Bayesian Assessment of Sample Size for Clinical Trials of Cost-Effectiveness

ANTHONY O'HAGAN, JOHN W. STEVENS

The authors present an analysis of the choice of sample sizes for demonstrating cost-effectiveness of a new treatment or procedure, when data on both cost and efficacy will be collected in a clinical trial. The Bayesian approach to statistics is employed, as well as a novel Bayesian criterion that provides insight into the sample size problem and offers a very flexible formulation. ***Key words:*** Bayesian statistics; cost-effectiveness; experimental design; net benefits; pharmacoeconomics; sample size. **(Med Decis Making 2001;21:219–230)**

1. Introduction

Traditionally, health economic evaluation of resource usage has been considered of secondary interest to the clinical objectives of confirmatory clinical trials regarding the efficacy of a treatment, mainly because health economic data are generally not required for the registration of a new treatment. However, Australia[1] and Canada[2] do require information on the cost-effectiveness of a new treatment before granting reimbursement, thereby raising the importance of health economic evaluation in the marketing of treatments. Other countries are also beginning to require information on the cost-effectiveness of treatments before granting reimbursement, including Norway,[3] which makes this a requirement as of January 2002, and NICE[4] in England and Wales.

Confirmatory clinical trials are generally powered to detect clinically relevant treatment effects based upon 1 primary endpoint. In addition, confirmatory clinical trials usually involve many centers across several countries to be able to recruit the required number of patients. This aspect of clinical study design does not usually cause any problems when estimating treatment effects providing that the treatment effect is a constant additive term across centers. However, different clinical practices across centers may lead to very different patterns of resource usage so that the pooling of resource usage and cost data is not straightforward. In practice, this problem is frequently ignored or glossed over. Where it is recognized, resource usage is often summarized separately by center in health economic evaluation.

Together, the increasing requirement to demonstrate the cost-effectiveness of new treatments and the need to consider clinical trial design issues means that health economists must address the ability of a study to provide sufficient information to support a health economic claim.

Much of the literature on comparing the cost-effectiveness of 2 treatments has been based on the incremental cost-effectiveness ratio, both from a theoretical perspective[5–7] and in practical applications.[8–10] However, Stinnett and Mullahy[11] argued for inference and decisions to be based instead on a net benefits approach; Van Hout and others[12] had earlier proposed the C/E Acceptability Curve (CEAC), which plots the probability of a positive net benefit against the threshold cost of a unit increase in efficacy. O'Hagan and others[13] endorsed the use of the CEAC and pointed out that its formulation is intrinsically Bayesian. Practical uses of this approach are now appearing.[14,15]

Analyses of power considerations and the choice of sample size have been published both in the context of subsequent analysis based on incremental cost-effectiveness ratios[16,17] and in the context of a net benefits formulation.[18–20] All of these methods are based on the classical approach of frequentist statistics. There is a considerable

Received 6 June 2000 from the University of Sheffield, United Kingdom (AO'H) and AstraZeneca R&D Charnwood, United Kingdom (JWS). Revision accepted for publication 10 January 2001.

Address correspondence and reprint requests to Dr. O'Hagan: Statistical Services Unit, University of Sheffield, Hicks Building, Sheffield S2 7RH United Kingdom; telephone: 0114 222 3900; fax: 0114 222 3909; e-mail: a.ohagan@sheffield.ac.uk.

literature on Bayesian design of experiments generally[21,22] and some work in the context of clinical trials.[23–25] The purpose of the present article is to give a novel alternative solution using Bayesian statistics and to demonstrate that this leads to a rather more flexible framework for considering both "power" and sample size. Our work is presented in the context of identifying sample size to demonstrate cost-effectiveness, although it is equally applicable to choosing the sample size to demonstrate efficacy.

2. Problem Formulation

2.1 MODEL

We suppose that the study to be designed will comprise a randomized clinical trial in which n_1 patients are given treatment 1 and n_2 are given treatment 2. The sample size problem is to determine n_1 and n_2 under some suitable specification of model and study objectives.

We assume that the data will yield a measure of efficacy and a cost for each patient, and we denote the observed efficacy for patient j given treatment i by e_{ij}, whereas the cost for that patient is denoted by c_{ij}. The range of subscripts is $i = 1,2$ and $j = 1,2 \ldots, n_i$. The expectation of e_{ij} is $E(e_{ij}) = \mu_i$, the population mean efficacy under treatment i. Similarly, $E(e_{ij}) = \gamma_i$ is the population mean cost under treatment i.

An important input to a sample size assessment is the population variances, $Var(e_{ij}) = \sigma_i^2$ and $Var(c_{ij}) = \tau_i^2$, since these determine the amount of information provided by a single observation. However, it is necessary in practice to allow for statistical dependence between the efficacy and cost measures, and so we define $Corr(e_{ij}, c_{ij}) = \rho_i$. Although dependence between efficacy and cost may be more complex than is measured by the correlation coefficient alone, our calculations depend only on the correlation coefficient (because of the linear nature of the net benefit measure introduced below).

We suppose that the objective of the study is to demonstrate that treatment 2 is more cost-effective than treatment 1 if this is the true situation. Let $\Delta_e = \mu_2 - \mu_1$ be the true (population) mean increment of efficacy for treatment 2 compared with treatment 1. Let $\Delta_c = \gamma_2 - \gamma_1$ denote the corresponding mean increment in cost. The cost-effectiveness assessment will be based on the net monetary benefit

$$\beta = K\Delta_e - \Delta_c = K(\mu_2 - \mu_1) - (\gamma_2 - \gamma_1)$$

of changing from treatment 1 to treatment 2, where K is the threshold unit cost. Thus, K is the maximum price (in appropriate monetary units) that the health care provider is prepared to pay to obtain a unit increase in efficacy. Treatment 2 is more cost-effective than treatment 1 if $\beta > 0$. Otherwise, treatment 1 is more cost-effective than treatment 2.

Note that we could have worked instead in terms of the net health benefit β/K, and the analysis that follows would have been unchanged.

2.2 SAMPLE SIZE—THE 2 STAGES

It is helpful to recognize 2 stages in the specification and solution of a sample size problem.

The problem specification in general requires 2 study objectives to be defined, which we will call the *analysis objective* and the *design objective*. The analysis objective specifies what we hope to achieve as a positive outcome of the analysis of whatever data are obtained in the study. Whether the analysis objective is achieved in practice depends on the data that we will obtain. In general, increasing the sample size produces stronger inferences and thereby a better chance of achieving a positive outcome. The design objective specifies in some sense how sure we wish to be of achieving the analysis objective.

The familiar frequentist approach to sample size determination is easily cast in this framework. In the cost-effectiveness situation, a positive outcome would be to demonstrate that treatment 2 is more cost-effective than treatment 1 by rejecting the null hypothesis of no difference with a significance level (P value) less than or equal to α. This is the analysis objective, which is therefore determined by α. Specifying a smaller α implies a more stringent objective that will be harder to achieve and thus will require a larger sample size.

For any given data obtained in the study, the analysis objective will either be achieved or not. The frequentist design objective, then, has 2 components. First, we specify a true effect that we

suppose holds, and in the present context this amounts to specifying a true net monetary benefit $\beta = \beta_0$. Then, we specify the power π that we require to have for detecting this supposed true effect. Thus, π is the probability of achieving the desired analysis objective (of rejecting the null hypothesis) if the true net monetary benefit is β_0.

This separation into an analysis objective (significance at level α) and a design objective (power π to detect a specific effect β_0) may not be familiar in frequentist work but will be very helpful in understanding how our proposed Bayesian solution relates to the frequentist approach.

The 2 stages of problem specification are reflected in the way the problem is solved. We 1st identify what data would lead to achieving the analysis objective, and then we select the sample size to give the design objective's required assurance of obtaining data that would achieve the analysis objective. In the standard frequentist paradigm, the analysis objective determines the critical region of the test. We then choose the sample size to obtain the desired probability (power) for the data to fall in the critical region. Bayesian solutions similarly have 2 stages (sometimes called the posterior and preposterior stages).

2.3 BAYESIAN STUDY OBJECTIVES

We formulate the study objectives as follows:

Analysis objective: We will regard the outcome of the study as positive if the data obtained are such that there is a posterior probability of at least ω that $\beta > 0$, that is, that treatment 2 is more cost-effective than treatment 1.

This Bayesian expression of the analysis objective is analogous to rejecting, at a P value of $\alpha = 1 - \omega$, the null hypothesis that $\beta = 0$ (or that $\beta \leq 0$) in a 1-sided frequentist significance test with alternative hypothesis that $\beta > 0$. Increasing ω is analogous to decreasing the frequentist significance level α and has the same effect of making the analysis objective more stringent.

Design objective: We require the sample size (n_1,n_2) to be large enough so that there is a probability of at least δ of obtaining a positive result.

We will refer to δ as the *assurance*. It is analogous to the frequentist's power π, but there is an important difference between the Bayesian and frequentist design objectives. The frequentist objective requires us also to specify a supposed true effect, in this case a supposed value β_0 of β, such that a specified power π is achieved when the true effect is β_0 or greater. The Bayesian objective does not require this. In a Bayesian analysis we specify a prior distribution for the unknown parameters, which implies a prior distribution for β. The Bayesian assurance δ is a kind of average power, averaged over this prior distribution for β.

Section 6 considers how the assurance δ might be specified in practice and contrasts it with the practical realities of power specification.

3. Prior Information

3.1 DESIGN AND ANALYSIS PRIOR INFORMATION

When considering a Bayesian method, it is often instructive to consider how one might obtain the same results as the standard frequentist solution to the same problem. Typically, the 2 approaches become broadly equivalent if, in the Bayesian method, we specify very weak prior information.

Thus, if we specify very weak prior information about the parameters in the model, then the posterior distribution will be based almost entirely on the evidence in the data. We will then find that the Bayesian posterior probability $1 - \omega$ that $\beta \leq 0$ becomes exactly the same as the frequentist P value for the 1-sided test of H_0:$\beta \leq 0$ against the alternative that $\beta > 0$. So the 2 approaches become equivalent at the analysis stage. However, it is not possible to obtain convergence at the design stage in this way. If we try to specify weak prior information at the design stage, then we will be allowing substantial probabilities that β is extremely large, either positive or negative, and it will become impossible to satisfy any realistic design objective. It is generally recognized that, to design any kind of experiment, it is necessary to specify nontrivial prior information. The frequentist approach effectively goes to the other extreme by supposing a

particular true effect. For the Bayesian analysis to mimic the standard frequentist solution at the design stage, we would need to specify very *strong* prior information that implies we are *certain* that $\beta = \beta_0$.

Now, it is not our intention simply to reproduce standard frequentist methods. On the contrary, we believe that a Bayesian approach has many benefits, not the least of which is to obtain stronger and more appropriate inferences through the use of substantive prior information. However, consideration of how we might mimic a frequentist analysis raises some important issues.

The most natural Bayesian analysis would use the same prior beliefs about the parameters in both stages of the calculation. This is the basis of almost all previous Bayesian work on all kinds of experimental design problems.[22] The analyst should honestly express his or her prior opinion in the form of a prior distribution for the parameters and then use this consistently throughout the analysis. But this raises the question of who the analyst is, or in other words, Whose prior information is to be used? We believe that the answer might very well be different in the analysis and design stages.

The analysis of a clinical trial, whether simply to prove efficacy or to demonstrate cost-effectiveness, will be presented to regulatory authorities or decision makers external to the company conducting the trial. (We envision principally the case of a pharmaceutical company wishing to demonstrate cost-effectiveness of a drug, but our use of "the company" is intended to convey a wider context.) To present a Bayesian analysis in which the company's own prior beliefs are used to augment the trial data will in general not be acceptable to an external agency. The appropriate form of prior specification in order for a Bayesian analysis to be acceptable to regulatory authorities is still a matter of debate. However, it seems likely that an impartial or even a pessimistic prior formulation may be needed until Bayesian methods become more accepted. As advocates of the Bayesian approach, we suggest that a prior distribution that represents the accepted knowledge base of experts in the relevant field might be most appropriate, and we hope that regulatory authorities and decision makers will soon recognize the value of incorporating genuine prior knowledge in a fully Bayesian analysis. However, we acknowledge that, to be sure of presenting an acceptable case in today's context, a company may also wish to present a non-Bayesian analysis corresponding to a weak prior specification.

Whether we adopt some form of consensus or cautious prior distribution for the analysis of the data, or whether one is required to use a weak prior specification at the analysis stage, the design stage is quite different. The choice of sample size is primarily a matter for the company, and it may be argued that at this stage the company should use all its prior information. We recognize that the design must also be approved by an ethics committee, but it is clear that the company's expectations of how the data are likely to turn out will be a valid input to the design.

We therefore develop a novel Bayesian solution based on assuming different prior beliefs at the analysis and design stages. Although some authors have adopted a similar approach of using different prior formulations at the 2 stages,[26,27] their justification and meaning is quite different from ours. No such distinction has previously been made in the design of clinical trials.

3.2 SOME NOTATION

To express our results concisely, it will be convenient to adopt a vector formulation. We define the vector of true parameter values to be $\boldsymbol{\xi} = (\mu_1,\gamma_1,\mu_2,\gamma_2)^T$. If we also define a vector $\mathbf{a} = (-K,1,K,-1)^T$, then we have $\beta = \mathbf{a}^T\boldsymbol{\xi}$.

We now define the 2 prior distributions. Let the prior expectation of $\boldsymbol{\xi}$ be $\mathbf{m}_a$ according to the analysis prior and $\mathbf{m}_d$ according to the design prior. Similarly, let the prior variance matrix of $\boldsymbol{\xi}$ be $\mathbf{V}_a$ according to the analysis prior and $\mathbf{V}_d$ according to the design prior.

In terms of our interpretation of the standard frequentist methods, we first achieve a weak prior distribution at the analysis stage by making the prior variances become extremely large, and this is represented in the limit by $\mathbf{V}_a^{-1} = \mathbf{0}$, a zero matrix. In the limit, it is not necessary to specify $\mathbf{m}_a$. Second, the very precise prior specification at the design stage is obtained by setting $\mathbf{m}_d$ to the assumed parameter values and $\mathbf{V}_d = \mathbf{0}$. The zero variance matrix here says that these assumed parameter values are the true values.

Our formulation thus includes the standard frequentist analysis as a special case. It includes

also the usual Bayesian experimental design approach by setting $\mathbf{m}_a = \mathbf{m}_d$ and $\mathbf{V}_a = \mathbf{V}_d$, but is much more general than either of these. In particular, it allows for the analysis prior specification to be weak, impartial or pessimistic according to the needs of an outside agency while allowing also for the company's genuine knowledge to be employed in the design prior.

We let $\bar{\mathbf{x}}$ denote the vector of sample means arranged in the corresponding order to the elements of $\boldsymbol{\xi}$. That is, its 1st element is the sample mean efficacy for treatment group 1, its 2nd is the sample mean cost for treatment group 1, and its 3rd and 4th elements are the sample mean efficacy and cost for treatment group 2.

Much of the statistical theory of cost-effectiveness analysis from clinical trial data has assumed that the sample mean costs and mean efficacies contained in $\bar{\mathbf{x}}$ are jointly normally distributed. Even if the individual observations are not normally distributed, this is justified as an approximation in sufficiently large samples by the central limit theorem. We shall adopt the same assumption here, assuming also approximate normality of prior distributions, and defer discussion of the realism of this assumption to Section 6.

4. Derivation of Sample Size

4.1 GENERAL BAYESIAN SOLUTION

It is shown in the appendix that the analysis objective leads to the following condition for a positive outcome:

$$\mathbf{a}^T\mathbf{V}^*(\mathbf{V}_a^{-1}\mathbf{m}_a + \mathbf{S}^{-1}\bar{\mathbf{x}}) \geq Z_{1-\omega}\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{a}}, \qquad (1)$$

where the matrix $\mathbf{V}^*$ is the posterior variance matrix of $\boldsymbol{\xi}$, given by

$$\mathbf{V}^* = (\mathbf{V}_a^{-1} + \mathbf{S}^{-1})^{-1}, \qquad (2)$$

where the matrix $\mathbf{S}$ is the sampling variance matrix of $\bar{\mathbf{x}}$, given by

$$\mathbf{S} = \begin{pmatrix} \sigma_1^2/n_1 & \rho_1\sigma_1\tau_1/n_1 & 0 & 0 \\ \rho_1\sigma_1\tau_1/n_1 & \tau_1^2/n_1 & 0 & 0 \\ 0 & 0 & \sigma_2^2/n_2 & \rho_2\sigma_2\tau_2/n_2 \\ 0 & 0 & \rho_2\sigma_2\tau_2/n_2 & \tau_2^2/n_2 \end{pmatrix}, \qquad (3)$$

and where Z_α is the upper 100α percentile of the standard normal distribution (so that, for instance, $Z_{0.025} = 1.96$). It is primarily through $\mathbf{S}$ that the sample sizes n_1 and n_2 appear in the calculations.

This condition depends on the sample data $\bar{\mathbf{x}}$, whose distribution depends on $\mathbf{S}$, and hence on n_1 and n_2. At the design stage, we need to choose n_1 and n_2 so that with probability δ we will obtain data satisfying eq. (1). It is shown in the appendix that n_1 and n_2 must therefore satisfy the condition

$$\mathbf{a}^T\mathbf{V}^*(\mathbf{V}_a^{-1}\mathbf{m}_a + \mathbf{S}^{-1}\mathbf{m}_d) \geq Z_{1-\omega}\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{a}} + Z_{1-\delta}\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{S}^{-1}(\mathbf{V}_d + \mathbf{S})\mathbf{S}^{-1}\mathbf{V}^*\mathbf{a}}. \qquad (4)$$

The problem is to choose n_1 and n_2 so as to give a matrix $\mathbf{S}$, and hence also a matrix $\mathbf{V}^*$ via eq. (2), to satisfy this inequality.

The condition is complex, and the matrix formulation may be particularly daunting to those unused to such notation. Before considering how it might be solved to yield sample sizes n_1 and n_2 in practice, it will be helpful to gain some understanding of the various terms in eq. (4).

On the right-hand side are 2 terms that clearly arise from the 2 stages of the solution. Each has the form of a normal Z significance point multiplied by a square root that corresponds to a standard error. Both standard errors will tend to decrease as the sample sizes n_1 and n_2 are increased. It is in this way that the condition determines sample sizes: They need to be large enough to reduce the right-hand side sufficiently to become smaller than the left-hand side. (The 1st standard error, $\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{a}}$, will actually go to 0 if the sample sizes go to infinity, but the 2nd will not, and this fact is related to a result that will be shown in subsection 4.4.)

The coefficients $Z_{1-\omega}$ and $Z_{1-\delta}$ are determined by ω and δ, which come from the analysis and design objectives, respectively. If we wish to strengthen either objective, then ω or δ will increase, and this will in turn increase either $Z_{1-\omega}$ or $Z_{1-\delta}$. In either case, this will increase the right-hand side of eq. (4), and so will necessitate larger sample sizes.

Turning now to the left-hand side of the inequality, we see that it involves the prior expectations $\mathbf{m}_a$ and $\mathbf{m}_d$ of $\boldsymbol{\xi}$, according to the analysis and design priors respectively. In fact, $\mathbf{a}^T\mathbf{m}_a$ and $\mathbf{a}^T\mathbf{m}_d$ are the prior expectations of the net benefit β, again according to the 2 priors, the left-hand side of eq. (4) is a kind of weighted average of these 2 expectations. If we genuinely expect the net benefit of treatment 2 over treatment 1 to be large, then the left-hand side will be large, and so we do not need such a large sample size to make the right-hand side small enough to satisfy the condition. Conversely, larger sample sizes will be needed if we expect treatment 2 to be of marginal net benefit.

Although both sides of this inequality are rather complicated expressions, they are straightforward to compute in practice. The matrices in question are only of dimensions 4×4, and so their inversion and multiplication involves little computation. Notice that we are assuming the sampling variances and correlations known, because it is assumed that we know the matrix $\mathbf{S}$. This assumption is discussed in Section 6.

Solution of eq. (4) will in general require an iterative search. However, we should note that unless a further condition is set there will not generally be a unique solution. Two natural additional constraints are, 1st, to require equal sample sizes in the 2 treatment groups via the condition $n_1 = n_2$ or, 2nd, to find the solution that minimizes the total sample size $n_1 + n_2$.

The remainder of this section will explore the implications of this general solution in some particular cases.

4.2 FREQUENTIST SOLUTION

If the analysis prior is weak, that is, $\mathbf{V}_a^{-1} = \mathbf{0}$, we have $\mathbf{V}^* = \mathbf{S}$. It is now easily shown that eq. (1) becomes exactly the frequentist 1-sided $100(1 - \omega)\%$ test of the null hypothesis that $\beta = 0$ (or $\beta \leq 0$) against the alternative hypothesis that $\beta > 0$. So, this weak analysis prior gives agreement between the Bayesian and frequentist approaches at the analysis stage, as anticipated.

If, in addition, we specify precisely $\boldsymbol{\xi} = \mathbf{m}_d$ at the design stage through $\mathbf{V}_d = \mathbf{0}$, we find that eq. (4) reduces to

$$\mathbf{a}^T\mathbf{m}_d \geq (Z_{1-\omega} + Z_{1-\delta})\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}. \tag{5}$$

If we now require equal sample sizes, applying the constraint that $n_1 = n_2$, we can solve eq. (5) explicitly. Letting the common sample size be n, the solution is found to be given by

$$n \geq \frac{(Z_{1-\omega} + Z_{1-\delta})^2\,\mathbf{a}^T\mathbf{S}_1\mathbf{a}}{(\mathbf{a}^T\mathbf{m}_d)^2}, \tag{6}$$

where $\mathbf{S}_1$ is the single-observation variance matrix obtained by setting $n_1 = n_2 = 1$ in eq. (3) for $\mathbf{S}$. The required sample size in each treatment group is the result of rounding up the right-hand side of eq. (6) to an integer.

This is exactly the solution already given independently for the frequentist sample size problem by Briggs and Tambour[19] and (in the special case of equal sampling variances and covariances in the 2 treatment groups) by Laska and others.[20]

We can also consider minimizing the total sample size $n_1 + n_2$ rather than first requiring $n_1 = n_2$. However, it will often be assumed in practice that $\sigma_1^2 = \sigma_2^2$, $\tau_1^2 = \tau_2^2$, and $\rho_1 = \rho_2$, and then it is easy to demonstrate that the optimal solution has equal sample sizes and hence that eq. (6) is again the required solution.

4.3 STRONG ANALYSIS PRIOR

If the prior information expressed in the analysis prior is sufficiently strong, then the required sample size may be 0. The reason is that then the prior information alone is enough to produce a positive outcome without needing any sample data at all. This arises when

$$\mathbf{a}^T\mathbf{m}_a \geq Z_{1-\omega}\sqrt{\mathbf{a}^T\mathbf{V}_a\mathbf{a}}.$$

It is clearly undesirable to express analysis prior information this strong. In the context where a clinical trial is contemplated, and in particular when the trial data are needed to formulate a case to an outside agency, it is unrealistic to suppose that prior information could legitimately be so strong. The situation can never arise with the conventional formulation of weak prior information via $\mathbf{V}_a^{-1} = \mathbf{0}$.

A referee has pointed out that, with a strong prior, the role of equipoise, usually considered as a necessary condition for randomization, becomes a difficult problem.

4.4 WEAK DESIGN PRIOR

Conversely, a problem may arise if the information expressed in the design prior is not sufficiently strong. The problem in this case is that even infinite sample sizes may not be large enough to achieve the required assurance δ of a positive outcome. This arises if the prior probability that treatment 2 is genuinely more cost-effective than treatment 1, according to the design prior, is less than δ.

To understand this point, it is helpful to consider the effect of infinite sample sizes. With infinite data we will find out exactly the true values of parameters, and hence we will find out whether treatment 2 is truly more cost-effective than treatment 1. We will be able to *demonstrate* that treatment 2 is more cost-effective if, and only if, it is *truly* more cost-effective. If, before I see any data, my prior probability that this is *truly* the case is 70%, say, then with infinite data I have exactly a 70% chance of being able to *demonstrate* that it is the case. The prior probability that $\beta > 0$ is therefore an upper bound on the achievable assurance δ. There is no point in trying to design a trial so that I can be, say, 90% sure of being able to demonstrate that treatment 2 is more cost-effective, if I am only 70% sure that it really is more cost-effective.

The prior probability in question is

$$P(\beta > 0) = \Phi(\mathbf{a}^T\mathbf{m}_d \big/ \sqrt{\mathbf{a}^T\mathbf{V}_d\mathbf{a}}) \qquad (7)$$

according to the design prior, where Φ denotes the standard normal distribution function. We cannot ask for δ to exceed this. Equivalently, we must have

$$\mathbf{a}^T\mathbf{m}_d > Z_{1-\delta}\sqrt{\mathbf{a}^T\mathbf{V}_d\mathbf{a}}. \qquad (8)$$

Again, the problem does not arise in the frequentist scenario, where we set $\mathbf{V}_d = \mathbf{0}$, since then the right-hand side of eq. (8) is 0. We therefore have a solution with finite sample sizes, provided only that the parameter values $\mathbf{m}_d$ are such that the supposed true effect $\beta_0 = \mathbf{a}^T\mathbf{m}_d$ is positive.

Table 1 • Sample Sizes with Frequentist Analysis

K	4000	5000	7000	10,000	20,000
n	1040	762	559	457	374

5. Examples

Briggs and Tambour[19] report an example previously analyzed by Briggs and Gray.[18] They present a frequentist analysis in which they assume that the proposed new treatment or procedure has a mean increase in efficacy of $\Delta_e = 0.8$ and a mean increase in cost of $\Delta_c = 1200$. They also assume that the sample variances and correlation are $\sigma_1^2 = \sigma_2^2 = 4.04^2$, $\tau_1^2 = \tau_2^2 = 8700^2$ and $\rho_1 = \rho_2 = 0$. These values give our matrix $\mathbf{S}_1$. The assumed values of Δ_e and Δ_c are enough to determine $\mathbf{a}^T\mathbf{m}_d = K\Delta_e - \Delta_c = 0.8K - 1200$ once K is specified. They set $\omega = 0.975$ (equivalent to their requirement of 5% significance in a frequentist 2-sided test) and one of the values that they consider for the power is 70%. It is now straightforward to apply eq. (6) for $\delta = 0.7$ and a range of values of K to obtain, for instance, the values in Table 1.

We now consider the same example but explore the extra flexibility offered by the Bayesian formulation of the sample size problem.

We first suppose that the analysis prior distribution is to be weak, representing the situation where a frequentist analysis (or equivalently a Bayesian analysis with noninformative prior distribution) is to be performed at the analysis stage. We will, however, consider incorporating proper prior information at the design stage. We suppose that the design prior mean is $\mathbf{m}_d = (5{,}6000{,}6.5{,}7200)^T$. The prior expectation of Δ_e is therefore 1.5, rather larger than the assumed value in the Briggs and Tambour [19] analysis, but the prior expectation of Δ_c remains at 1200. We set the design prior variance matrix to

$$\mathbf{V}_d = \begin{pmatrix} 4 & 0 & 3 & 0 \\ 0 & 10^7 & 0 & 0 \\ 3 & 0 & 4 & 0 \\ 0 & 0 & 0 & 10^7 \end{pmatrix}.$$

Thus, we have a variance of 4 for the mean efficacy under each treatment, but the covariance of 3 implies that the variance of Δ_e is $2 \times (4 - 3) = 2$. This gives a prior probability of $\Phi(1.5/\sqrt{2}) = 0.8555$, or 85.5%, that treatment 2 is more effective than treatment 1. This defines substantive prior information about mean efficacies. In contrast, it is usual to have less prior information about costs, and this is reflected in the relatively large variances for mean costs. Similarly, lack of knowledge about how mean costs might be related suggests that we give them 0 covariances.

Noting the discussion in subsection 4.4, consider now the prior probability, eq. (7), that treatment 2 is actually more cost-effective than treatment 1. For various values of K, we find the figures in Table 2.

The limit as K goes to infinity is 85.5%, the probability that treatment 2 is more effective, according to the design prior distribution. We can therefore achieve the previously set value of $\delta = 0.7$ for all K in the range of interest. The required sample sizes if we specify $n_1 = n_2$, for suitable values of K, are given in Table 3. In fact, since both the prior information and the sampling variances are symmetric between the 2 treatments, the values minimizing $n_1 + n_2$ also satisfy $n_1 = n_2$.

We see that for lower values of K, we are not so sure a priori that treatment 2 is more cost-effective than treatment 1, the desired δ is not much more than the maximum achievable, and sample sizes are larger than under the frequentist analysis. However, for larger values of K, we have smaller sample sizes than in Table 1.

Figure 1 plots the assurance δ against n in the case $K = 5000$. It is plotted on a logarithmic scale for n for greater clarity. As $n \to 0$, the assurance tends to the prior probability, according to the analysis prior, that $\beta > 0$. Because we have a weak analysis prior distribution, this left-hand limit is 0. To the right, as $n \to \infty$, the limit is the probability that $\beta > 0$ according to the design prior, which is given in Table 2 for $K = 5000$ as 0.774. The line crosses the required $\delta = 0.7$ at $n = 1048$, the sample size given in Table 3.

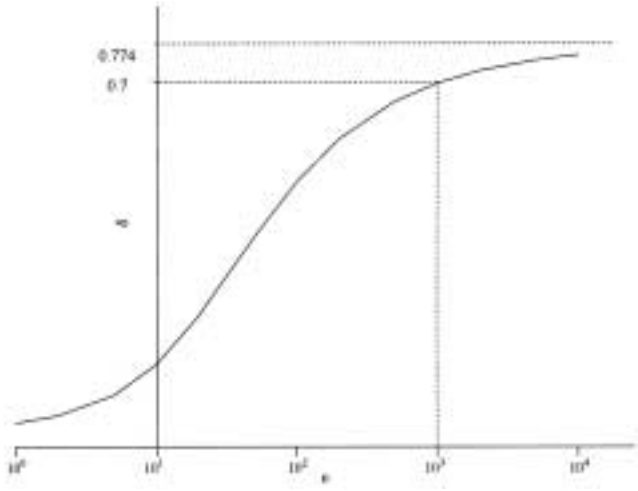

FIGURE 1. Assurance δ achieved by sample sizes $n_1 = n_2 = n$ for the example with weak analysis prior and $K = 5000$.

Sample sizes will be reduced further if we incorporate prior information also into the analysis stage. The values in Table 4 result when we make the analysis prior the same as the design prior.

6. Discussion

6.1 FLEXIBILITY OF THE BAYESIAN APPROACH

We have presented a Bayesian approach to the design of a clinical trial for the purposes of cost-effectiveness evaluation. To formulate this approach, we were led to recognize the importance of allowing different expressions of prior information at the design and analysis stages. As a result, our results encompass those previously derived following the frequentist approach to statistics but are considerably more flexible.

A key benefit of the Bayesian approach is avoiding the need to fix arbitrary parameter values, leading to an assumed value β_0 of the true net

Table 2 • Maximum Values for δ in Bayesian Analysis with Weak Analysis Prior

K	4000	5000	7000	10,000	20,000
$P_d(\beta > 0)$	74.7%	77.4%	80.4%	82.4%	84.3%

Table 3 • Sample Sizes in Bayesian Analysis with Weak Analysis Prior

K	4000	5000	7000	10,000	20,000
$n_1 = n_2$	2543	1048	541	382	285

Table 4 • Sample Sizes in Bayesian Analysis with Proper Analysis Prior

K	4000	5000	7000	10,000	20,000
$n_1 = n_2$	901	513	348	279	224

monetary benefit β at which a specific power π is desired. Practical guidance on the specification of π and β_0 is far from clear, and practice evidently varies widely. It is generally acknowledged that the power should be "high," but there is no consistency in assigning it a value. Concerning β_0, we can note that in the context of a conventional clinical trial, it is often said that the assumed effect should be the smallest "clinically meaningful" effect. But in cost-effectiveness studies, what would be a clinically meaningful net monetary benefit? In practice, using the smallest clinically meaningful effect and high power will usually lead to prohibitively large sample sizes, and various adjustments may be applied by the company. Simon,[28] for example, says

> The sample size is usually established to obtain a specified statistical power for rejecting the null hypothesis when a specified alternative hypothesis is true. This formalism is often abused by specifying unrealistically large alternative hypotheses for the power calculation. This is done in order to attempt to justify doing a trial by an organization that does not have sufficient patient accrual potential to conduct independent clinical trials.

It is true that in the Bayesian approach, one must instead specify a prior distribution, and that this leads to a design objective expressed in terms of the assurance probability δ rather than the power π. δ is an *overall* power requirement, averaged over the possible values of those parameters according to our prior beliefs. However, we believe that this formulation is both more natural and less arbitrary than the frequentist approach.

Consider, for instance, a new drug that could be highly beneficial and hence also highly profitable for the company, but such that the company's experience to date suggests that it is at least as likely to turn out to be *less* cost-effective than the comparator drug as to be more cost-effective. The frequentist approach now requires the company to imagine parameter values for which its drug would be clearly more cost-effective and then to say just how good a chance it wishes to have of demonstrating the drug to be more cost-effective in the (unlikely) event that these are the true values. It is far more natural to express actual beliefs about the parameters in a prior mean vector $\mathbf{m}_d$ and a prior variance matrix $\mathbf{V}_d$, even if it is recognized that together these give a good chance a priori that the new drug is actually less cost-effective than the comparator. And it is natural then to say what *overall* assurance δ is required of demonstrating that the new drug is more cost-effective, acknowledging that this chance cannot be as high as is conventionally set in frequentist power calculations. In such a context, even a small chance of demonstrating improved cost-effectiveness might be worthwhile for the company. It is natural for the company to think in terms of overall probability of success rather than in terms of an unrealistically high power requirement attached to a true outcome with unrealistically good parameter values.

6.2 THE INTERPRETATION OF ASSURANCE

The above discussion makes it clear how the Bayesian notion of assurance is actually quite different from the frequentist power. The conventional wisdom is that high power is needed, and a confirmatory trial designed to achieve power much less than 80% would generally seem inadequate. But the power is relative to the parameter values that we choose to assume for the purpose, and it is not unknown for power calculations to be "reverse engineered," that is, for the sample size to be chosen by other means, perhaps economic, and then assumed parameter values to be chosen at which the power would be considered acceptable.

In the Bayesian framework, assurance can and should be selected on economic grounds, reflecting the relative benefit to the company of being able to demonstrate cost-effectiveness versus the cost of the trial itself. We should stress that there is nothing irrational or unethical in this.

The level of assurance chosen may in practice be much lower than conventional wisdom of power would suggest. In the case of Figure 1, we see that a sample size of more than 1000 (in each treatment group) is needed to achieve 70% assurance but that greater than 50% assurance would be obtained with a sample size of 100. Depending on the cost per patient of running the trial, the company may very well be satisfied with 50% assurance and a small, cheap trial. A compromise of $n_1 = n_2 = 220$, which gives $\delta = 0.6$, could be even more attractive.

Through the Bayesian formulation, the trial design can be genuinely responsive to the priorities of the company without compromising the wider needs of the regulatory environment at the analysis stage.

6.3 SAMPLE SIZE VERSUS POWER

At this point, we should note that our analysis does not follow the pure Bayesian approach of Lindley.[22] A fully Bayesian analysis would not specify δ as the design objective but would instead consider the relative costs of sampling versus making wrong decisions and set a design objective of minimizing the expected cost. Such a decision-theoretic analysis would be the ideal, and the field of health economics, which has already grappled with balancing costs against effectiveness of drugs and treatments, might surely be able to address the easier task of balancing sampling costs against the resulting net monetary benefits of good or bad cost-effectiveness decisions. Such a balancing is hinted at in the preceding discussion of assurance. In practice, though, the realities of clinical trial design would not readily admit such an analysis.

Although demonstrating cost-effectiveness is becoming increasingly important, the planning of confirmatory clinical trials is still typically done primarily to demonstrate efficacy, with collection of cost data being in some sense a side consideration. In this context, it is likely that sample size will be chosen simply for the purposes of giving sufficient power to show efficacy. It is nevertheless very useful for the health economist to be able to show the consequences of the selected sample sizes in the context of demonstrating cost-effectiveness. Instead of using the design criterion shown in eq. (4) to derive suitable sample sizes n_1 and n_2, we simply calculate the assurance $\delta = \Phi(b)$, where b is given in the appendix by eq. (11) for the already chosen sample sizes. Such a calculation may act as a warning against going ahead with a trial design that gives far too little value in cost-effectiveness terms or, conversely, it may confirm that the chosen sample size would also be adequate to demonstrate cost-effectiveness.

These discussions emphasize the importance of using the company's best understanding of likely efficacies and costs when formulating the design prior. Although we strongly believe that genuine prior information should also play a substantial role in the analysis stage and in presenting that analysis to regulatory authorities and decision makers, our methods allow for the analysis prior to be different and even to be essentially noninformative.

6.4 THE NORMALITY AND KNOWN VARIANCE ASSUMPTIONS

We consider now the assumption made throughout Section 4 that (a) both cost and efficacy data are normally distributed and (b) the sample variances and correlations are known.

The assumption of normality will clearly be technically inappropriate in practice. Efficacy data are often binary responses or failure times, for instance, and costs are almost always highly skewed. It would in principle be possible to formulate the sample size problem more generally to recognize a range of distributional assumptions. Existing frequentist theory relating to analysis of cost-effectiveness from clinical trial data is based either on normal distributions or on nonparametric analysis by application of bootstrap methods.

We have recently presented a more general framework for a Bayesian analysis allowing for more realistic distribution assumptions for the data and showing the impact that these have on the posterior probability of cost-effectiveness.[29] Nevertheless, it is not clear to us whether the extra complexity in Bayesian sample size determination would produce meaningfully different sample sizes.

The normality assumption is usually justified not by arguing that the raw data will be normally distributed but on the basis that sample means will be approximately normal because of the central limit theorem. Similarly, although the true distributions may not be at all normal, we might suppose that sample sizes derived assuming normality would be approximately valid under other assumptions. In Bayesian statistics, this may be supported by reference to the Bayes linear methods,[30] which are analogous to the classical Gauss-Markov theory. The analysis stage outlined in the appendix, leading to eq. (1), can be justified approximately by Bayes linear theory without assuming normally distributed data.

In spite of these comments, we feel that further research is needed based on alternative models to ascertain the extent to which sample sizes may be robust to the underlying distributions.

Finally, the assumption of known variances and correlations is expressed in assuming that we know the matrix **S** given by eq. (3). In effect, we will specify these as prior estimates (and in practice they are specified similarly in frequentist sample size calculations). Ignoring uncertainty about these values will probably lead to recommendations of somewhat smaller sample sizes than should really be required. However, the difference is unlikely to be serious in practice.

We gratefully acknowledge the helpful suggestions of 2 referees on an earlier version of this paper. Their comments have contributed particularly to the presentation and clarity of the final version.

APPENDIX

We present here some mathematical derivation of the results given in subsection 4.1.

The posterior distribution of $\boldsymbol{\xi}$ is (approximately) normal. Its mean vector is

$$\mathbf{m}^* = E(\boldsymbol{\xi}|\mathbf{x}) = \mathbf{V}^*(\mathbf{V}_a^{-1}\mathbf{m}_a + \mathbf{S}^{-1}\bar{\mathbf{x}}). \quad (9)$$

In this formula, the symbol **x** represents all the sample data. We have defined $\mathbf{m}_a$ and $\mathbf{V}_a$ to be the mean vector and variance matrix of the analysis prior distribution. The sampling variance matrix of $\bar{\mathbf{x}}$ is the matrix **S** shown in eq. (3). Then, the posterior variance matrix of $\boldsymbol{\xi}$ is the matrix $\mathbf{V}^*$ in eq. (2).

It follows that the posterior mean and variance of $\beta = \mathbf{a}^T\boldsymbol{\xi}$ are

$$E(\beta \mid \mathbf{x}) = \mathbf{a}^T\mathbf{m}^*, \quad Var(\beta \mid \mathbf{x}) = \mathbf{a}^T\mathbf{V}^*\mathbf{a}.$$

Therefore, the criterion for a positive outcome, that is, that $P(\beta > 0 \mid \mathbf{x}) \geq \omega$ corresponds to the condition

$$\mathbf{a}^T\mathbf{m}^* \geq Z_{1-\omega}\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{a}}. \quad (10)$$

Substituting the formula in eq. (9) for $\mathbf{m}^*$ into this equation yields eq. (1) in the main text.

At the design stage, we do not know what data will be observed, and in particular we do not know $\bar{\mathbf{x}}$. According to the design prior distribution, the (unconditional, or preposterior) distribution of $\bar{\mathbf{x}}$ is normal with mean and variance

$$E(\bar{\mathbf{x}}) = \mathbf{m}_d, \quad Var(\bar{\mathbf{x}}) = \mathbf{V}_d + \mathbf{S}.$$

(The variance derives from the random variation in the data, given by **S**, and uncertainty about the true value of $\boldsymbol{\xi}$, given by $\mathbf{V}_d$.) It follows that $\mathbf{a}^T\boldsymbol{m}^*$ is also normally distributed, with

$$E(\mathbf{a}^T\mathbf{m}^*) = \mathbf{a}^T\mathbf{V}^*(\mathbf{V}_a^{-1}\mathbf{m}_a + \mathbf{S}^{-1}\mathbf{m}_d),$$

$$Var(\mathbf{a}^T\mathbf{m}^*) = \mathbf{a}^T\mathbf{V}^*\mathbf{S}^{-1}(\mathbf{V}_d + \mathbf{S})\mathbf{S}^{-1}\mathbf{V}^*\mathbf{a}.$$

Hence, the prior probability (according to the design prior) that the condition shown in eq. (10) will hold is $\Phi(b)$, where Φ denotes th
e standard normal distribution function and where

$$b = \frac{\mathbf{a}^T\mathbf{V}^*(\mathbf{V}_a^{-1}\mathbf{m}_a + \mathbf{S}^{-1}\mathbf{m}_d) - Z_{1-\omega}\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{a}}}{\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{S}^{-1}(\mathbf{V}_d + \mathbf{S})\mathbf{S}^{-1}\mathbf{V}^*\mathbf{a}}}. \quad (11)$$

The power requirement for the design stage specifies that this probability should be at least δ, that is, $b \geq Z_{1-\delta}$. Using eq. (11) and multiplying both sides of the inequality by $\sqrt{\mathbf{a}^T\mathbf{V}^*\mathbf{S}^{-1}(\mathbf{V}_d + \mathbf{S})\mathbf{S}^{-1}\mathbf{V}^*\mathbf{a}}$ (which is necessarily positive) results in the condition shown in eq. (4) in the main text.

References

1. Australia Commonwealth Department of Human Services and Health. Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: Including Major Submissions Involving Economic Analysis. Canberra: Australia Government Publishing Service, 1995.
2. Canadian Coordinating Office for Health Technology Assessment. Guidelines for Economic Evaluation of Pharmaceuticals: Canada (2nd ed.). Ottawa: Canadian Coordinating Office for Health Technology Assessment, 1997.
3. Norwegian Medicines Control Authority. Norwegian Guidelines for Pharmacoeconomic Analysis in Connection with Application for Reimbursement. Norway: The Norwegian Medicines Control Authority, Department of Pharmacoeconomics, 1999.

4. National Institute for Clinical Excellence. Technical guidance for manufacturers and sponsors [online]. Available from: http://www.nice.org.uk/. Accessed 2001.
5. Wakker P, Klaassen MP. Confidence intervals for cost/effectiveness ratios. Health Econ. 1995;4:373–81.
6. Willan AR, O'Brien BJ. Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. Health Econ. 1996;5:297–305.
7. Laska EM, Meisner M, Siegel C. Statistical inference for cost-effectiveness ratios. Health Econ. 1997;6:229–42.
8. Lacey L, Mauskopf J, Lindrooth R, Pham S, Saag M, Sawyer W. A prospective cost-consequence analysis of adding lamivudine to zidovudine-containing antiretroviral treatment regimens for HIV infection in the US. PharmacoEconomics. 1999;15(Suppl 1):23–37.
9. Pieters WR, Lundbäck B, Johansson G, et al. Cost-effectiveness analyses of salmeterol/fluticasone propionate combination product and fluticasone propionate in patients with asthma. II: study methodologies. PharmacoEconomics. 1999;16(Suppl. 2): 9–14.
10. Tennvall G, Norinder A, Ohlin B. Cost effectiveness of helicobacter pylori eradication therapies in patients with duodenal ulcer: an analysis of triple therapy versus two dual therapy alternatives. PharmacoEconomics. 1999;16:297–306.
11. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making. 1998;18(Suppl): S68–S80.
12. Van Hout BA, Al MJ, Gordon GS, Rutten F. Costs, effects and C/E ratios alongside a clinical trial. Health Econ. 1994; 3:309–19.
13. O'Hagan A, Stevens JW, Montmartin J. Inference for the C/E acceptability curve and C/E ratio. PharmacoEconomics. 2000;17:339–49.
14. Raikou M, Gray A, Briggs A, et al. (on behalf of the UK Prospective Diabetes Study Group). Cost effectiveness analysis of improved blood pressure control with type 2 diabetes: UKPDS 40. Br Med J. 1998;317:720–6.
15. Sculpher M, Poole L, Cleland J, et al. (on behalf of the ATLAS Study Group). Low doses versus high doses of the angiotensin converting enzyme inhibitor lisinopril in chronic heart failure: a cost-effectiveness analysis based on the Assessment of Treatment with Lisinopril and Survival (ATLAS) study. Eur J Heart Failure. 2000;2:447–54.
16. Willan, AR, O'Brien BJ. Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. Health Econ. 1999;8:203–11.
17. Gardiner JC, Huebner M, Jetton J, Bradley CJ. Power and sample size assessments for tests of hypotheses on cost-effectiveness ratios. Health Econ. 2000;9:227–34.
18. Briggs A, Gray AM. Sample size and power calculations for stochastic cost-effectiveness analysis. Med Decis Making. 1998;18(Suppl.):S81–S92.
19. Briggs A, Tambour M. The design and analysis of stochastic cost-effectiveness studies for the evaluation of health care interventions (Working Paper Series in Economics and Finance No. 234). Stockholm, Sweden: Stockholm School of Economics, 1998.
20. Laska EM, Meisner M, Siegel C. Power and sample size in cost-effectiveness analysis. Med Decis Making. 1999;19: 339–43.
21. Chaloner K, Verdinelli I. Bayesian experimental design: a review. Stat Sci. 1995;10:273–304.
22. Lindley DV. The choice of sample size. The Statistician. 1997;46:129–38.
23. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Stat Med. 1986;5:1–13.
24. Hutton JL, Owens RG. Bayesian sample size calculations and prior beliefs about child sexual abuse. The Statistician. 1993;42:399–404.
25. Lee SJ, Zelen M. Clinical trials and sample size considerations: another perspective. Stat Sci. 2000;15: 95–103.
26. Lindley DV, Singpurwalla ND. On the evidence needed to reach agreed action between adversaries, with application to acceptance sampling. J Am Stat Assoc. 1991;86:933–7.
27. Etzioni R, Kadane JB. Optimal experimental design for another's analysis. J Am Stat Assoc. 1993;88:1404–11.
28. Simon R. Comment. Stat Sci. 2000;15:103–5.
29. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. Health Econ. 2000.
30. O'Hagan A. Kendall's Advanced Theory of Statistics. Volume 2B, Bayesian Inference. London: Arnold, 1994. Paragraphs 6.48–6.54.