Collection of Biostatistics Research Archive
COBRA Preprint Series

Simple, Defensible Sample Sizes Based on Cost Efficiency – With Discussion and Rejoinder

Peter Bacchetti[*]   Charles E. McCulloch[†]   Mark R. Segal[‡]

Richard Simon[**]   Peter Muller[††]   Gary L. Rosner[‡‡]

James A. Hanley[§]   Stan Shapiro[¶]

[*]University of California, San Francisco, peter@biostat.ucsf.edu

[†]University of California - San Francisco, chuck@biostat.ucsf.edu

[‡]University of California, San Francisco, mark@biostat.ucsf.edu

[**]NIH, rsimon@nih.gov

[††]University of Texas M.D. Anderson Cancer Center, pm@odin.mdacc.tmc.edu

[‡‡]University of Texas, M. D. Anderson Cancer Center, glr@odin.mdacc.tmc.edu

[§]McGill University, james.hanley@mcgill.ca

[¶]McGill, stan.shapiro@mcgill.ca

Simple, Defensible Sample Sizes Based on Cost Efficiency – With Discussion and Rejoinder

Peter Bacchetti, Charles E. McCulloch, Mark R. Segal, Richard Simon, Peter Muller, Gary L. Rosner, James A. Hanley, and Stan Shapiro

Abstract

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.

Simple, Defensible Sample Sizes
Based on Cost Efficiency

With Discussion and Rejoinder

Peter Bacchetti
Charles E. McCulloch
Mark R. Segal
Division of Biostatistics, Department of Epidemiology and Biostatistics
University of California, San Francisco, CA 94143-0560 USA
peter@biostat.ucsf.edu

Richard Simon
National Cancer Institute
9000 Rockville Pike
Bethesda MD 20892-7434
rsimon@nih.gov

Peter Müller and Gary L. Rosner,
Department of Biostatistics,
The University of Texas, M. D. Anderson Cancer Center
Houston, TX 77030, U.S.A

James A Hanley and Stanley H Shapiro
Department of Epidemiology, Biostatistics and Occupational Health
McGill University
Montreal

1

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.

Key words: 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.

 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

2

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

3

"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.

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

4

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,

5

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

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

and

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

then

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

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:

$$f(n) = n \tag{1}$$

and

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

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 $n_{\text{root}} \leq n_{\min}$. 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.

6

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 $c > 0$, can only decrease n_{min}, n_{root}, and n_{opt}, never increase them.

Proposition 5. If $c_n = F + nc$, then $n_{\mathrm{root}} = F/c$ 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 $\{(F + nc)/(2F)\}\{v_{n_{\mathrm{root}}}/v_n\}$.

Condition B_{min} implies $v_{n_{\mathrm{root}}}/v_n \geq n_{\mathrm{root}}/n$. Letting $r = n/n_{\mathrm{root}}$ 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}.
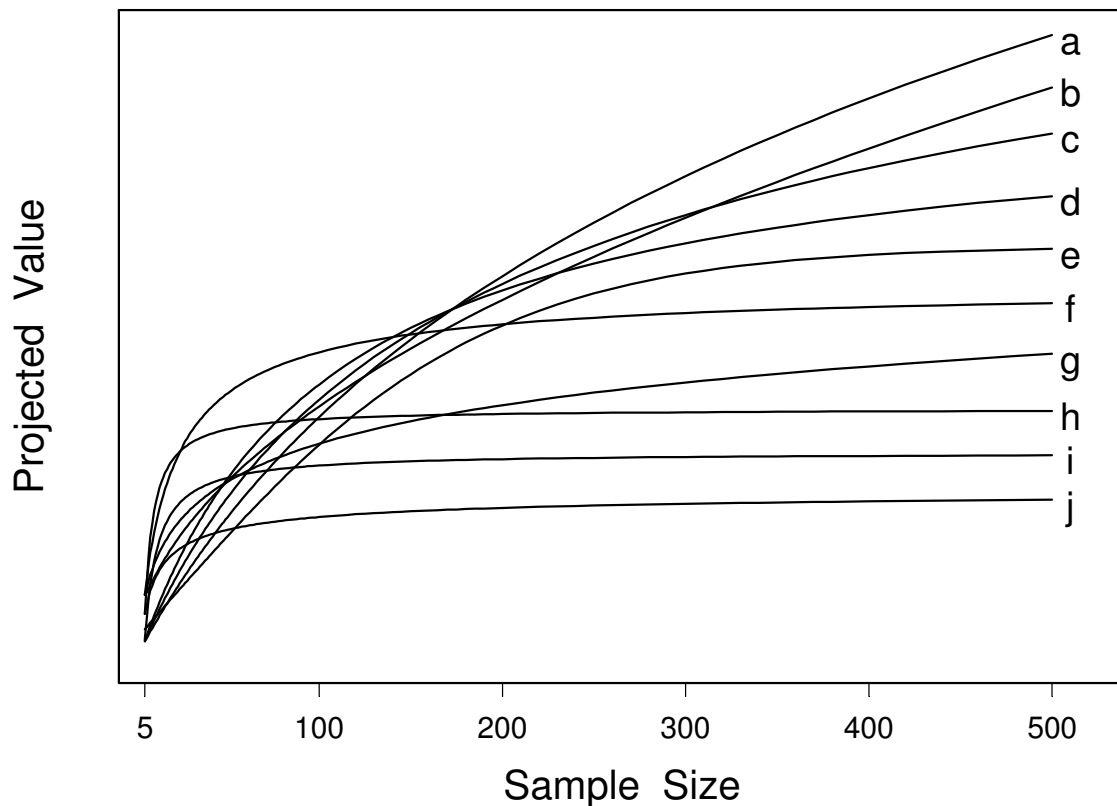
7

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.

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

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

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:

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

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

9

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-τn 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: $10 > n_{root} \geq n_{opt}$. 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

Table 2.

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

| Cost Structure | Optimal | % of optimal cost efficiency, by method* | | |
| | | | 80% power for 40% vs | |
Measure of study value proportional to	n	n_{root}	50%	60%	80%
Fixed costs are 1000 times per-subject cost, $n_{root}=1000$					
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, $n_{root}=100$					
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%	99%
Shannon information, $n_0=200$	232	87%	78%	99%	60%
Fixed costs are 20 times per-subject cost, $n_{root}=20$					
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.

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 ≥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.

13

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/BackgroundUpdatesa ndTimeline/FinalPhase1Report.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

14

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 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.

References

Bacchetti, P., Wolf, L.E., Segal, M.R., and McCulloch, C.E. (2005) Ethics and sample size. *American Journal of Epidemiology*, 161, 105-110.

Bernardo, J.M. (1997), Statistical inference as a decision problem: the choice of sample size, *The Statistician*, 46, 151-153.

Claxton, K., and Posnett, J. (1996) An economic approach to clinical trial design and research priority-setting. *Health Economics*, 5: 513-514.

Detsky, A.S. (1990) Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Statistics in Medicine*, 9, 173-184.

Freedman, B. (1987) Equipoise and the ethics of clinical research. *N Engl J Med*, 317, 141-145.

Gittins, J. and Pezeshk, H. (2000a) How large should a clinical trial be? *The Statistician*, 49, 177-197.

Gittins, J. and Pezeshk, H. (2000b) A behavioral Bayes method for determining the size of a clinical trial. *Drug Information Journal*, 34, 355-363.

Goodman, S.N., and Berlin, J.A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results, *Annals of Internal Medicine*, 121, 200-206.

Graham, R.L., Knuth, D.E., and Patashnik, O. (1994) *Concrete Mathematics: A Foundation for Computer Science, 2nd Ed.*, Reading, MA: Addison-Wesley.

Halpern, J., Brown, B.W., and Hornberger, J. (2001) The sample size for a clinical trial: a Bayesian-decision theoretic approach, *Statistics in Medicine*, 20, 841-858.

Halpern, S.D., Karlawish, J.H.T., and Berlin, J.A. (2002) The continuing unethical conduct of underpowered clinical trials, *Journal of the American Medical Association*, 288, 358-362.

Horrobin, DF. (2002) Peer review of statistics in medical research - Rationale for requiring power calculations is needed. *British Medical Journal* 325: 491-492.

Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. (2001) Designing Clinical Research: An Epidemiological Approach, 2nd Ed. Philidelphia: Lippincott, Williams, and Wilkins.

Joseph, L., and Belisle, P. (1997) Bayesian sample size determination for normal means and differences between normal means, *The Statistician*, 46, 209-226.

Joseph, L., Du Berger, R., and Belisle, P. (1997) Bayesian and mixed Bayesian/likelihood criteria for sample size determination, *Statistics in Medicine*, 16, 769-781.

Lee, S.J., and Zelen, M. (2000) Clinical trials and sample size considerations: another perspective, *Statistical Science*, 15, 95-103.

Lindley, D.V. (1997) The choice of sample size, *The Statistician*, 46, 129-138.

Miller, F.G., and Brody, H. (2003). A critique of clinical equipoise: therapeutic misconception in the ethics of clinical trials. *Hastings Center Report*, 33, 19-28.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549-556.

O'Hagan A, Stevens JW, and Campbell MJ. (2005) Assurance in clinical trial design. *Pharmaceutical Statistics*, 4, 187-201.

Pham-Gia, T. (1997) On Bayesian analysis, Bayesian decision theory and the sample size problem, *The Statistician*, 46, 139-144.

Rothman, K.J., and Michels, K.B. (1994). The continuing unethical use of placebo controls. *N Engl J Med*, 331, 394-398.

Tan, S.B., and Smith, A.F.M. (1998) Exploratory thoughts on clinical trials with utilities, *Statistics in Medicine*, 17, 2771-2791.

Walker, S.G. (2003) How many samples?: a Bayesian nonparametric approach, *The Statistician*, 52, 475-482.

Willan AR, Pinto EM. (2005) The value of information and optimal clinical trial design. *Statistics in Medicine*, 24: 1791-1806.

Yokota F, Thompson KM. (2004). Value of information analysis in environmental health risk management decisions: past, present, and future. *Risk Analysis*, 24, 635-650.

16

Comment on

"Simple Defensible Sample Sizes Based on Cost Efficiency"

Richard Simon
National Cancer Institute
9000 Rockville Pike
Bethesda MD 20892-7434
rsimon@nih.gov

Bacchetti, McCulloch and Segal (BMS) argue for an approach to sample size planning for clinical trials based on maximizing "cost efficiency" defined as v_n/c_n where v_n is the expected scientific, clinical, or practical value of the study if the sample size is n and c_n is the cost of conducting a study with sample size n. They give examples of plausible definitions of v_n which increase less than linearly with n. For such functions, the increase in value per unit increase in cost ultimately declines, although it may increase with n for small sample sizes because of a fixed study set-up cost. For example, in comparing a new treatment to control with a normally distributed endpoint and known variance, one might define v_n as the power for detecting a mean difference δ at significance level α. This can be written

$$v_n = \Phi\left\{\frac{\delta}{2\sigma}\sqrt{n} - k_{1-\alpha}\right\} \tag{1}$$

where Φ denotes the standard normal cdf, $k_{1-\alpha}$ denotes the 1-α percentile of the standard normal distribution and n is the total sample size for the two arms of the study. Figure 1 shows this power as a function of n for $\delta/2\sigma = 0.5$ and α=0.025. Figure 2 shows $v_n /(10+n)$ which would be the form of the cost efficiency function if the set-up cost were equivalent to the cost of treating 10 patients. In this case, the optimum cost efficiency is seen to occur at a sample size of about 20 patients, although the power for such a study would only be about .61 even for a large treatement effect $\delta/2\sigma = 0.5$.

Statisticians are very familiar with the fact that power increases at a decreasing rate with increasing sample size. The same is true for the width of a confidence interval for treatment difference. Nevertheless, the argument that such a "cost efficiency" measure is appropriate for sample size planning is not compelling. For example, a company with a single indication product to clinically develop may be best served by funding the largest study that they can afford in order to show that the product is statistically significantly better than the control. Obtaining regulatory approval for marketing based on large power for a small treatment effect may be in the financial interest of the company, not performing a small "cost efficient" study. Even in cases where the value measure v_n is an appropriate reflection of utility of the study, maximizing value subject to restrictions on resources may result in very different decisions than maximizing the "cost efficiency" ratio described by BMS. In general, study value and study costs are incommensurate and not usefully combined except in the sense that the studies that maximize value within the limits of resources are those that should be undertaken (Anscombe, 1963).

The proposal for using sample sizes based on cost efficiency reflects an attempt to improve the allocation of resources among competing opportunities. Allocation of public funds for biomedical research should take into account the opportunity costs. That is, the funds provided for one proposal are not available for other proposals. This is also true for much private sector

research. Effectively allocating resources among disparate projects must, however, consider the disparate objectives of the projects and the likelihood that they will achieve their objectives. For therapeutics development, effective resource allocation strategies should utilize sequentially accumulating information; this is the purpose of phased clinical development. Phase I and II clinical trials are undertaken to determine which treatments warrant phase III clinical trials and to optimize the regimens and target populations for those that are undertaken. Several phase II design strategies have been proposed to take opportunity costs into account directly. Strauss and Simon (1995) proposed a sequence of two-arm randomized phase II trials, selecting the "winner" at each trial for carry-over to the next trial, to optimally utilize a horizon of N patients in a manner that maximized the expected effectiveness of the treatment selected at the end. They showed that the best sample size in each trial depended on the distribution of treatment effects; if there are treatments with large effects, it is better to perform a large number of small phase II studies than a small number of large phase II studies. Whitehead (1985, 1986), Yao et al. (1996), Yao and Venkatraman (1998), and Wang and Leung (1998) described similar approaches based on a series of single arm pilot trials.

There can be no meaningful one-size fits all rule for sizing randomized clinical trials. Randomized clinical trials, including those involving randomization between a new regimen and a control have an important role as developmental phase II studies. This is increasingly true in oncology today where endpoints such as time-to-disease progression and the contribution of new drugs to existing regimens cannot be meaningfully evaluated in single arm trials (Korn et al., 2001). The endpoints, significance threshold for judging a treatment as promising, and sample size planning for such developmental trials should be based on different considerations than for phase III trials whose objectives are to potentially change medical practice. For example, Simon et al. proposed using ranking and selection theory (Simon, Wittes and Ellenberg, 1985). Simon et al. (2001) and Rubinstein et al. (2005) proposed using elevated thresholds for statistical significance in some randomized developmental trials. Randomized developmental trials of therapeutic regimens with a strong biological basis should be encouraged, not prevented because they do not fit the mold appropriate for sizing phase III trials.

When it comes time to conduct a phase III trial, there is usually strong interest in obtaining a definitive result. Large multi-center clinical trials are often difficult to organize, in part because many investigators would rather perform their own single center studies (Simon, 1994). Many published studies report "suggestive" non-statistically significant treatment effects but have too small a sample size to be considered negative. Such trials may be of value in guiding further research or in contributing to meta-analyses, but are in themselves not interpretable for medical decision making. BMS are correct that the usual approach to sizing phase III trials is imperfect on many counts. It certainly is subjective with regard to the size of treatment effect to be detected. In comparing a new treatment to a control, there is usually a difference δ_{min} of the primary endpoint considered to be the minimal medical benefit, in the sense that smaller differences for that endpoint would not be worth the known and unknown risks of adverse effects for other endpoints. Even if there were no subjectivity in specifying δ_{min} , providing 90% power for rejecting the null hypothesis when the treatment effect is δ_{min} does not ensure that treatments found statistically significant have effectiveness greater than δ_{min} .There are also rather serious concerns about whether the standard frequentist approach provides an appropriate context for utilizing previous information (Spiegelhalter, Freedman and Parmar, 1994) . For example, for some confirmatory trials, the usual 0.05 level may be too stringent (Parmar, Ungerleider and Simon, 1996). On the other hand, for a treatment regimen with a small prior probability of success, even a statistically significant (p<0.05) result will be associated with a large posterior

probability of being a false positive and will require confirmation in a follow-up study (Simon, 2000). Using the Bayesian paradigm and planning sample size to assure that results are convincing to both skeptics and enthusiasts will often require much larger sample sizes than the standard frequentist approach (Spiegelhalter et al., 1994). On the other hand, maximizing utility to patients included in the trial and the horizon of patients who may utilize a new approved treatment in some cases supports smaller clinical trials; large treatment effects are easy to detect with small sample sizes and treatments with small effects have limited impact on utility except when large numbers of patients are influenced by the trial (Anscombe, 1963; Begg and Mehta, 1979; Colton, 1963).

The large randomized phase III clinical trial has served society very well. It is not, however, a proper prescription for all situations. The future of medicine will emphasize therapeutics molecularly targeted to the intrinsic causes of disease pathogenesis and utilize predictive diagnostics to select treatments for patients. The implicit assumption of past clinical trials that qualitative treatment by covariate interactions are unlikely will be inappropriate for many new clinical trials; in many cases, qualitative interactions will be likely. New clinical trial designs and analysis strategies will be needed and biostatisticians will need to learn to move from a retrospective correlative science mode of operation to a prospective predictive medicine mode (Freidlin and Simon, 2005; Jiang, Freidlin and Simon, 2007; Simon, 2004, 2007; Simon and Maitournam, 2006)}. Many of the small treatment effects observed in past studies were probably due to the heterogeneity of the patients being treated in the same clinical trial and the absence of biological tools for adequately diagnosing and characterizing disease. The wealth of new tools offered by genomics and biotechnology provide great opportunities to dramatically improve therapeutics. Randomized clinical trials will continue to be important, but new ways of designing them and sizing them will be developed. The exciting scientific challenges available to biostatisticians involved in therapeutics today require clarity about the objectives of specific trials and innovative thinking about design and analysis strategies appropriate to those objectives.

REFERENCES

Anscombe, R. F. (1963). Sequential medical trials. *Journal of American Statistical Association* **58**, 365-383.

Begg, C. B., and Mehta, C. R. (1979). Sequential analysis of comparative clinical trials. *Biometrika* **66**, 97-103.

Colton, T. (1963). A model for selecting one of two medical treatments. *Journal of American Statistical Association* **58**, 388-400.

Freidlin, B., and Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11**, 7872-7878.

Jiang, W., Freidlin, B., and Simon, R. (2007). Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute* **99**, 1036-1043.

Korn, E. L., Arbuck, S. G., Pluda, J. M., Simon, R., Kaplan, R. S., and Christian, M. C. (2001). Clinical trial designs for cytostatic agents: Are new approaches needed? *Journal of Clinical Oncology* **19**, 265-272.

Parmar, M. K. B., Ungerleider, R. S., and Simon, R. (1996). Assessing whether to perform a confirmatory randomised clinical trial. *Journal of the National Cancer Institute* **88**, 1645-1651.

19

Rubinstein, L. V., Korn, E. L., Freidlin, B., S.Hunsberger, S.P.Ivy, and M.A.Smith (2005). Design issues of randomized phase 2 trials and a proposal for phase 2 screening trials. *Journal of Clinical Oncology* **23**, 7199-7206.

Simon, R. (1994). Randomized clinical trials in oncology: principles and obstacles. *Cancer* **74**, 2614.

Simon, R. (2000). Commentary on "Clinical and sample size considerations: Another perspective". *Statistical Science* **15**, 95-110.

Simon, R. (2004). An agenda for clinical trials: Clinical trials in the genomic era. *Clinical Trials* **1**, 468-470.

Simon, R. (2007). New challenges for 21st century clinical trials. *Clinical Trials* **4**, 167-169.

Simon, R., and Maitournam, A. (2006). Correction & Supplement: Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* **12**, 3229.

Simon, R., Wittes, R. E., and Ellenberg, S. S. (1985). Randomized phase II clinical trials. *Cancer Treatment Reports* **69**, 1375-1381.

Simon, R. M., Steinberg, S. M., Hamilton, M.*, et al.* (2001). Clinical trial designs for the early clinical development of therapeutic cancer vaccines. *Journal of Clinical Oncology* **19**, 1848-1854.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Series A, General* **157**, 357-387.

Strauss, N., and Simon, R. (1995). Investigating a sequence of randomized phase II trials to discover promising treatments. *Statistics in Medicine* **14**, 1479-1489.

Wang, Y. G., and Leung, D. H. (1998). An optimal design for screening trials. *Biometrics* **54**, 243-250.

Whitehead, J. (1985). Designing phase II studies in the context of a program of clinical research. *Biometrics* **41**, 373-383.

Whitehead, J. (1986). Sample sizes for phase II and phase III clinical trials: an integrated approach. *Statistics in Medicine* **5**, 459-464.

Yao, T. J., Begg, C. B., and Livingston, P. O. (1996). Optimal sample size for a series of pilot trials of new agents. *Biometrics* **52**, 992-1001.

Yao, T. J., and Venkatraman, E. S. (1998). Optimal two-stage design for a series of pilot trials of new agents. *Biometrics* **54**, 1183-1189.
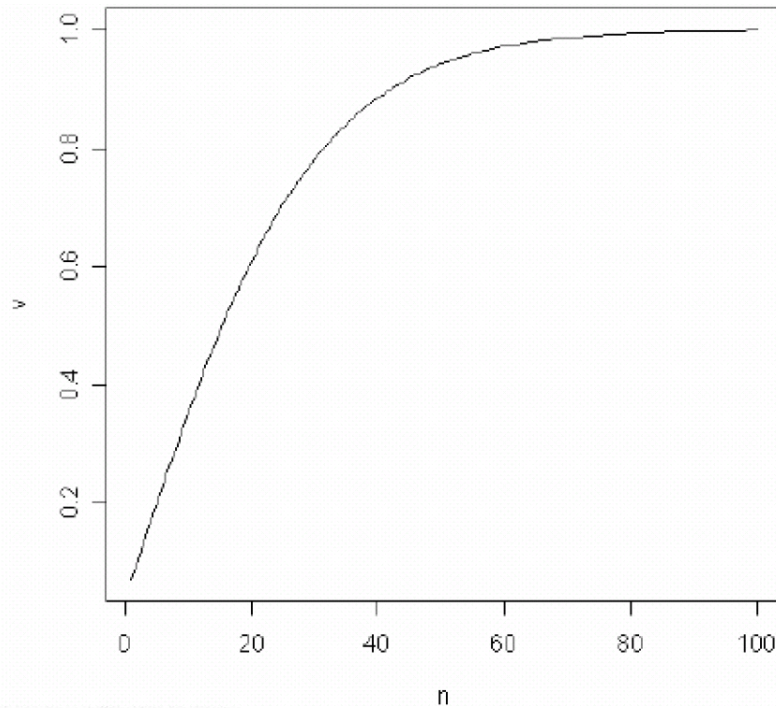
Legend for Figure 1

Power for comparing two treatment groups. Endpoint is normally distributed with equal variance. Horizontal axis is total number of cases, distributed equally between the groups. Vertical axis is power as computed from expression (1) with $\delta/2\sigma = 0.5$ and $\alpha = 0.025$.
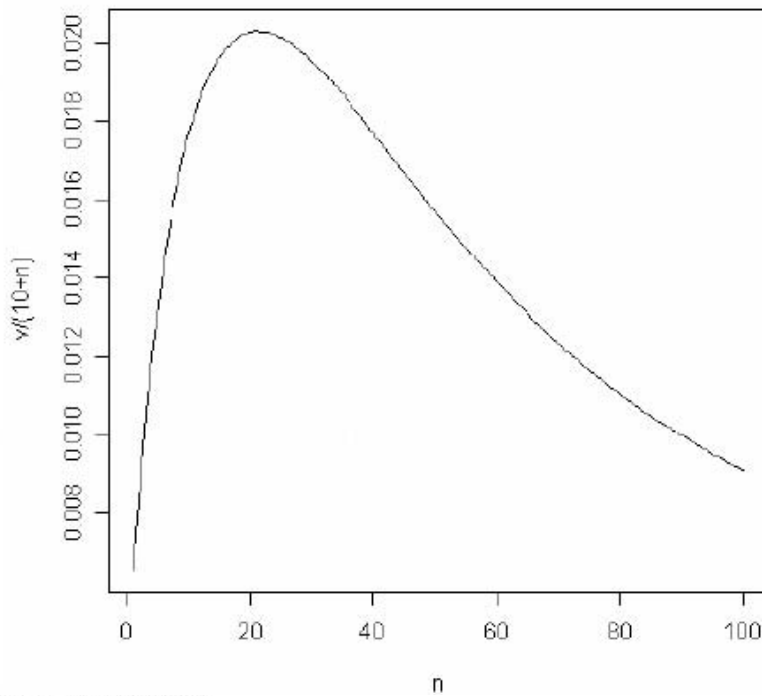
Legend for Figure 2

Cost efficiency as function of sample size $v_n/(10+n)$ when study value corresponds to power in Figure 1 and cost is linear in number of patients with set-up cost equal to the cost of treating 10 patients

21

Discussion of "Simple, Defensible Sample Sizes Based on Cost Efficiency," by

P. Bacchetti, C.E. McCulloch, and M.R. Segal

Peter Müller and Gary L. Rosner,

Department of Biostatistics,

The University of Texas, M. D. Anderson Cancer Center

Houston, TX 77030, U.S.A

Bacchetti, McCulloch, and Segal (BMS) criticize the current emphasis on power when choosing a sample size for research involving human subjects. The authors argue that focusing solely on power considerations may prevent researchers from carrying out studies that would produce useful data. Rather than focusing on power, BMS contend that one should focus on return on investment. BMS give rules that apply when the functions governing the cost and projected return from a study satisfy certain criteria. These criteria basically imply the condition that per-person cost increases with sample size faster than does per-person study reward. Based on this condition, a smaller study may well be more *cost efficient* and, therefore, justifiable, when compared to a larger but less cost efficient study.

We congratulate these authors for bringing forward a new critical approach for evaluating study proposals. We agree that power considerations are not the only measures that we statisticians can bring to bear when considering whether or not a study is worth doing. We have a few points to add.

First, let us reconsider the motivating example BMS describe. Three reviewers are evaluating a proposed randomized clinical study. The study proposes a new form of therapy that the investigators think will cure some of the patients with a disease. It turns out that the disease goes away on its own in 40% of the patients. The proposal hypothesizes that the new treatment may cure one-third of the remaining 60% of patients with the disease. If so, then the overall fraction cured will increase by 20% from 40% to 60%. The proposal calls for enrolling 97 patients per treatment group in order to have 80% power to detect this alternative.

The three reviewers each have different ideas about the propriety of the proposed sample size. While one reviewer agrees with the study investigators, the other two reviewers suggest that the study be powered to have 80% to detect smaller treatment differences. BMS go on to present the sample sizes that correspond to each reviewers wish. They then consider the future value of the study if the treatment is adopted and the three respective treatment effects are true. Using some simple cost considerations and assuming that each alternative treatment effect has probability 25% of being correct, BMS show that the original proposal with the smallest total sample size is the most cost efficient. Since the smallest sample size is most cost efficient, the three reviewers should be willing to accept the proposal as originally presented. In other words, cost efficiency rather than power should be the criterion by which to judge the proposal.

There are some problems with this approach. Consider the paper's Table 1, in which BMS compare the smaller proposed study to the larger studies suggested by two of the reviewers. The table shows that the smaller study is more cost efficient than the larger ones, in that the smaller study is associated with a greater number of expected cures per $100,000 spent. Of course, there is nothing to stop a comparison with a fourth reviewer who feels that a doubling of the cure percentage is the only level of efficacy that is worth detecting. A study with just 22

patients per treatment arm will have 80% power to detect a difference between 0.4 and

0.8. If this study costs just $40,000, and if we still assuming probability of 25% that the treatment is effective at the alternative rate, then this study would lead to 20,000 expected cures per $100,000 spent, making it more cost efficient than even the proposed study. Even if one believes that there is only a 10% chance that the new treatment cures 80% of the patients, the expected number of additional cures is 3200 at $13 per expected cure or 8000 expected cures per $100,000 spent. Even if this study with just 22 patients per treatment arm would cost $100,000— one-half the cost of the proposed study—this small study would lead to 3200 expected additional cures per $100,000, making it even more cost efficient than the proposal's study. One wonders whether any study is so small that it would not pass the BMS criterion of being cost efficient.

Possible alternative strategies can be based on acknowledging additional uncertainties and the nature of the study. For example, one could remove the arbitrariness in the choice of the point alternative by including prior judgement on the unknown parameter. This allows the investigator to consider Bayesian power, based on averaging power with respect to the prior (Spiegelhalter et al., 2004, Section 6.5.3). Another direction of addressing the issues highlighted in the motivating example is to recognize that the study is not carried out in isolation. In most settings the new therapy is competing for resources with other studies. There is a tradition of literature that explicitly addresses such an extended setting and considers a sequence of studies in the same disease area, with comparable eligible population etc. See, for example, Stout and Hardwick (2005) and Ding et al. (2007), and references therein.

Another aspect of the proposed approach is highlighted by the implication on posterior probabilities. The ratio of power and significance level is essentially the Bayes factor (conditioning on the event that the data falls into the rejection region of the test). Under the Bayesian paradigm the Bayes factor summarizes the evidence in favor of the alternative hypothesis. A method that does not protect power might recommend treatments with very small evidence in favor of the recommended treatment.

Basically, BMS substitute one possibly short-sighted criterion, namely power, for another criterion, one that is quite a lot vaguer. It would seem more appropriate to consider utility associated with each action (sample size), accounting for the possible benefits, costs, and weighting each by the a *prior*i probability, uncertainty, or belief associated with each possible treatment effect.

Another problem with this example is that the new treatment will not necessarily be adopted. Adoption likely depends to a large extent on the results of the clinical trial. The clinical community will almost certainly not adopt the new treatment if the study does not show a significant difference. If the minimum difference that is clinically meaningful is a 10% difference in cures, then having high probability of detecting this difference may make sense. Furthermore, not every treatment difference is as likely as the other.

The two examples in the paper do not really help bolster the arguments, either. The first example concerns a presumably clinical study of ten patients with Type I diabetes who will undergo islet cell transplantation. The treatment is expensive, necessitating the small sample size. Thus, the arguments in favor of the study should reflect the potential benefit in terms of knowledge and development of future therapies benefitting patients with Type I diabetes. While BMS indicate that there were criticisms of the proposal's lack of a power calculation, they do not tell us if the grant received funding. If it did, then the reviewers clearly were not considering the worth of a study based solely on power.

The second example involves a proposal to take hair samples from selected individuals in an existing cohort of study participants. It is not clear whether the hair samples are or will be available for study as part of the existing cohort study's protocol. Neither of these examples are randomized comparative clinical trials for which questions of *equipois*e are relevant. Instead,

they are examples of situations in which one would argue that some information is better than no information and, possibly in the second case, the data are already available.

Of course we agree with BMS in their criticism of a rigid insistence on an absolute 80% power requirement. We would rather avoid addressing the identified limitations by replacing them with another set of rigid requirements. Instead we recommend an approach that acknowledges the context of a study and tries to build a justification that is based on the inferential or decision problem that the study addresses. The two actual grant proposals reported by BMS are good examples in which to apply such a strategy. For the diabetes trial with the high per-patient cost and n = 10, one could focus on the nature of the study as an exploratory study that will hopefully provide the information to facilitate a later confirmatory trial. Taking this perspective, we would be led to consider a decision-theoretic approach with a utility function that includes a stylized description of a future confirmatory trial and the possible outcomes. The utility function would include a large reward in the event of a significant outcome of the future trial, as well as a sampling cost for the future trial. The sample size of the future trial would be determined by the usual power argument. The final reward and the sampling cost would be appropriately weighted and averaged with respect to both the posterior predictive distribution at the end of the current trial and the prior predictive for the outcomes of the 10 patients. The averaging with respect to the future sample size would include zero when results from the current study do not support launching a follow-up study. The description sounds far more complicated than the actual implementation. See, for example, the utility function that is used in Rossell et al. (2007) for details.

In the case of the HIV study, one could recognize the aim of the study as scientific learning about the antiretroviral drug levels. If we formalize the goal of scientific learning as reducing the length of a confidence interval, then we could justify the sample size by the need to achieve a desired confidence interval width. As BMS point out in the discussion of other measures of projected value, this view is consistent with their proposal.

In summary, we congratulate BMS on a stimulating and very necessary critical discussion of the unreflected use of traditional power requirements. We agree with many of their arguments but would like to caution against an uncritical use of any single alternative criterion. We expect BMS might agree with this advice.

Additional References

Ding, M., Rosner, G., and Müller, P. (2007). Bayesian Optimal Design for Phase II Screening Trials. Biometrics, to appear.

Rossell, D., Müller, P. and Rosner, G. (2007). Screening Designs for Drug Development. Biostatistics 8, 595-608.

Spiegelhalter, D. J, Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trails and health care evaluation*, John Wiley and Sons: Chichester, UK.

Stout, Q., and Hardwick, J. (2005). Optimal screening designs with flexible cost and constraint structures. Journal of Statistical Planning and Inference 132, 149–162.

Discussion of

"Simple defensible sample sizes based on cost efficiency"

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

James A Hanley and Stanley H Shapiro
Department of Epidemiology, Biostatistics and Occupational Health
McGill University
Montreal

When introducing the notion of sample size estimation we sometimes like to share the story of the statistician whose response to the question: "How many subjects do I need for my study?" was a Socratic-like: "How many bricks does it take to build a wall?" The answer of course depends on what kind of wall one seeks to build.

Bacchetti, McCulloch and Segal (BMS) put forward a new architecture for wall building. Their paper is a welcome call to others to more explicitly consider cost perspectives when planning the size of trials. The blueprints are impressive and the wall is touted as being designed in a cost-effective manner. However, at the end of the day one is left with a nagging concern about its functionality.

Different types of studies have different purposes. For example, a clinical trial might be considered either exploratory or confirmatory depending upon whether it is an early phase study to generate data which will support further investigation or a late phase study designed to corroborate promising preliminary results. The requirement for the latter is typically demanding in that its intent is to affect medical practice and a strong wall is needed to support that enterprise. As Peto Collins and Gray (1995) note, "The medical importance of treatment effects that are only moderate in size implies the need for large-scale randomized evidence (…) Reliable detection or refutation of moderate differences requires negligible biases and small random errors."

Given finite resources, and the fact that few treatments pay for themselves, the minimum 'clinically significant' delta is sometimes considered as "the degree of treatment benefit that, *if it were true*, would make adopting the new treatment worth the extra "cost". The pre-trial assessment as to how large a benefit the treatment <u>might</u> actually produce is a separate matter, arrived at from biological and empirical considerations. In BMS's example in section 2, the investigator's (and reviewer 1's) 'best estimate' -- a 20 percentage point treatment benefit -- was based solely on biological theory; reviewer 2's was also, but was more conservative at 14 points. Reviewer 3, on the other hand, focused on a (minimum) "clinically significant difference," a benefit of 10 percentage points (all agreed on this value) and considered the (higher) estimates provided by all the experts to be irrelevant. Presumably, the projected "costs" of the treatment – as it would be used in practice -- had already been considered when arriving at this minimum clinically significant difference.

The sharp contrast between the viewpoint of reviewer 3 and that of the others should not be taken as a reason to reject both approaches, and thus leave the matter unresolved. Rather, this

25

divergence of viewpoints could be taken as a call to clarify how to proceed in such circumstances. We expect that the rational choice can be shown to be a function of the phase of the investigation, whether there are likely to be other trials of this treatment if the proposed study is "negative" or "inconclusive," and other (non-mathematical) considerations. The consequences arising from uncertainty in the clinical and the biological deltas could be built into the sample size calculations using a Bayesian approach, as well as monitored through the use of interim analysis.

BMS do not further address these issues. Rather the remainder of their paper is used to arrive at and use a metric to justify that a sample size that has less than 50% power against a given delta may be as (or more) defensible as one that has 80% power against this same delta. Their focus is on the costs of doing the study and on robust (omnibus) value functions. We fear that investigators may adopt this focus merely to avoid confronting the unpleasant realities regarding sample sizes. We believe that investigators should plan the size of the study so as to have a high probability of influencing the subsequent course of action.

We agree that the most commonly used sample size procedures do not explicitly consider a study's value or its costs, whereas their procedure attempts to do so. However, the way in which the procedure does this needs to be carefully considered. In particular they state (section 3 second paragraph) " …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 most consumers of a study's results, that is a very narrow perspective, likely too narrow. However, we are told that the neglected costs can be considered to be part of a study's value.

Although BMS state that there are circumstances where they remain tolerant of designs with larger sample sizes (second paragraph, section 6.2) they also state (same location) that their smaller sample size choices should be " … acceptable whenever a larger sample size would be acceptable." Since accrual is typically a challenge in most RCTs, adoption of the proposed methods would seem to raise the potential for a lowest common denominator effect with investigators planning large studies being challenged for doing so. One is reminded of the article by Gehan and Freireich (1974) over 30 years ago which questioned the appropriateness (again largely on the grounds of efficiency) of randomization in certain cancer clinical trials. Meier (1975) recounts that several investigators who were planning what seemed to him to be well-conceived randomized trials, questioned him on whether it had been shown that they were mistaken. Meier went on to show that the allure of the Gehan and Freireich approach was quite limited in that it held only under rather rare circumstances. BMS maintain that their approach holds for any projected value measure that exhibits diminishing marginal returns with increasing sample size. However, since we still seem to be struggling with the issue of how we can improve the translation of research findings into practice, the question of realistic measures of projected value remains an open one for many of us.

BMS tell us their approach is based on "return on investment" but it is not always clear who is investing, what other proposals are competing for the funds, and to whom the returns accrue. In many countries, the funds for health research are raised from the same general taxes that are used to pay for universal health care. However, the sample size and economic considerations are quite different if viewed from the *perspective* of the *developer* of a treatment -- who stands to benefit financially if it is shown to be successful -- than from that of the government sponsor of the study who is also the *payer* of the costs of the treatment if it is adopted.

26

Some less-statistically sophisticated investigators treat sample size considerations as an exercise in creative arithmetic. Although impressed with the mathematical aesthetics of the current paper, we worry about giving such investigators arguments to be even more creative in mathematically defending even smaller sample sizes.

References

Gehan, E.A. and Freireich E.J. (1974). Non-randomized controls in cancer clinical trials. *N Engl J Med*. **290**(4), 198–203.

Meier, P. (1975). Statistics and medical experimentation. *Biometrics* **31**, 511-529.

Peto, R., Collins, R. and Gray, R. (1995). Large-scale randomized evidence: Large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* **48**(1), 23-40

Rejoinder for discussions of "Simple, Defensible Sample Sizes Based on Cost Efficiency"

We thank the discussants for considering our paper and offering their insights, and we appreciate this opportunity to respond to their concerns. They do not appear to question the logic behind our proposals or the case we made for the applicability of conditions B_{min} and B_{root}. The only aspect of our argument that is directly questioned appears to be the relevance of cost efficiency. We also note some concern about seemingly unappealing hypothetical cases and possible negative consequences. Before addressing specific points, we first turn to a cross-cutting issue that seems relevant for many of the concerns about our proposal.

We carefully described our proposed sample sizes as "acceptable" rather than "better" throughout the paper, and we explicitly stated in Section 6.2 that "we advocate more tolerance of different sample size planning approaches, not adoption of the proposals here as a new rigid standard." Nevertheless, the discussants (especially Müller and Rosner) all seem to have some concern about what our proposals might rule out. We therefore re-emphasize here that our proposed methods are not designed to produce the one correct sample size, and they do not imply that all other choices are unacceptable. They provide a simple way to determine a sample size that cannot be validly criticized as too small. This does not imply that any larger sample size must be too large. We explicitly noted in Section 6.2 that both larger *and* smaller studies can be justified, and we now provide Figure 1 to illustrate this for larger studies.
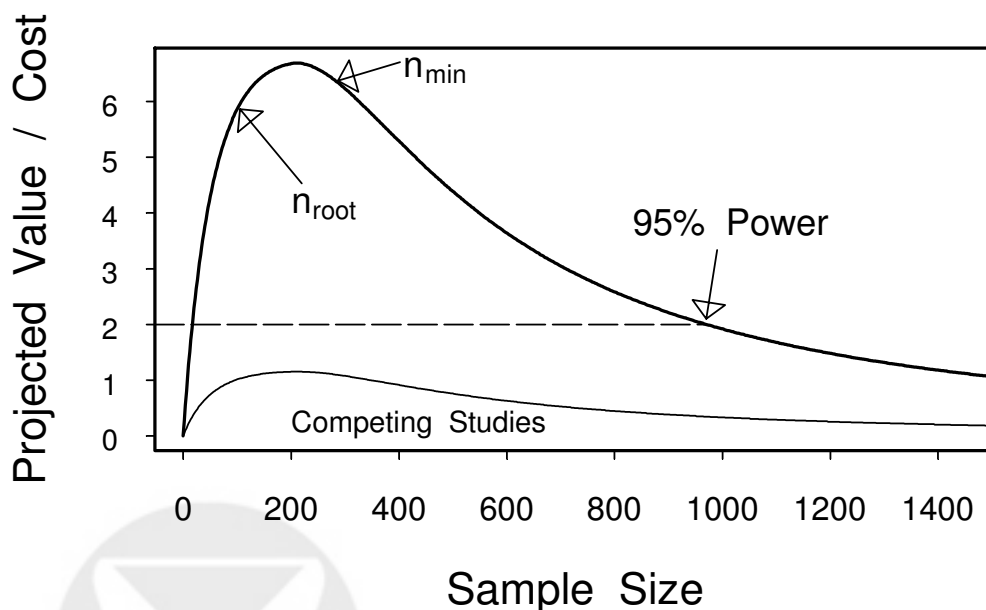

Figure 1. Illustration of how sample sizes much larger than that with optimal cost efficiency can be justifiable. The thick solid line shows the cost efficiencies attained by different sample sizes for a very valuable study, while the "competing studies" line shows the cost efficiencies for other studies on less important topics or based on ideas with less potential. The sample size with 95% power is well beyond the peak in the cost efficiency curve, but it is still projected to produce twice as much value as it costs (horizontal dashed line), a 100% return on investment. It therefore may be justifiable, because the spending to exceed optimal cost efficiency produces more value than saving that money and instead using it to fund the competing studies, which are only projected to return about 15% more value than their cost even at their most cost efficient sample sizes.

Our methods assume that a sample size should be considered acceptable if it is more cost efficient than another sample size that is considered acceptable. This seems unassailable, and we tried to make this very clear and concrete in the example of Section 2: How can obtaining

28

expected cures at a cost of $500 each be acceptable while obtaining them at $279 each is a waste of taxpayer money? Nevertheless, Simon questions the relevance of cost efficiency, citing as support an example where it could be desirable to go beyond the most cost efficient sample size. To clarify why this does not invalidate consideration of cost efficiency, we add some more detail to flesh out an extreme version of what we believe he has in mind.

Suppose the company has only one potential product and is planning a make or break study that will lead either to marketing the product at huge profits or to bankruptcy. They believe that spending 1/10 of what they can afford will produce 99% power, but they have absolutely no other possible use for the remaining money that will produce anything of value to them. Under these extreme (and unrealistic) assumptions, spending the remaining 9/10 of the money on the study could in theory be justified, even though the rate of return can only be about 1/900 of the return on the first 1/10, because failing to do so means that the remaining 9/10 will produce no value at all. But this does not mean that the study using only 1/10 of the funds would be a poor investment and that it is so flawed that it would be better to do no study at all. Our methods only require that better cost efficiency should protect a study from the all-too-common charge of being fatally flawed due to "inadequate" sample size, not that the sample size must be considered better than any other choice.

Our methods do not address the issue Simon raises of maximizing value subject to restrictions. If the resources for a given study have already been determined, then of course the investigators should just do the best they can within that constraint, and the problem of choosing a sample size does not arise (although they would still need to justify their sample size in terms of 80% power under current standards). Adjusting sample sizes of multiple studies so that they fit together optimally is beyond the scope of our methods—and also not typically addressed by investigators or reviewers. Our methods also do not directly address allocating funds among multiple studies, but we believe that they could help improve the cost efficiency of such allocations over current practices. They provide investigators with a way to find and justify reasonably cost efficient sample sizes. Perhaps more importantly, if accepted in practice they would often remove distracting quibbles about assumptions and power from the review process, permitting more attention to focus on more important determinants of whether a study is worth funding, such as the importance of the topic area for public health and/or scientific knowledge, the quality and innovation of the basic idea behind the proposal, and whether the design is vulnerable to serious bias. Our impression is that reviewers often implicitly assume that all of a study's potential value will be realized if power is at least 80% and none of it will be realized if power is any less. This is so inaccurate that it seems bound to distort the allocation of research funds.

We are not sure whether Simon intends his Figures 1 and 2 as a counterexample showing that our methods do not work well. If producing a sample size with only 61% power is considered to invalidate our approach, then it would appear that circular logic will leave us locked into the current absolute requirement for 80% power. We believe instead that his Figures illustrate a situation where 61% power is acceptable because of cost efficiency considerations.

Müller and Rosner extend our example of Section 2 to an even smaller sample size, raising the spectre that our methods can justify any sample size no matter how small. We agree that their very small study could be justified if it would only cost $40,000 ($909 per subject), because this is less per subject than the other sample sizes and Proposition 1 would apply. This is not the case, however, when the very small study costs $100,000 ($2273/subject, much more than the $1031/subject for the study with 97 per arm). Their calculations showing better cost efficiency only apply if the cure rate of 80% versus 40% is assumed to be the correct alternative. If we instead assume 60% versus 40%, then the very small study is only 65% as cost efficient as

the one with 97 per arm. Such reliance on specific assumptions about projected value is exactly what our proposed methods avoid. They remain reliable for *any* assumed alternative.

Müller and Rosner appear to skirt close to defending the status quo when they write, "A method that does not protect power might recommend treatments with very small evidence in favor of the recommended treatment." This appears to be concerned with a decision-theoretic point of view, in which case what matters is the action taken and its consequences, not the evidence for the action taken. We note that the chance of incorrectly recommending the alternative is the same regardless of power—it is $(1-\theta)\alpha/2$ in our notation from Section 4.1—and the real-world consequences of this are not mitigated or exacerbated according to whether the power would have been high or low if the alternative had been true instead of the null. If the concern is instead with how much information the study will provide, then some of the other measures of value we examined are relevant. All of those have the properties needed by our methods.

For our examples in Section 5.2, Müller and Rosner suggest alternative approaches to justifying sample size based on the knowledge to be gained and the fact that some information is better than none. We agree that these are reasonable arguments to make, but we suspect that many investigators would regard using these instead of power-based arguments to be peer-review suicide—and they might be right. Even if reviewers do not regard perceived flaws in sample size justifications as fatal, they may rate proposals lower, often with tangible consequences. Strong defensibility, such as our methods provide, seems to be necessary in the current climate.

Many of the methods that Müller and Rosner advocate appear to be along the lines of the Bayesian MEU/VOI (maximum expected utility/value of information) methods we discussed in Section 1 and the value measures we mentioned in Section 4.2. As stated, we regard such methods as more thoughtful than the prevailing power-based approaches. Unfortunately, such approaches have had seemingly little impact on actual practice. The apparent reluctance of investigators to use such methods may stem from the need for many specific assumptions about benefits, costs, and *a priori* probabilities, along with how vulnerable such assumptions are to disagreement from reviewers. With complete specification of assumptions, cost efficiency could be precisely defined. We argued in Sections 3 and 4 and Web Appendix A, however, that our methods can do reasonably well without relying on such specific assumptions. We do not understand why Müller and Rosner appear to equate this robustness with vagueness. This is a key strength of our approach—it avoids the difficulties and controversies that come with trying to specifically quantify projected value.

Hanley and Shapiro object to how we have handled the cost of the treatment, should it be adopted, in our definition of cost efficiency. We regarded the cost of the treatment as one determinant of its potential net benefit, and its potential net benefit as the source of the study's projected value. This matches the handling of costs and net benefits in a recent retrospective study of the cost effectiveness of clinical trials (Johnston, et al., 2006). We are unclear on exactly how Hanley and Shapiro would modify this framework.

Hanley and Shapiro also worry that our methods could cause large studies to be challenged. This relates to the second paragraph, above. We also remark that studies seeking huge commitments of resources *should* be challenged and very carefully justified, and the effort to thoughtfully perform a MEU/VOI analysis seems most justified for such proposals. They also worry that our methods could help investigators justify "even smaller" sample sizes. We agree that in some circumstances our methods *will* justify smaller sample sizes. Of course, we feel that this is correct for the reasons advocated in our paper. We also note that investigators already have ample incentives to make their sample sizes as large as they can, including a better chance of publishing in more prestigious journals and less chance of having the sample size challenged.

30

We agree that unpleasant realities about sample size planning should be confronted, but we believe that these go well beyond difficulties in specifying an effect size and that they should be confronted by statisticians and reviewers rather than only by investigators who are simply trying to do what they have been taught. One such reality is the wide acceptance of the myth that a study with less than 80% power is fatally flawed. Another is that this misconception forces investigators toward dishonesty by making them claim 80% power instead of disclosing the real reasons for their chosen sample sizes. A prominent article concerning sample size, co-authored by a leader of the CONSORT group (which sets standards for clinical trials), recently defended obtaining a desired sample size via manipulation of assumptions as "an operational solution to a real problem" (Schulz and Grimes, 2005, p. 1351). This suggests to us that facing these realities is quite urgent.

We hope that these remarks will reduce concern about our methods, and we are pleased that none of the discussants advocate an absolute requirement for 80% power in all circumstances. We nevertheless fear that our proposal and the insights from the discussants will have no impact on day to day practices concerning sample size. We therefore end by calling on readers to stop enforcing a rigid 80% power requirement when they review proposals and to be open to using—or at least tolerating—other methods, such as those we have proposed.

Additional References

Johnston, S.C., Rootenberg, J.D., Katrak, S., Smith, W.S., Elkins, J.S. (2006) Effect of a US National Institutes of Health programme of clinical trials on public health and costs, *Lancet*, 367, 1319-1327.
Schulz, K.F., and Grimes, D.A. (2005) Sample size calculations in randomised trials: mandatory and mystical, *Lancet*, 365, 1348-1353.