is to choose the sample size that minimizes the average cost per subject. The second is to choose sample size to minimize total cost divided by the square root of sample size. This latter method is theoretically more justifiable for innovative studies, but also performs reasonably well and has some justification in other cases. For example, if projected study value is assumed to be proportional to power at a specific alternative and total cost is a linear function of sample size, then this approach is guaranteed either to produce more than 90% power or to be more cost efficient than any sample size that does. These methods are easy to implement, based on reliable inputs, and well justified, so they should be regarded as acceptable alternatives to current conventional approaches.

Keywords

Innovation; Peer review; Power; Research funding; Study design

1. Introduction

The conventional approach to choosing a sample size is to specify a goal and then calculate the sample size needed to reach that goal, with no explicit consideration of the cost implications. For health-related research involving human subjects, reviewers expect the goal to be 80% or 90% power, and some even regard use of this approach and this goal as necessary for a study to be ethical (Halpern, Karlawish, and Berlin., 2002; Horrobin, 2002). Despite these rigid expectations, we see no justification for ignoring costs. When studies compete for limited resources, with many studies going unfunded despite being judged to be excellent in a rigorous peer review process, cost efficiency seems important. We propose here new methods for choosing sample size that utilize cost information, and we explain why they should be considered acceptable approaches.

Peter Bacchetti (e-mail: peter@biostat.ucsf.edu), Charles E. McCulloch (e-mail: chuck@biostat.ucsf.edu), and Mark R. Segal (e-mail: mark@biostat.ucsf.edu), are Professors of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560, USA.

Supplementary Materials

Web Appendices and Tables referenced in Section 4 are available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

A number of fully Bayesian methods proposed for selecting sample size, usually known as maximization of expected utility (MEU) or value of information (VOI) methods, do take costs into account (Detsky, 1990; Claxton and Posnett, 1996; Bernardo, 1997; Lindley, 1997; Tan and Smith, 1998; Gittins and Pezeshk, 2000a,b; Halpern, Brown, and Hornberger, 2001; Walker, 2003; Willan and Pinto, 2005). These attempt to maximize the study's projected value minus its cost, which requires quantifying value and cost on the same scale, along with specifying priors that quantify uncertainty about the state of nature. Unfortunately, these more thoughtful methods have not been widely used (Yokota and Thompson, 2004). We propose here new and simple-to-implement approaches to sample size planning using costs. In contrast to the MEU/VOI approaches to date, these are based on cost efficiency, meaning the ratio of the study's projected scientific and/or practical value to its total costs—where MEU/VOI seeks to maximize a quantity analogous to “gross profit,” we focus on “return on investment.” Our proposed choices are more cost efficient than any larger sample size, which provides investigators using them with a ready defense against the common charge of “inadequate” sample size. If a reviewer or funder believes that a larger study would be worth doing, then a study that produces more value per unit cost must also be worthwhile. This high degree of defensibility is a crucial property that investigators need before they dare to depart from the current rigid conventions.

This article is organized as follows. We first provide a motivating example in Section 2, and then develop our methods in Section 3, along with some of their properties and conditions that justify them. In Section 4, we examine different measures of study value. We provide further examples in Section 5 and conclude with discussion in Section 6.

2. Motivating example: review of proposals

Because of the rigid expectations noted above, essentially every proposal seeking funding for clinical research includes a claim that its sample size will provide at least 80% power. Indeed, some have complained of a “sample size game” (Goodman and Berlin, 1994), in which investigators choose a sample size based on feasibility and cost considerations and then “invert the equation” (Detsky, 1990) to subsequently find assumptions that produce a calculation showing 80% power. Many reviewers guard against this practice by scrutinizing and criticizing the sample size justifications in grant proposals. We illustrate this dynamic with a simple example.

An investigator proposes to test a safe and inexpensive new treatment for a condition that resolves spontaneously in 40% of patients. As there is no existing treatment, she plans a randomized, double-blind, placebo-controlled trial. The proposal argues that if the biological theory underlying the new treatment is correct, then it should produce a cure in at least 1/3 of the patients who would not have had spontaneous resolution, an additional 20% or a difference of 40% versus 60%. A standard formula (Hulley, et al., 2001) shows that the proposed sample size of 97 per arm produces 80% power. Reviewer 1 finds this acceptable. Reviewer 2 feels that a rate of 60% is fairly likely if the biological theory is correct, but lower rates might be possible. He proposes that it would be safer to power the study for a difference of 40% versus 54%, requiring about a doubling of the proposed sample size. Reviewer 3 believes that the likely rate is irrelevant, because the study must be powered based on the minimum clinically significant difference, which all reviewers agree is about 10%. To obtain 80% power for this difference of 40% versus 50% would require about a quadrupling of the proposed sample size. Reviewer 3 further notes that anything less than 80% power for the minimum clinically significant difference would be unethical, citing a paper in a leading medical journal (Halpern, et al. 2002). This situation highlights a common difficulty in performing power calculations: what difference or effect size should be assumed? This assumption is very influential, because changes in the assumed difference are magnified in the resulting sample size, as noted above.

But precisely specifying the difference is difficult. Even the conceptual basis for picking the difference is unclear: should it be a difference considered to be likely to exist, or the smallest difference that would be clinically important?

We now consider a line of reasoning that could justify the investigator's proposed sample size while avoiding these difficulties. Despite the reviewers' concerns, an analysis of cost efficiency can show that if they agree that a larger study would be worth conducting, then so is the study as proposed. Suppose that the study as proposed costs \$200,000, but substantially larger sample sizes would cost more per patient due to the need to have multiple clinical sites with increased set-up, overhead, and coordination costs that are not completely offset by economies of scale: doubling the sample size would cost \$500,000, and quadrupling the sample size would cost \$1 million. In order to concretely illustrate the cost efficiencies involved, we need to compare these costs to a particular measure of the projected value that the study can be expected to produce. For illustrative purposes, we define a measure based on the predominant frequentist paradigm in which a 2-sided p-value < 0.05 in favor of the new treatment will lead to "rejection" of the null hypothesis that it is not effective and therefore to its adoption in practice. Based on how long it may be until other, better treatments are developed, suppose that an expected 100,000 future patients will be treated according to the result of the study. Also assume a 0.25 probability that the treatment is effective at the assumed alternative rates. (We can also assume a 0.25 probability that the treatment might prevent an equal proportion of spontaneous cures. This creates pre-study equipoise, but does not materially impact our calculations, because the chance of wrongly adopting the treatment when it is really harmful is very small for the sample sizes considered here.) The expected number of future cures is then calculated as 100,000 times the proportion of patients cured by treatment (the assumed resolution rate with treatment minus the background resolution rate of 40%) times the probability that the treatment is effective (assumed here to be 0.25) times the power. Under these assumptions, Table 1 shows the expected performance of the three sample sizes.

These calculations show that, with regard to expected clinical benefit, the smallest proposed sample size is the most cost efficient under *all* the assumed cure rates, despite having low power for some. For example, under Reviewer 3's assumption of 50% resolution with treatment, the advocated sample size of 388 per arm requires spending \$500 for each expected future cure produced, but the proposed sample size of 97 per arm obtains future cures at a cost of only \$279 each. *Clearly, if the value of a cure is enough to justify the larger study, then the smaller study is also acceptable.* We believe that reviewers should accept this conclusion, because a cure cannot be worth more than \$500 but less than \$279. Importantly, we note that cures cost more for the larger studies regardless of the assumed number of future patients or the assumed probability that the treatment is effective (because these only scale the expected number of cures up or down proportionally under all sample sizes). Furthermore, the larger studies would cost more per cure even if they could be done at the same cost per patient as the small one.

Although we focus above on a particular definition of projected value, similar results hold for *any* measure that exhibits diminishing marginal returns with increasing sample size, a property that holds for many measures of projected study value that have been proposed in connection with sample size planning, as we will establish in Section 4. But we first turn to developing a framework for exploiting this reliable general property.

3. A framework for choosing sample size

This section presents methods for identifying sample size choices that must be more cost efficient than any larger choices. These are potentially appealing because they can be defended against the frequent charge of being "inadequate"—if a larger study would be worth doing,

then so is one that is more cost efficient. Assume that all aspects of a study's design have been fixed and only sample size remains to be determined. Let

c_n = the cost of conducting the study if sample size is n

v_n = the expected scientific/clinical/practical value of the study if the sample size is n

The projected cost efficiency is then v_n/c_n .

The cost c_n consists of the resources that must be committed in order for the study to be completed. Some types of costs may be considered relevant from some perspectives but irrelevant from others. Our methods can be used with whatever definition of c_n is considered most meaningful to those who are choosing the sample size or those who must be convinced that it is an acceptable choice. The validity of our proposed methods therefore does not depend on our advocating or precisely defining any particular perspective on what costs should be considered or how they should be quantified. One important perspective is the broad societal perspective that would be relevant for considering studies to be funded by governments, which we believe would entail consideration of both financial costs and other aspects of the study's conduct such as risks and inconvenience to subjects. Narrower definitions could also be used; for example, some potential funders might consider only their own direct financial costs. Regardless of the perspective, we emphasize that c_n is the cost of the study itself and does not include costs that may be incurred later as a consequence of the study's results. For example, the cost of a treatment once it is adopted is not part of c_n , and neither is the cost of further research that is stimulated by the study. These would instead factor into its projected value.

Precisely quantifying v_n is difficult and can also depend on a particular perspective. In the previous section, we assumed that the projected value of a study was the expected number of additional cures produced, and we knew the parameters needed to calculate this. But in practical situations, an exact definition of study value will be less clear, and the knowledge needed to accurately project it is typically unavailable. Even in the simple case of expected cures, the projected value of the study would usually have to be modified to reflect the possibility of unanticipated side effects and how costly the treatment will be if it is adopted. Both would modify the net benefit of an efficacious treatment, and consequently modify the projected value of a study that might lead to its adoption. In addition, from a broad societal perspective, such as would be considered by reviewers of proposals for government funding, other possible benefits will often need to be considered. For example, much of the value of studies often lies in what they contribute to an accumulating body of knowledge, making a simple decision theoretic definition as in Section 2 inapplicable and leaving the difficult question of what some incremental information is worth. In particular, some studies may produce most of their value via the new ideas or further research that they stimulate, a potential source of value that seems particularly difficult to quantify in advance. These issues often make the projected value of a study very difficult to calculate. They also may help explain why the MEU/VOI methods mentioned in the Introduction have rarely been used in practice. We believe that such methods can work well if skilfully done, but they may appear too difficult and risky to investigators, because reviewers can easily disagree with the many specific, quantitative assumptions that are needed.

We therefore develop here methods that *completely avoid* the need to quantify projected value. These are easy to implement and rely only on study cost and on general properties of how sample size influences projected value. (We argue in Section 4 that the needed general properties can be relied on without specific verification for each study.) The idea is to use a simple stand-in function for v_n that increases with n at least as fast as any reasonable definition of v_n , and then to optimize costs relative to this function.

Proposition 1

If there is a positive function $f(n)$ and a value n^* such that

$$\frac{c_{n^*}}{f(n^*)} \leq \frac{c_n}{f(n)} \text{ for all } n > n^* \quad (\text{A})$$

and

$$\frac{v_{n^*}}{f(n^*)} \geq \frac{v_n}{f(n)} \text{ for all } n \geq n^*, \quad (\text{B})$$

then

$$\frac{v_{n^*}}{c_{n^*}} \geq \frac{v_n}{c_n} \text{ for all } n \geq n^*.$$

This follows immediately from multiplying the smaller terms from A and B together and the larger terms from A and B together, and then simplifying the resulting inequality. This leads directly to the following proposition, which is the basis for our proposed methods.

Proposition 2

Suppose $f(n)$ can be chosen so that condition (B) holds for any sample size under consideration. Then choosing n^* to minimize $c_n/f(n)$ selects the smallest sample size that meets both (A) and (B) and guarantees that the most cost efficient sample size is met or exceeded.

We propose two choices of $f(n)$ for implementing this strategy:

$$f(n)=n \quad (1)$$

and

$$f(n)=\sqrt{n}. \quad (2)$$

These lead to the following two sample size choices:

Definition 1

$n_{\min}$ is the smallest sample size that minimizes total study cost divided by sample size, i.e., the cost per subject.

Definition 2

n_{root} is the smallest sample size that minimizes total study cost divided by the square root of the sample size.

With choice (1), $n_{\min}$ is the smallest n^* that meets condition (A), and condition (B), which we denote as $B_{\min}$, follows if projected study value per subject is non-increasing beyond $n_{\min}$.

With choice (2), n_{root} is the minimum n^* meeting condition (A). In this case, condition (B), denoted B_{root} , is stronger than B_{min} , and $n_{\text{root}} \leq n_{\text{min}}$. In cases where B_{root} holds for $n^* = n_{\text{root}}$, it therefore follows that n_{root} will be a more cost efficient choice than n_{min} .

Our results do not depend on specifying the value of the study, but do depend on specifying the costs as a function of sample size and whether condition (B) holds. So it is important to understand how the most cost efficient sample size behaves as a function of the cost structure, and whether our easy-to-implement methods are likely to overshoot the most cost efficient sample size substantially. Some relevant results are given below. These are easily verified, so we provide a proof only for Proposition 7.

Proposition 3

Let n_{opt} denote the sample size maximizing $v_n c_n$. Replacing a given cost structure c_n with $c'_n = F + c_n$, for fixed costs $F > 0$ that do not depend on n , can only increase n_{min} , n_{root} , and n_{opt} , never decrease them.

Proposition 4

Replacing a given cost structure c_n with $c'_n = c_n + nc$, for per-subject costs $c > 0$, n can only decrease n_{min} , n_{root} , and n_{opt} , never increase them.

Proposition 5

If $c_n = F + nc$, then $n_{\text{root}} = F/c$ and total cost at n_{root} is $2F$.

Proposition 6

If $c_n = F + nc$, then n_{root} is at least half as cost efficient as any smaller sample size for any measure of value that is non-decreasing in n . This follows immediately from Proposition 5.

Proposition 7

If $c_n = F + nc$, then n_{root} is more than half as cost efficient as any larger sample size for any measure of value such that B_{min} holds at n_{root} .

The cost efficiency of n_{root} relative to any larger sample size n is $\{(F + nc)/(2F)\} \{v_{n_{\text{root}}}/v_n\}$.

Condition B_{min} implies $v_{n_{\text{root}}}/v_n \geq n_{\text{root}}/n$. Letting $r = n/n_{\text{root}}$ so that $nc = rF$, we see that the relative cost efficiency therefore is no smaller than $(1+r)/(2r) > 0.5$.

4. Influence of sample size on projected study value

We propose that use of n_{min} or n_{root} is reasonable in typical situations because conditions B_{min} and B_{root} hold under wide circumstances. In this section we consider many definitions of study value that have been proposed in connection with sample size planning, showing that condition B_{min} holds even at very small sample sizes, and B_{root} holds at small sample sizes for low-prior- information situations, when little is already known about the issue under study. We consider frequentist and Bayesian measures of value based on decision theory, interval estimation, information theory, and point estimation with squared error loss. Because of the predominant role of power and classical statistical hypothesis testing in current sample size planning conventions, we show some details for frequentist decision theory here (but with most mathematical detail deferred to Web Appendix A). We acknowledge that other measures may be regarded as more sensible; results for these are summarized in Section 4.4, with full details given in Web Appendix A. Figure 1 illustrates the shapes of the sample size-value relationships. These all show the concave shape that establishes condition B_{min} .

4.1 Decision theory using frequentist hypothesis tests

We start with a simple situation where both the state of reality—null or alternative—and the action to be taken as a result of the study are dichotomous, similar to a framework proposed by Lee and Zelen (2000). Obtaining a 2-sided p-value $< \alpha$ in favor of the alternative will result in “rejecting” the null in favor of the alternative, producing value k_1 if the alternative is true and value k_2 if the null is true. Otherwise, we “accept” the null, producing value k_3 if the null is true and value k_4 if the alternative is true. The constants k_1 to k_4 may be any values, but it is better to be right than wrong, so $k_1 \geq k_4$ and $k_3 \geq k_2$. With the type I error rate fixed at α for all n and power with a sample size of n denoted by P_n , the expected value following a two-sided statistical hypothesis test based on a sample size of n is then

$$k_1 \theta P_n + k_2 (1 - \theta) \alpha / 2 + k_3 (1 - \theta) (1 - \alpha / 2) + k_4 \theta (1 - P_n), \quad (3)$$

where θ is the prior probability that the alternative is true. The expected value of the study is (3) minus its value with 0 in place of P_n , because $n=0$ means that we accept the null with certainty.

This difference simplifies to:

$$v_n = \theta (k_1 - k_4) \left(P_n - \frac{\alpha}{2} \frac{1 - \theta}{\theta} \frac{k_3 - k_2}{k_1 - k_4} \right). \quad (4)$$

If the decision does not matter under the null, then $k_3 = k_2$ and value is proportional to power, similar to the situation in Section 2. If there is pre-study equipoise in the sense that actions corresponding to the null and alternative are equally desirable before the study because $k_1 \theta + k_2 (1 - \theta) = k_3 (1 - \theta) + k_4 \theta$, then $v_n = \theta (k_1 - k_4) (P_n - \alpha / 2)$ and value is nearly proportional to power for small α . We also have $v_n = \theta (k_1 - k_4) (P_n - \alpha / 2)$ if the action in the absence of a study is randomized so that $P_0 = \alpha / 2$. These results support the conventional focus on power as the relevant quantity for sample size planning for frequentist hypothesis testing. We therefore examine the consequences for our proposed methods of assuming that projected study value is proportional to power.

In Web Appendix A, we provide a proof of the following proposition for the simple case of a one-sample Z-test.

Proposition 8—For v_n proportional to power with the conventional choice of $\alpha=0.05$ most often used for sample size planning, $B_{\min}$ holds for all n^* regardless of the effect size used to calculate power.

For other situations, power calculations using normal approximations (the vast majority of power calculations done in practice) produce similar results. Bacchetti, et al. (2005) examined value proportional to power for the two-sample Z-test and discussed extensions to comparison of proportions and to log rank tests, confirming condition $B_{\min}$. Condition B_{root} , however, does not hold at small sample sizes.

4.2 Low prior information

Having only two possibilities might be characterized as a high-prior-information situation, because it assumes only one possible important departure from the null and that it has been correctly specified. In practice, there is always uncertainty about the size of any departure from the null that may exist. Although the concept of prior information is foreign to frequentist

hypothesis testing, we consider an example of a low-information situation in order to illustrate how this can lead to B_{root} holding at small sample sizes. Suppose that pre-study uncertainty about the standardized effect size Δ follows a normal distribution with mean 0, corresponding to equipoise, and standard deviation σ , and define projected study value as proportional to $\Delta P_n(\Delta)$, where $P_n(\Delta)$ is the probability of rejecting the null hypothesis $\Delta=0$ by a 2-sided, one-sample Z-test, in favor of $\Delta>0$. This might apply when Δ quantifies how much a new treatment improves on a standard in a crossover trial. This is analogous to a measure of value proposed by Detsky (1990) for studies with a binary outcome. It is also similar to the concept of “assurance” defined by O’Hagan, Stevens, and Campbell (2005) as the average power over a range of alternatives, but here the value depends not only on whether the study produces $p<0.05$ but also on the size of the true treatment effect. For this situation, B_{root} holds for $n^* \geq 3$ if $\sigma = 1.0$, for $n^* \geq 5$ if $\sigma = 0.75$, and for $n^* \geq 11$ if $\sigma = 0.5$. Thus, B_{root} holds down to fairly small sample sizes unless there is considerable prior evidence that Δ is likely to be small.

4.3. Performance of n_{root} with value proportional to power

Because power versus a single, known alternative plays such a central role in sample size planning conventions, and value based on this does not satisfy B_{root} , we note some results that nevertheless hold for n_{root} when study value is proportional to power at a specific alternative with $\alpha=0.05$ and the cost structure is linear, $c_n = F + nc$. In Web Appendix B, we show that the following two propositions hold:

Proposition 9—Either n_{root} produces more than 90% power, or it is more cost efficient than any sample size that does.

Proposition 10— n_{root} is never less than 81% as cost efficient as any larger sample size. This sharpens the general bound given by Proposition 7. Similar to Proposition 8, Propositions 9 and 10 also hold regardless of the assumed effect size used for calculating power.

4.4 Other measures of projected value

Despite the predominance of power-based sample size planning in applied studies, many other measures of projected study value have been proposed for use in sample size planning. Web Appendix A assesses conditions B_{min} and B_{root} for several such alternatives, mainly focusing on simple cases for normal and Bernoulli outcomes. These simple cases accord with the assumptions typically made for sample size planning in actual practice. We note the main results below.

For value proportional to reduction in Bayesian credible interval width from its prior width (Joseph and Belisle, 1997; Lindley, 1997; Pham-Gia, 1997), B_{min} holds for all n^* and B_{root} holds for $n^* > 1.6n_0$, where n_0 is the sample size equivalent of the prior information. For value inversely proportional to frequentist confidence interval width, B_{min} and B_{root} hold for all n^* . For value proportional to the reduction in squared error loss versus using the prior mean, B_{min} holds for all n^* and B_{root} holds for $n^* > n_0$. For value proportional to gain in Shannon information (Bernardo, 1997), B_{min} holds for all n^* and B_{root} holds for $n^* > 3.9n_0$. These results formalize the notion of B_{root} holding at small sample sizes for innovative studies, because such studies will have low values of n_0 . For Bayesian decision theory (Tan and Smith, 1998; Gittins and Pezeshk, 2000a,b; Halpern, et al., 2001; Willan and Pinto, 2005), the low-prior-information case is pre-study equipoise, by which we mean equal prior means for the outcome under the two competing treatments. Under this definition of equipoise, B_{min} holds for the normal case as we show in Web Appendix C. Numerical evaluation of many realistic cases, presented in Web Tables 1 and 2, suggests that both B_{min} and B_{root} hold down to small sample sizes for both the normal and Bernoulli cases under equipoise or when prior uncertainty about one or

both treatments is large. Severe enough departures from equipoise can, however, produce violations of $B_{\min}$ at large sample sizes.

Other proposed measures of value also appear to meet $B_{\min}$. Bacchetti et al. (2005) show that $B_{\min}$ holds for value proportional to the probability of seeing a rare side effect at least once. Walker (2003) proposes a nonparametric Bayesian approach requiring Markov chain Monte Carlo methods; the example, his Figures 1(a)–1(c), shows concavity of $v_n - \tau n$ for cost τ per subject, which implies concavity of v_n and condition $B_{\min}$. Tan and Smith (1998) consider a wide variety of utility functions, including some that address side effects as well as efficacy, and all of them show concavity (their Figures 2, 3, 5, 8–10, 12, 13).

4.5. Summary and discussion

We suggest that the above results make it reasonable to use $n_{\min}$ without explicitly verifying $B_{\min}$ for each specific study. $B_{\min}$ holds for all the measures considered and therefore for any weighted combination of them. The only exceptions that we found were for decision theoretic measures when there are departures from pre-study equipoise. Although the need for equipoise is controversial (Freedman, 1987; Rothman and Michels, 1994; Miller and Brody, 2003), the severe departures from our specific decision theoretic equipoise conditions that would be needed to make $n_{\min}$ fall substantially short of optimal cost efficiency seem likely to raise clear ethical problems and to be rare in practice. In addition, $B_{\min}$ holds for (4) not only under equipoise, but also if a departure from equipoise favors the alternative; substantial departures in favor of the null would usually mean that a study is of low interest at any sample size. Finally, we note that the decision theoretic measures seem the least compelling in our view, because they artificially dichotomize the outcome of the study.

The consistency of our results concerning when B_{root} holds also suggest no need for specific verification for each study, beyond simply confirming that there is little known about the issue the study addresses. Several Bayesian measures provide formal results indicating that B_{root} holds when there is little prior information, and Section 4.2 and Web Tables 1 and 2 provide additional evidence that this is a consistent phenomenon. In addition, Propositions 9 and 10 provide additional reassurance that n_{root} will not perform poorly, although only for the standard measure of projected value, power.

Thus, while it may be possible to construct specialized measures of value and situations where our methods would perform poorly, we believe that they should work well in typical situations.

5. Further examples

5.1. Comparison of two treatments' cure rates

Table 2 extends the 2-treatment comparison situation from Section 2 to seven different measures of study value, showing the cost efficiency attained by n_{root} or by requiring 80% power for various alternatives. We now assume a linear cost structure, which is more favorable toward larger sample sizes because it implies that cost per subject decreases with increasing sample size. We focus only on the ratio of fixed to per-subject costs, because the specific magnitude of the costs does not impact the performance of one choice relative to optimal or to another. As expected from Proposition 2, n_{root} never falls short of the most cost efficient sample size when there is low prior information, $n_0=4$ or equipoise. It also avoids very poor cost efficiency in all situations, in accordance with Propositions 6 and 7, and it compares well to larger sample sizes on the power-based measure, in accordance with Proposition 10. In contrast, the conservative power-based choice assuming an alternative cure rate of 50% is poor for many measures under the middle cost structure and very poor in the bottom cost structure, even for value proportional to power. The power-based choice with the alternative specified as 60% is also poor for the bottom cost structure for the situations with low prior information, precisely

where n_{root} performs well. The over-optimistic choice assuming an alternative cure rate of 0.8 performs very poorly in the top cost structure for many measures, notably value proportional to power.

5.2. Actual grant proposals

We recently collaborated on two proposals at opposite extremes in terms of how cost influenced sample size choice. One concerned innovative methods for islet cell transplantation to treat Type I diabetes with serious complications, a procedure that requires approximately \$100,000 in clinical costs per patient. Because of these costs, the proposed sample size was only ten. This could be justified using the ideas presented here. With little already known about these new methods, condition B_{root} is likely to hold down to very small sample sizes. But ten is already more than n_{root} , because fixed costs are less than \$1 million (see Proposition 5). So ten should not fall short of the most cost efficient sample size: $10 > n_{\text{root}} \geq n_{\text{opt}}$. Despite this, ten would seem inadequate by a conventional power analysis, and the proposal was criticized for not including one.

The other study concerned measurement of antiretroviral drug levels in hair specimens that are already collected as part of an existing cohort study of HIV-infected women. Fixed costs that would be incurred regardless of sample size included considerable effort for assay development, setting up data management and other procedures, data analysis, scientific leadership and guidance, and presentation and publication of results. Per-specimen costs included technician time and supplies for running the assays, entry and cleaning of data, shipping, and project monitoring. Without formal analysis, the investigators proposed the simple choice of studying all 5700 specimens projected to become available over the study period. This reflected the clear reality that set-up costs would be high, per-specimen costs would be low, and a wealth of concomitant information would be collected by the parent study anyway, leveraging the value of each specimen.

Developing a new source of additional specimens would be very expensive, so the proposed sample size of 5700 was in fact equal to n_{min} , implying that it was more cost efficient than any larger sample size. Clearly, increasing sample size was not viable, and decreasing it much would also appear likely to worsen cost efficiency, failing to fully exploit the development of the assays and the freely available concomitant information on the subjects. The choice of n_{min} reflects these facts and provides justification for the proposed sample size. In contrast, power-based sample size calculations would be very unlikely to produce this obvious choice unless specifically manipulated to do so. This proposal was also criticized for failing to show 80% power for one of its aims. These two cases exemplify reviewers' current tendency to criticize sample size choices, even when cost considerations leave little real doubt about what is best.

6. Discussion

Setting a goal that must be reached at any cost is often impractical. If there were a minimum sample size that was necessary for a study to have any value at all, then a study should either be done right or not at all, and costs would not be relevant. But power and other measures of projected value that have been proposed for use in sample size planning do not produce any such necessary minimum. We therefore see no justification for making $\geq 80\%$ power an absolute requirement, which would preclude use of our approach or other alternatives. Instead, these measures all exhibit the properties that justify our simple approach for incorporating costs in sample size planning. We have shown that planners can rely on n_{min} being more cost efficient than any larger sample size. For innovative studies, n_{root} will also have this very desirable property. Our approach also has the following important strengths.

6.1. Strengths of the proposed approach

Reliable—By choosing sample size based on properties that hold for any of the measures of projected value considered here, we avoid relying entirely on a single definition and avoid the risk of being misled by incorrectly specified inputs, such as the assumed difference for calculating power. In addition, costs are a requisite ingredient in most trial planning and seem easier to reliably project in detail than eventual scientific or practical value, which depends on the unknown factors that are to be studied. Also, improved cost efficiency is an acceptable goal in a wide variety of situations. Thus, uncertainty or disagreement in any of these areas will not invalidate our approach. Furthermore, Propositions 6, 7, and 10 provide assurance that n_{root} will not be disastrously bad if the cost structure is linear. Although the bound of 50% of optimal may seem weak, the daunting uncertainties often present in sample size planning situations make cost efficiencies of much less than 50% a real danger under conventional approaches. For example, in Table 2 the power-based choices dip well below 50% of optimal in many situations and one reaches as low as 7% of optimal.

Defensible—Our method justifies the choice of small sample sizes in situations such as high per-subject cost and low prior information, where such a choice is often unavoidable on practical grounds but cannot be justified by power considerations. Combined with the results in Section 4, Proposition 1 ensures that n_{min} is more cost efficient than any larger sample size, as is n_{root} when the study is innovative in the sense of having low prior information. Furthermore, Proposition 9 shows that n_{root} will generally be more cost efficient than the sample size producing 90% power, even under a cost structure and definition of value that are favorable toward larger sample sizes. Because reviewers frequently criticize sample size as too small, these properties make n_{min} and n_{root} viable choices for investigators fearful of such criticism.

Easy to use—Proposition 5 shows that n_{root} is particularly easy to calculate if costs are linear. Finding n_{min} will often just require estimating how many subjects can be studied before encountering a cost barrier, such as the need to open another site. We suspect that investigators are already familiar with using such projections to pick sample sizes (this is the first step in what Goodman and Berlin (1994) described as a “sample size game”). Costs usually must be projected as part of a study proposal anyway. We do note that costs to society not included in proposals may be important, such as risks to study subjects, time spent planning the study, and publishing costs. Propositions 3 and 4 indicate the qualitative impacts that these should have on sample size choice. These wider costs still seem easier to quantify than projected value. A possible exception is risks to subjects, but if these are ignored in calculating n_{min} or n_{root} , then Proposition 4 ensures that our approach is still guaranteed not to fall short of the most cost efficient sample size defined with risks included.

Promotes innovation—Astute clients sometimes ask, “How can we calculate a sample size when no one has studied this before?” Conventional approaches provide no good answer, instead resorting to arbitrary values or meaningless standardized effect sizes that have nothing to do with the study. Preliminary data can mitigate this problem but may be unavailable for very innovative studies. Indeed, an overemphasis on preliminary data has been recognized as a barrier to innovation by a National Institutes of Health task force, which wrote: “an obsession with preliminary data discriminates against bold new ideas, against young scientists, and against risk taking. For new ideas, little or no preliminary data may be required” (<http://cms.csr.nih.gov/NewsandReports/ReorganizationActivitiesChannel/BackgroundUpdatesandTimeline/FinalPhaseIReport.htm>, accessed 10 March 2007). Unfortunately, reduced expectations concerning preliminary data still would not help innovators with the practical problem of choosing and justifying a sample size. But n_{root} is a

viable solution: it is especially applicable to innovative studies and provides a sensible and defensible choice without a need for any preliminary data.

6.2. Additional remarks

Following much of the literature on sample size planning using costs (e.g., Detsky, 1990; Bernardo, 1997; Halpern, et al., 2001; Willan and Pinto, 2005), we focus on linear cost structure in Table 2 and some propositions. This simple assumption implies constant marginal costs as sample size increases, which may be reasonably realistic for many studies, but may in some cases be unduly favorable toward larger sample sizes. In practice, there may be a limited number of the most convenient, easily studied subjects, and costs will increase more than linearly when sample size exceeds that number.

Our analysis does not imply that very large studies cannot be justified. For a study addressing an important issue, a very large sample size may produce value greater than cost even when it is much larger than the most cost efficient sample size. If the issue is important enough, the resulting return on investment, while less than the maximum possible, may nevertheless exceed what is possible for investigations on other topics that might be funded instead. In this case, money spent to exceed the most cost efficient sample size will still be well spent. Universal use of $n_{\min}$ or n_{root} therefore would not produce a society-wide optimum return on research spending. We do not advocate that our proposed choices be mandatory for every study, only that they should be regarded as acceptable whenever a larger sample size would be acceptable. We also note that the first example in Section 5.2 illustrates how our ideas could be used to defend a sample size that exceeds n_{root} . On the other hand, we also emphasize that nothing we have proposed prevents use of *smaller* sample sizes if these are suggested by other approaches or considerations. In general, we advocate more tolerance of different sample size planning approaches, not adoption of the proposals here as a new rigid standard.

We have focused here on studies with fixed sample sizes; consideration of adaptive designs is beyond the scope of this paper. Although group sequential designs are a long-established approach for potentially improving cost efficiency, they are not feasible in all situations, and many studies are carried out with fixed sample sizes. This is particularly true for small, innovative studies where n_{root} is most applicable. Funding for the maximum sample size must be committed in full at the outset for sequential studies, and costs when stopping early will likely exceed what they would have been for a study with a planned end at the same point. In addition, sequential studies may still be inherently less cost efficient than smaller studies. In the example from Section 2, adding standard interim analyses (O'Brien and Fleming, 1979) to the studies with doubled or quadrupled sample size would not reduce their expected cost per expected cure down to that of the smallest study, even if they could be carried out at the same cost per subject. Finally, we note that sequential or other adaptive designs could still be used when the maximum possible sample size has been chosen as $n_{\min}$ or n_{root} instead of by a conventional power calculation..

6.3. Conclusion

The proposals here run counter to some established ideas about sample size planning, but we believe that our approach is nevertheless reasonable and promising. It has the potential to improve allocation of research funding, promote innovation, and save much investigator, reviewer, and statistician time that is currently wasted justifying and evaluating sample size under standards that are often unrealistic and too inflexible. We hope that this paper will lead to more tolerance of alternatives to the rigid conventions that seem to have a stranglehold on practices concerning sample size planning.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *American Journal of Epidemiology* 2005;161:105–110. [PubMed: 15632258]
- Bernardo JM. Statistical inference as a decision problem: the choice of sample size. *The Statistician* 1997;46:151–153.
- Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. *Health Economics* 1996;5:513–514. [PubMed: 9003938]
- Detsky AS. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Statistics in Medicine* 1990;9:173–184. [PubMed: 2111932]
- Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141–145. [PubMed: 3600702]
- Gittins J, Pezeshk H. How large should a clinical trial be? *The Statistician* 2000a;49:177–197.
- Gittins J, Pezeshk H. A behavioral Bayes method for determining the size of a clinical trial. *Drug Information Journal* 2000b;34:355–363.
- Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 1994;121:200–206. [PubMed: 8017747]
- Graham, RL.; Knuth, DE.; Patashnik, O. *Concrete Mathematics: A Foundation for Computer Science*. Vol. 2. Reading, MA: Addison-Wesley; 1994.
- Halpern J, Brown BW, Hornberger J. The sample size for a clinical trial: a Bayesian-decision theoretic approach. *Statistics in Medicine* 2001;20:841–858. [PubMed: 11252007]
- Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association* 2002;288:358–362. [PubMed: 12117401]
- Horrobin DF. Peer review of statistics in medical research - Rationale for requiring power calculations is needed. *British Medical Journal* 2002;325:491–492. [PubMed: 12211232]
- Hulley, SB.; Cummings, SR.; Browner, WS.; Grady, D.; Hearst, N.; Newman, TB. *Designing Clinical Research: An Epidemiological Approach*. Vol. 2. Philadelphia: Lippincott, Williams, and Wilkins; 2001.
- Joseph L, Belisle P. Bayesian sample size determination for normal means and differences between normal means. *The Statistician* 1997;46:209–226.
- Joseph L, Du Berger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* 1997;16:769–781. [PubMed: 9131764]
- Lee SJ, Zelen M. Clinical trials and sample size considerations: another perspective. *Statistical Science* 2000;15:95–103.
- Lindley DV. The choice of sample size. *The Statistician* 1997;46:129–138.
- Miller FG, Brody H. A critique of clinical equipoise: therapeutic misconception in the ethics of clinical trials. *Hastings Center Report* 2003;33:19–28. [PubMed: 12854452]
- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–556. [PubMed: 497341]
- O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceutical Statistics* 2005;4:187–201.
- Pham-Gia T. On Bayesian analysis, Bayesian decision theory and the sample size problem. *The Statistician* 1997;46:139–144.
- Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med* 1994;331:394–398. [PubMed: 8028622]
- Tan SB, Smith AFM. Exploratory thoughts on clinical trials with utilities. *Statistics in Medicine* 1998;17:2771–2791. [PubMed: 9881422]

- Walker SG. How many samples?: a Bayesian nonparametric approach. *The Statistician* 2003;52:475–482.
- Willan AR, Pinto EM. The value of information and optimal clinical trial design. *Statistics in Medicine* 2005;24:1791–1806. [PubMed: 15806619]
- Yokota F, Thompson KM. Value of information analysis in environmental health risk management decisions: past, present, and future. *Risk Analysis* 2004;24:635–650. [PubMed: 15209935]

Figure 1.

Shapes of the relation between projected value and sample size for 10 measures of study value and situations. For visual clarity and because only the shapes are of interest, the vertical scale varies for different curves. Shown are curves for value proportional to: **a)** gain in Shannon information with $n_0=100$, where n_0 is the sample size equivalent of the prior information; **b)** reciprocal of confidence interval width; **c)** reduction in Bayesian credible interval width when $n_0=100$; **d)** reduction in squared error versus using prior mean when $n_0=100$; **e)** power for a standardized effect size of 0.2; **f)** additional cures from a Bayesian clinical trial with prior means (SDs) for cure rates of 0.4 (0.05) versus 0.4 (0.1); **g)** gain in Shannon information with $n_0=2$; **h)** reduction in squared error versus using a single observation; **i)** reduction in squared error versus using prior mean when $n_0=2$; **j)** reduction in Bayesian credible interval width when $n_0=2$.

Table 1

Influence of sample size on power, projected study value, and cost efficiency. Calculations assume an expected 100,000 persons will receive the new treatment if the study achieves $p < 0.05$, and that there is a 0.25 probability that the treatment has the assumed cure rate.

Performance measure	Sample size per arm	Performance with assumed cure rate with new treatment equal to:		
		50%	54%	60%
Power	97	29%	50%	80%
	196	51%	80%	98%
	388	80%	98%	>99%
Expected additional cures	97	717	1741	4002
	196	1280	2784	4897
	388	2002	3414	4999
Cost per expected cure (\$)	97	279	115	50
	196	391	180	102
	388	500	293	200
Cures per \$100,000 spent	97	358	870	2001
	196	256	557	979
	388	200	341	500

Table 2

Comparison of cost efficiencies of methods for choosing sample size under different cost structures and measures of projected study value.

Cost Structure	% of optimal cost efficiency, by method*				
	80% power for 40% vs				
Measure of study value proportional to	Optimal n	n_{root}	50%	60%	80%
Fixed costs are 1000 times per-subject cost, $n_{\text{root}}=1000$					
Power for cure rate 60% versus 40%	300	69%	78%	93%	34%
Reduction in credible interval width, $n_0=4$	120	64%	71%	98%	93%
Reduction in credible interval width, $n_0=200$	600	95%	99%	77%	29%
Reciprocal of confidence interval width	1000	100%	99%	74%	40%
Bayesian decision, means=40%, SDs 5%, 10%	152	64%	74%	99%	86%
Shannon information, $n_0=4$	296	83%	89%	98%	71%
Shannon information, $n_0=200$	914	100%	99%	63%	21%
Fixed costs are 100 times per-subject cost, $n_{\text{root}}=100$					
Power for cure rate 60% versus 40%	158	93%	41%	98%	65%
Reduction in credible interval width, $n_0=4$	30	80%	21%	58%	98%
Reduction in credible interval width, $n_0=200$	174	94%	64%	100%	67%
Reciprocal of confidence interval width	100	100%	64%	95%	92%
Bayesian decision, means=40%, SDs 5%, 10%	38	86%	24%	65%	100%
Shannon information, $n_0=4$	54	94%	35%	76%	95%
Shannon information, $n_0=200$	232	87%	78%	99%	60%
Fixed costs are 20 times per-subject cost, $n_{\text{root}}=20$					
Power for cure rate 60% versus 40%	88	84%	29%	86%	95%
Reduction in credible interval width, $n_0=4$	12	95%	7%	26%	71%
Reduction in credible interval width, $n_0=200$	76	75%	44%	87%	95%
Reciprocal of confidence interval width	20	100%	31%	58%	93%
Bayesian decision, means=40%, SDs 5%, 10%	14	96%	9%	31%	77%
Shannon information, $n_0=4$	18	100%	15%	41%	87%
Shannon information, $n_0=200$	96	71%	59%	94%	92%

* Chosen total n for 80% power for 40% versus 50% is 776, for 40% versus 60% is 194, for 40% versus 80% is 44. These remain the same for all cost structures.