*Statistical Science*
2000, Vol. 15, No. 2, 95–110

# Clinical Trials and Sample Size Considerations: Another Perspective

**Sandra J. Lee and Marvin Zelen**

*Abstract.* We propose a Bayesian formulation of the sample size problem for planning clinical trials. The frequentist paradigm for calculating sample sizes for clinical trials is to prespecify the type I and II error probabilities. These error probabilities are conditional on the true hypotheses. Instead we propose prespecifying posterior probabilities which are conditional on the outcome of the trial. Our method is easy to implement and has intuitive interpretations. We illustrate an application of our method to the planning of cancer clinical trials for the Eastern Cooperative Oncology Group (ECOG).

*Key words and phrases:* Type I and II error probabilities, posterior error probabilities, clinical trials, Bayesian inferences.

## 1. INTRODUCTION

This paper discusses the use of Bayesian ideas in the formulation and calculation of sample sizes for planning clinical trials. Our view is that there are special features of the clinical trials setting that require a Bayesian formulation. However the implementation of our ideas has a frequentist perspective. The discussion will be formulated in the context of comparing two treatments, but the ideas are easily generalized to compare more than two treatments.

Nearly all sample size calculations for planning clinical trials follow the usual frequentist ideas by choosing a fixed type I error ($\alpha$) and calculating a sample size consistent with a prespecified power $(1-\beta)$ to detect a prespecified noncentrality value $\delta$. Ordinarily the value of $\alpha$ is taken to be $\alpha = 0.05$ or in rarer cases $\alpha = 0.01$. It is highly unusual to have values of $\alpha > 0.05$, although it is common to have $0.05 < \beta \leq 0.2$. There is no logic in the widespread use of $\alpha = 0.05$ except that there is general agreement that it should not be large. Intuitively it is clear that $\beta$ should not be large, but there is no general agreement on the widespread use of a fixed $\beta$ value. A referee has recommended that the selection of $(\alpha, \beta)$ should be based on the relative costs of making a wrong decision. However the consideration of relative costs is rarely done. We have no knowledge of any such application in the clinical trial setting.

There have been recent papers in the literature putting forth Bayesian ideas in the planning of clinical trials. The papers by Berger, Boukai and Wang (1997, 1999), although not specifically directed at the planning of clinical trials, attempt to reconcile frequentist and Bayesian ideas that have implications for the interpretations of results from clinical trials. Spiegelhalter and Freedman (1986) propose using the posterior distribution to plan clinical trials when the magnitude of the noncentrality parameter is obtained from subjective clinical opinions (see Discussion). Recent papers on this topic are Joseph and Belisle (1997), Joseph, Wolfson and du Berger (1995), Lindley (1997) and Pham-Gia (1997). However these efforts have for the most part been ignored by those engaged in the planning of clinical trials. We also note that Peto et al. (1976) briefly discuss the advantages and disadvantages of interpreting a $p$-value together with prior opinions.

Nevertheless, most statistical practitioners of clinical trials acknowledge that the formal procedures for planning sample sizes in clinical trials have many subjective elements. Among these are the choice of the $\alpha$ level and the sensitivity of the trial to have acceptable power for detecting (say) a specified noncentrality parameter. Our approach to the sample size problem leads to specification of

*Sandra J. Lee is Research Scientist, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts 02115* (*e-mail: sjlee@jimmy.harvard.edu*) *and Marvin Zelen is Professor, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts 02115.*

$(\alpha, \beta)$, but requires other subjective elements that may be easier to estimate or accept.

The basic idea underlying the frequentist theory of planning studies is to control the false positive and false negative error rates by specifying them in advance of the study. These are probabilities that are conditional on the true state of the hypotheses under consideration. We believe this methodology is inappropriate for the planning of clinical trials. Our views are motivated by the following considerations. If the outcome of the clinical trial is positive (one treatment is declared superior), then the superior therapy is likely to be adopted by other physicians. Similarly if the outcome is neutral (treatments are comparable), then considerations other than outcomes are likely to be important in the adoption of therapy.

The consequences of false negative and positive decisions after the clinical trial has been completed are conditional on the outcome of the clinical trial. Hence the false positive and negative error rates should also be conditional on the trial outcome. In our view these latter probabilities should be used in the planning of clinical trials. The contrast is that there are two classes of error rates: the frequentist type I and II error rates (that are conditional on the true hypotheses) and the posterior error rates (that are conditional on the outcome of the trial). It is the latter rates that are important in assessing the consequences of wrong decisions. In "plain" language, we believe the two fundamental issues are (a) if the trial is positive, "What is the probability that the therapy is truly beneficial?" and (b) if the trial is negative, "What is the probability that the therapies are comparable?" The frequentist view ignores these fundamental considerations and can result in positive harm because of the use of inappropriate error rates. The positive harm arises because an excessive number of false positive therapies may be introduced into practice. Many positive trials may be unethical to duplicate and, even if replicated, could require many years to complete. Hence a false positive trial outcome may generate many years of patients' receiving nonbeneficial therapy.

The above discussion essentially casts the fundamental interpretation of a clinical trial in a Bayesian setting. This requires that the determination of the required sample size also be viewed from a Bayesian framework. It is the purpose of this paper to reformulate the sample size problem in the context of planning clinical trials and to illustrate how it can be used in a practical way. Because of the widespread use of the 0.05 significance level we believe that an unacceptably high proportion of possibly ineffective therapies may have been routinely adopted. We do not intend to be pejorative with respect to the use of well conducted clinical trials, but feel that the interpretation of results from such trials is not well understood.

The outline of this paper is that Section 2 formulates the problem, Section 3 contains illustrations of the planning of trials and Section 4 concludes with a general discussion.

## 2. NOTATION AND PROBLEM FORMULATION

Consider a phase III clinical trial for comparing two groups. To motivate ideas, suppose the trial is comparing an experimental therapy with the best available therapy. It is assumed that the experimental therapy has been evaluated in phase II clinical trials that have demonstrated some efficacy. The available evidence allows the possibility that the experimental therapy may be comparable to, or potentially of even greater benefit than, the best available therapy.

Let $\delta$ be a noncentrality parameter that is a function of the parameters so that the hypothesis testing situation can be formulated in the traditional frequentist paradigm; that is, $H_0$: $\delta = 0$ versus $H_1$: $\delta \neq 0$ (two-sided alternative). (Later in this section, we discuss carrying out a trial with a one-sided alternative hypothesis together with the associated ethical issues.) A test statistic $T(Y)$, that is a function of the observations, will be used to carry out an appropriate statistical test. Without loss of generality we will take the mean and variance of the test statistic to be $E(T(Y)) = \delta$ and $\text{var}(T(Y)) = 2\sigma^2/n$ where $\sigma^2$ is defined in an appropriate manner and $n$ is the sample size for each group. For example if the test statistic is the difference of two sample averages, $T(Y) = \bar{Y}_1 - \bar{Y}_2$, then $E(\bar{Y}_1 - \bar{Y}_2) = m_1 - m_2 = \delta$ and $\text{var}(\bar{Y}_1 - \bar{Y}_2) = \sigma_1^2/n + \sigma_2^2/n = (2/n)(\sigma_1^2 + \sigma_2^2)/2 \equiv 2\sigma^2/n$. If two proportions are being compared, averages are replaced by sample proportions. It will be assumed that the sample sizes are sufficiently large so that the test statistic has an asymptotic normal distribution. However, this assumption is not necessary for our main development.

We shall consider a two-sided alternative hypotheses, that is, $H_0$: $\delta = 0$ versus $H_1$: $\delta \neq 0$. In order for this trial to be ethical, it is necessary that, a priori, there is no reason to favor one therapy over the other. We assign a prior probability of $\theta$ to the joint event $\delta > 0$ or $\delta < 0$. Hence there will be a prior probability $(1 - \theta)$ of the null hypothesis being true, and $\theta/2$ will be the prior probability for each alternative $\delta > 0$ and $\delta < 0$. If the prior probabilities are not equal for $\delta > 0$ versus $\delta < 0$,

then the trial would be unethical as it will be advantageous for the patient to be assigned to the treatment with the larger prior probability of being superior. Our view is that if a physician has a prior belief that one of the treatments is "likely" to be better for a particular patient, then the physician cannot ethically enter that patient in a clinical trial.

The quantity $\theta$ essentially summarizes the prior evidence and/or subjective assessment in favor of differences between treatments. It also reflects the level of clinical innovation that motivated the trial. Numerical values of $\theta$ are difficult to estimate. In practice it may only be necessary to have a range of values.

In some applications it may be possible to order different experimental situations with regard to $\theta$. For example, combining two drugs into a combination, in which each drug alone has been shown to be clinically ineffective, may have a lower prior probability of success when compared to a situation in which one is evaluating a two-drug combination in which each drug has shown benefit. Another possibility is that a series of pilot or preclinical studies may be available showing that a new treatment is beneficial. If the disease that is being treated has no accepted standard treatment, or the standard treatment is regarded as having very modest benefit, then it may be unethical to carry out a clinical trial comparing the new treatment to a control or the standard treatment, as the prior probabilities may not be equal for $\delta > 0$ compared to $\delta < 0$.

The null hypothesis $H_0$: $\delta = 0$ is a "shorthand" expression that $\delta$ is in the neighborhood of $\delta = 0$. There is an "indifference" region in which the treatments are regarded as comparable. The hypothesis testing situation can be reformulated by denoting an indifference region $|\delta| < \delta_0$ and a region of importance $|\delta| > \delta_1$. Then the null and alternative hypotheses can be stated as $H_0$: $|\delta| \le \delta_0$ and $H_1$: $|\delta| > \delta_1$, with the region $\delta_0 < |\delta| < \delta_1$ being regarded as an "indecisive" region. A special case in the formulation is to take $\delta_0 = \delta_1$. The usual formulation is to set $\delta_0 = \delta_1 = 0$. We note that the main ideas of this paper can be applied to this more general formulation of the hypothesis testing problem. However, we do not address these modifications further as it detracts from the main theme of this paper.

As a special case of the hypothesis testing context described above, we distinguish between the two sets of one-sided hypothesis testing situations: (a) $H_0$: $\delta \le 0$ versus $H_1$: $\delta > 0$ and (b) $H_0$: $\delta = 0$ versus $H_1$: $\delta > 0$. The hypotheses denoted by (a) require that equal prior probabilities ($\theta = 1/2$) be assigned to the null and alternative hypotheses. Otherwise the trial would be unethical. In the setting of (b), we assume that the prior probability of $\delta < 0$ is zero. An example of this situation is when a new treatment is being evaluated against no treatment or placebo and the new treatment will not result in negative benefit. Our thinking is that the hypotheses denoted by (b) should never be tested in a clinical trial because in that context it is always advantageous to the patient to be assigned to the new treatment. These considerations show that one-sided alternative hypotheses require different considerations with respect to the ethical basis of a clinical trial as well as the choice of the prior distribution.

The outcome of the clinical trial will be idealized as having a positive or negative outcome. The positive outcome refers to the conclusion that $\delta \neq 0$, whereas a negative outcome refers to the conclusion $\delta = 0$. Define $C$ to be a binary random variable which reflects the outcome of the clinical trial; that is, $C = +$ or $-$. Also define $T$ to be an indicator random variable which denotes the true state of the hypothesis under evaluation; that is, $T = -$ refers to $\delta = 0$ and $T = +$ signifies $\delta \neq 0$.

Our assessment of $T$ will be identified with the prior probability associated with $\delta \neq 0$; that is, $\theta = \Pr(T = +)$. Also define the usual frequentist probabilities of making false positive and false negative conclusions from the data by

$$\alpha = \Pr(C = +|T = -),$$
$$\beta = \Pr(C = -|T = +).$$

In an analogous way define

$$\alpha^* = \Pr(T = +|C = -),$$
$$\beta^* = \Pr(T = -|C = +).$$

These are the posterior probabilities of the true situation being opposite to the outcome of the trial. They are the posterior false positive and false negative error probabilities.

In the context of applications it is sometimes convenient to refer to the complement of the posterior error probabilities. For this purpose define $P_1$ and $P_2$ to be $P_1 = 1 - \alpha^* = \Pr(T = -|C = -)$ and $P_2 = 1 - \beta^* = \Pr(T = +|C = +)$. These quantities are functions of $(\alpha, \beta, \theta)$. A direct application of Bayes Theorem results in expressing $(P_1, P_2)$ in terms of $(\alpha, \beta, \theta)$; that is,

$$\begin{aligned} P_1 &= 1 - \alpha^* = \Pr(T = -|C = -) \\ &= (1 - \alpha)(1 - \theta)/\big[(1 - \alpha)(1 - \theta) + \beta\theta\big], \end{aligned} \tag{1}$$

$$P_2 = 1 - \beta^* = \Pr(T = +|C = +) = (1-\beta)\theta / [(1-\beta)\theta + \alpha(1-\theta)]. \quad (2)$$

Note that if $P_1 = P_1(\theta, \alpha, \beta)$ and $P_2 = P_2(\theta, \alpha, \beta)$, then $P_2 = P_1(1-\theta, \beta, \alpha)$ and $P_1 = P_2(1-\theta, \beta, \alpha)$. The quantities $P_1$ and $P_2$ also arise in the evaluation of diagnostic tests. They are referred to as the negative and positive predictive values. We prefer to refer to them as posterior probabilities to avoid any confusion about diagnostic tests versus hypothesis testing applications.

Alternatively $(\alpha, \beta)$ can be written as a function of $(\theta, P_1, P_2)$; that is,

$$\alpha = (1-P_2)(\theta + P_1 - 1)/(1-\theta)(P_1 + P_2 - 1), \quad (3)$$

$$\beta = (1-P_1)(P_2 - \theta)/\theta(P_1 + P_2 - 1). \quad (4)$$

Equations (3) and (4) require $P_1 > 1-\theta$ and $P_2 > \theta$ or $P_1 < 1-\theta$ and $P_2 < \theta$. Otherwise $\alpha$ and $\beta$ could be negative. Note that $\alpha = \beta = 1-P$ if $P_1 = P_2 = P$ and $\theta = 0.5$. The relations $P_1 > 1-\theta$ and $P_2 > \theta$ ensure that the posterior probabilities are larger than the prior probabilities and this hypothesis test is informative. The opposite is true for $P_1 < 1-\theta$ and $P_2 < \theta$. This latter condition seems unreasonable and we will only require that $P_1 > 1-\theta$ and $P_2 > \theta$.

The frequentist method of calculating sample size in a clinical trial is to specify $(\alpha, \beta, \delta)$. This specification is sufficient to allow the calculation of the sample size. We believe a more relevant way of calculating sample size is to specify $(\theta, P_1, P_2)$, or equivalently $(\theta, \alpha^*, \beta^*)$, yielding the values of $(\alpha, \beta)$ by use of equations (3) and (4).

Given the value of $(\alpha, \beta)$, we then have the large sample relationship for two-sided tests,

$$n(\delta/\sigma)^2 = 2(z_{\alpha/2} + z_\beta)^2, \quad (5)$$

where

$$Q(z_\gamma) = \int_{z_\gamma}^{\infty} (2\pi)^{-1/2} \exp(-t^2/2)\, dt = \gamma.$$

The relationship given by (5) is a suitable approximation for the two-sided alternative if the type I error is in the neighborhood of $\alpha \le 0.05$. Otherwise it may be necessary to use the more accurate normal approximation to the power given by

$$1 - \beta = Q\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma\sqrt{2}}\right) + Q\left(z_{\alpha/2} + \frac{\delta\sqrt{n}}{\sigma\sqrt{2}}\right). \quad (6)$$

Conditional on the value of $(\alpha, \beta)$, there is a tradeoff between sample size $(n)$ and $\delta/\sigma$ as given by (5) or its more complicated version (6). Thus for fixed $(\alpha, \beta)$ and given $\theta$, one can tabulate values of $n$ and $\delta/\sigma$ which satisfy the error probabilities. In the above it is assumed that $\sigma$ is approximately known or that $\delta$ is expressed as a multiple of $\sigma$.

TABLE 1
*Type I and II errors for specified $P_1$, $P_2$ and $\theta$*

| θ | $P_1$ | $P_2$ | α | β |
|---|---|---|---|---|
| 0.25 | 0.85 | 0.95 | 0.0083 | 0.5250 |
| 0.25 | 0.90 | 0.95 | 0.0118 | 0.3294 |
| 0.25 | 0.95 | 0.95 | 0.0148 | 0.1556 |
| 0.50 | 0.85 | 0.95 | 0.0438 | 0.1688 |
| 0.50 | 0.90 | 0.95 | 0.0471 | 0.1059 |
| 0.50 | 0.95 | 0.95 | 0.0500 | 0.0500 |
| 0.75 | 0.85 | 0.95 | 0.1500 | 0.0500 |
| 0.75 | 0.90 | 0.95 | 0.1529 | 0.0314 |
| 0.75 | 0.95 | 0.95 | 0.1556 | 0.0148 |

We have derived all of the elements for estimating sample size. Instead of choosing $(\alpha, \beta)$, the investigator chooses $(P_1, P_2, \theta)$ subject to $P_1 > 1-\theta$ and $P_2 > \theta$. This in turn specifies $(\alpha, \beta)$ using equations (3) and (4), and a table of $n$ versus $\delta/\sigma$ can be calculated by making use of (5) or (6). If there is some uncertainty about $\theta$, the entire calculation may be repeated for different values of $\theta$. The final sample size is chosen corresponding to a fixed value of $\delta/\sigma$.

The procedure described above requires selecting values of $(P_1, P_2)$ prior to the trial. The trial will be planned so that $(P_1, P_2)$ are the posterior probabilities after the trial is completed. It is clear that these posterior probabilities should be reasonably high. Intuitively, our view is that a positive outcome from a clinical trial should have a relatively high posterior probability of being true. We recommend that $P_2 = 0.95$ or higher. We believe that values of $P_1$ can be lower, and could range from 0.85 to values close to unity.

Table 1 summarizes values of $(\alpha, \beta)$ for $P_1 = 0.85$, 0.90, 0.95 and $P_2 = 0.95$ for values of $\theta = 0.25$, 0.50, 0.75 and Figure 1 shows how these quantities change over a range of prior probabilities. In general, as $\theta \to 1$, $\alpha$ increases and $\beta \to 0$; similarly as $\theta \to 0$, $\beta$ increases and $\alpha \to 0$. When $P_1 = 0.95$ and $P_2 = 0.95$, both $\alpha$ and $\beta$ remain under 0.156 for $\theta$ between 0.25 and 0.75.

We note that if the alternative hypothesis is specified by $H_1$: $|\delta| > \delta_1$ and a prior distribution $P(\delta)$ is available for $\delta$, then an "average" sample size may be calculated by the expression $E_\delta(n) = \int_{|\delta|>\delta_1} n(\delta)P(\delta)\, d\delta$.

## 3. PLANNING OF CLINICAL TRIALS

### 3.1 ECOG Studies

In this section we illustrate one application of our methods to the planning of cancer clinical trials for the Eastern Cooperative Oncology Group (ECOG).

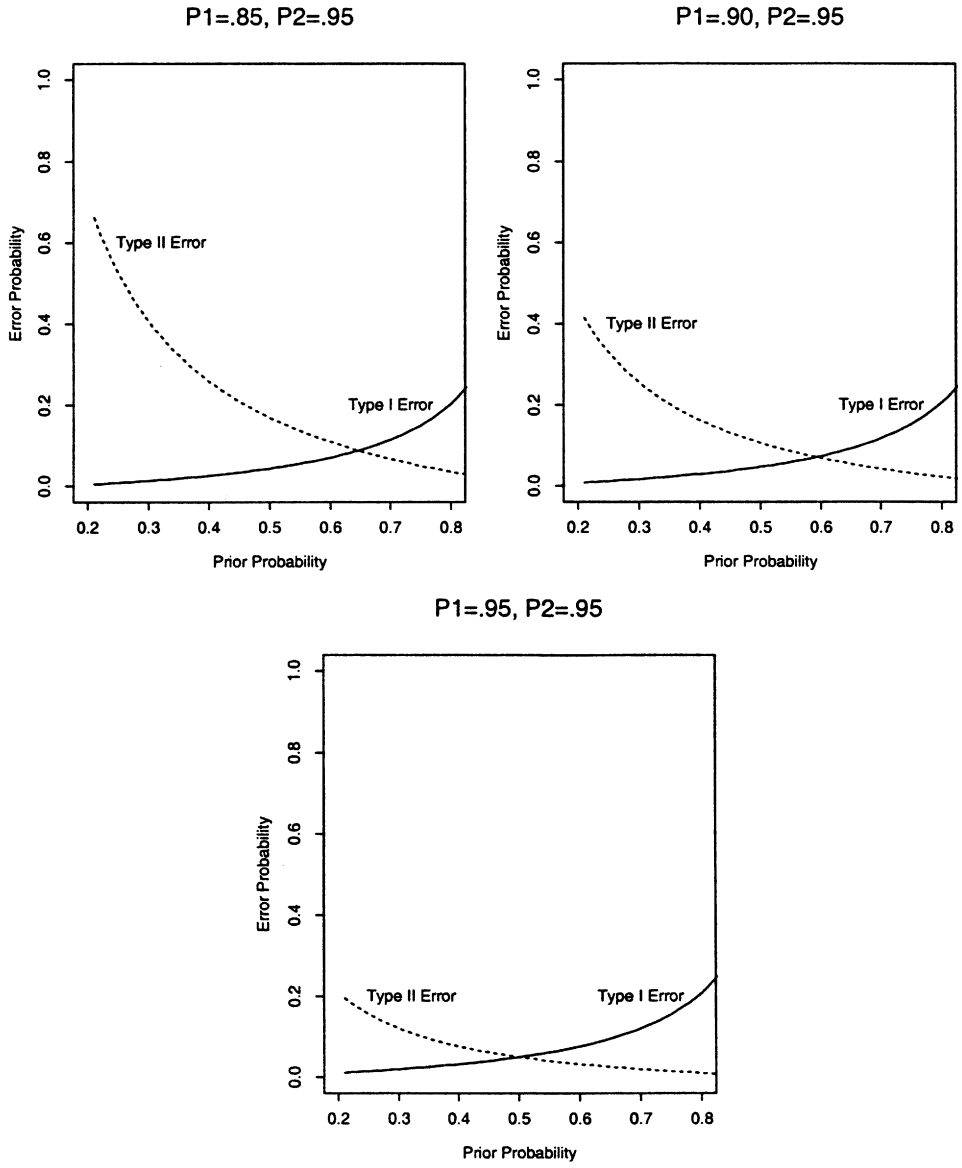


FIG. 1. *Type I and type II error probabilities versus prior probabilities* ($\theta$) *for selected values of* $P_1$ *and* $P_2$.

The ECOG is a collection of major cancer treatment centers, medical school hospitals and community hospitals (mainly located throughout the United States) which enter cancer patients into common therapeutic clinical trials. It has been in existence since 1955.

In order to estimate the quantity $\theta = \Pr(T = +)$, we evaluated outcomes of all phase III clinical trials conducted by ECOG during a recent 15-year period. Between 1980 and 1995, 98 studies were activated and completed. Among these studies, 87 trials had a report on the final outcomes. If any of the major endpoints such as response rate, overall survival or disease-free survival was declared significant, we considered that the study had a positive outcome. Most studies used $\alpha = 0.05$ and $0.10 \leq \beta \leq 0.20$. Among the 87 studies, 25 had significant outcomes. Hence an estimate of the probability of having a positive clinical outcome was $\Pr(C = +) = 25/87 = 0.29$. This quantity ranged from 0.25 to 0.43 in major disease sites such as Breast (0.38), GI (0.33), GU (0.31), Leukemia (0.40), Lymphoma (0.43) and Melanoma (0.25). In the discussion of the ECOG studies, we will use the overall estimate of $\Pr(C = +) = 0.29$. However in planning studies for a particular disease site, the marginal probability for that site should be utilized.

By using the relationship

$$\Pr(C = +) = \Pr(C = +|T = -)\Pr(T = -) + \Pr(C = +|T = +)\Pr(T = +),$$

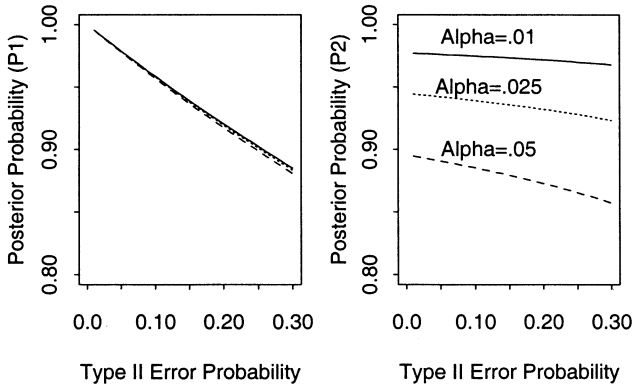


FIG. 2. *ECOG posterior probabilities and type II error probability for* $\alpha = 0.01, 0.025, 0.05$ (*prior probability* $\theta = 0.30$).

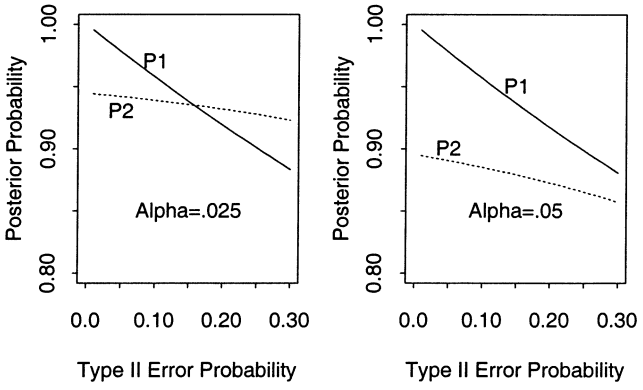


FIG. 4. *ECOG posterior probabilities* ($\theta = 0.30$) *for* $\alpha = 0.025$, 0.05.

we have

$$\theta = \frac{\Pr(C = +) - \alpha}{(1 - \alpha - \beta)}, \tag{7}$$

thus permitting an estimate of $\theta$ from knowledge of $\Pr(C = +)$. Under the assumption of $\alpha = 0.05$ and $0.10 \leq \beta \leq 0.20$, $\theta$ ranges from 0.28 to 0.32. This leads to the posterior probabilities $(P_1, P_2)$ ranging from (0.90, 0.88) to (0.96, 0.88). In other words, the negative clinical trials have a posterior probability of being true negatives within the range [0.90, 0.96]; the positive outcomes have a probability of being true positives equal to 0.88. As a result, among the 25 positive outcomes, 12% (or three trials) are expected to be false positive trials. Similarly among the 62 negative (or neutral) outcomes, 4–10% (two to six trials) are expected to be false negative trials.

Figure 2 displays the relationships between the posterior probabilities and $\beta$ at selected levels of $\alpha$. Figure 3 is a similar plot displaying the relationships between the posterior probabilities and $\alpha$ at selected levels of $\beta$. These calculations were made using a value of $\theta = 0.30$. Note that the posterior negative probability $P_1$ is insensitive to the type I error probability $\alpha$, whereas the posterior positive probability $P_2$ is insensitive to the type II error probability $\beta$. In order to have the posterior positive probability $P_2 > 0.90$, the type I error should be in the neighborhood of [0.025, 0.030].

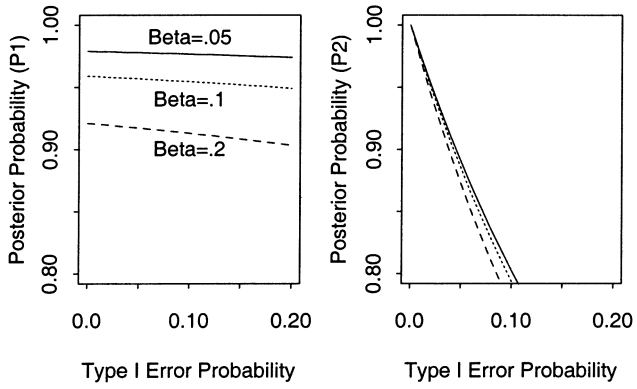


FIG. 3. *ECOG posterior probabilities and type I error probability for* $\beta = 0.05, 0.1, 0.2$ (*prior probability* $\theta = 0.30$).

Figure 4 shows how the posterior probabilities change over a range of $\beta$ values when $\alpha = 0.025$ and $\alpha = 0.05$. Note that values of $\beta$ in the neighborhood of 0.20 result in relatively high posterior probabilities. Because the ECOG experience is comparable to that of other cancer cooperative clinical control groups, we recommend that future trials be planned to have a type I error of less than 0.03 and a power of at least 80%.

### 3.2 Sample Size Calculations

In this section, we illustrate the sample size calculations for two common experimental cases. One case is the comparison of means and the other is the comparison of survival distributions. We illustrate calculating sample size by specifying $(\theta, P_1, P_2)$. First, equations (3) and (4) are used to calculate $(\alpha, \beta)$. Using the large sample approximations described by (5), a table of sample size $(n)$ versus $\delta/\sigma$ can be generated for specified $(\theta, P_1, P_2)$. We consider $P_1 = P_2 = 0.95$, $\theta = 0.25$, 0.5, 0.75 and assume equal variances for the two treatment groups. Table 2 summarizes calculated sample sizes for a range of $\delta/\sigma$. As indicated in Table 1, $P_1 = P_2 = 0.95$ results in different values of $(\alpha, \beta)$ which depend on $\theta$. When $\theta = 0.5$, $\alpha = \beta = 0.05$. The quantity $\theta = 0.25$ leads to values of $(\alpha, \beta) = (0.015, 0.156)$; these values are reversed for $\theta = 0.75$. For example, a value of

TABLE 2
*Sample size versus* $(\delta/\sigma)$ *for* $P_1 = P_2 = 0.95$

| θ | α | β | (δ/σ) | Sample Size (*n*) |
|---|---|---|---|---|
| 0.5 | 0.05 | 0.05 | 0.1 | 2599 |
| | | | 0.2 | 650 |
| | | | 0.3 | 289 |
| | | | 0.4 | 162 |
| | | | 0.5 | 104 |
| 0.25 | 0.015 | 0.156 | 0.1 | 2371 |
| | | | 0.2 | 593 |
| | | | 0.3 | 263 |
| | | | 0.4 | 148 |
| | | | 0.5 | 95 |
| 0.75 | 0.156 | 0.015 | 0.1 | 2576 |
| | | | 0.2 | 644 |
| | | | 0.3 | 286 |
| | | | 0.4 | 161 |
| | | | 0.5 | 103 |

TABLE 3
*Number of events versus* $(\Delta)$ *for* $P_1 = P_2 = 0.95$

| θ | α | β | (Δ) | Number of events ($d_1$) |
|---|---|---|---|---|
| 0.5 | 0.05 | 0.05 | 0.4 | 31 |
| | | | 0.5 | 54 |
| | | | 0.6 | 100 |
| | | | 0.7 | 205 |
| | | | 0.8 | 522 |
| 0.25 | 0.015 | 0.156 | 0.4 | 28 |
| | | | 0.5 | 50 |
| | | | 0.6 | 91 |
| | | | 0.7 | 187 |
| | | | 0.8 | 476 |
| 0.75 | 0.156 | 0.015 | 0.4 | 31 |
| | | | 0.5 | 54 |
| | | | 0.6 | 99 |
| | | | 0.7 | 203 |
| | | | 0.8 | 517 |

$\theta = 0.25$ generates $\alpha = 0.015$, $\beta = 0.156$ and leads to $n(\delta/\sigma)^2 = 23.7$. Any choice of $(n, \delta)$ satisfying $n(\delta/\sigma)^2 = 23.7$ can be used to plan the clinical trial. The entries in Table 2 indicate the range of values of $n$ and $\delta/\sigma$.

These procedures can be extended to calculate the approximate sample size calculation for time-to-event outcomes. We assume that the time-to-event outcome follows the exponential distribution with failure rate $\lambda$, and interest is in comparing $\lambda_1$ and $\lambda_2$ between two treatment groups. In this setting, the total sample size is equivalent to the total number of events from both treatment groups. We denote the number of events by $d_1$ and $d_2$ in each treatment group. Again using the large sample approximation, we have

$$\frac{d_1 d_2}{d_1 + d_2}(\ln \Delta)^2 = (z_{\alpha/2} + z_\beta)^2 \tag{8}$$

with $\Delta = \lambda_1/\lambda_2$. In (8), the type I error is $\alpha$ for two-sided tests. For the purpose of illustration, we assume that $d_1 = d_2$. Calculated total number of events $(d_1)$ under a range of $\Delta$ are presented in Table 3 for $P_1 = P_2 = 0.95$.

An alternative treatment of sample size calculations for exponentially distributed survival time is to carry out a test using information on the number of failures. If the two groups have equal person years of follow-up, then the total number of observed events to achieve a level $\alpha$ test (two-sided) with a power of $(1 - \beta)$ is given by

$$d = \frac{\left\{\frac{z_{\alpha/2}}{2} + z_\beta \sqrt{\Delta/(1+\Delta)^2}\right\}^2}{\left(\frac{\Delta}{1+\Delta} - 0.5\right)^2}. \tag{9}$$

The above formula was derived by taking the failures to follow a Poisson distribution with the same person years of follow-up time for each group. Equation (9) immediately follows by conditioning on the sum of two Poisson random variables $(d_1 + d_2 = d)$ which has a binomial distribution with sample size $d$ and success probability $\Delta/(1 + \Delta)$. The value of $d$ in (9) is approximately twice of the value of $d_1$ in Table 3.

## 4. DISCUSSION

We have proposed that the calculation of sample sizes for clinical trials be formulated by prespecifying the posterior error rates. In our view, the prespecification of the posterior probability error rates is more appropriate in the clinical trials setting than the specifications of the usual type I and II error probabilities. The devotion to a type I error of $\alpha = 0.05$ has no empirical or theoretical basis. However, specifying the posterior error rates generates the appropriate type I and II error probabilities. Armed with the resulting $(\alpha, \beta)$, it is then possible to determine the trade-off between the sample size and noncentrality parameters for large samples. The analysis of the data may proceed in the usual frequentist way in which the type I error rate is used to "judge" statistical significance.

Spiegelhalter and Freedman (1986) have also formulated the clinical trial sample size problem as a Bayesian formulation. They recommend finding a prior distribution for $\delta$ by interviewing physicians. The prior distribution is then used to calculate an average power curve based on a two-sided confidence interval. This leads to an appropriate sample size. Their formulation also allows the possibility

of reaching no conclusion. One difference between their formulation and ours is that they keep the confidence coefficient fixed (their example uses a 95% confidence coefficient) and do not allow the posterior or prior probabilities to affect the choice of the confidence coefficient.

Although we have only discussed fixed sample size calculations, our ideas are easily adapted without any change to trials which allow early stopping. One simply uses the calculated $(\alpha, \beta)$ generated by the prespecified posterior error probabilities to determine the early stopping rules.

Our development makes use of both Bayesian and frequentist ideas. It may be viewed as a compromise between these two "Schools of Inference." However we believe that the reformulation of the sample size calculation problem in the clinical trials setting requires specifying the posterior error probabilities at the planning stage of the trial.

The current frequentist practice of choosing $\alpha = 0.05$ in the context of calculating sample sizes is clearly subjective. One may argue that choosing the posterior error probabilities in advance is also subjective. However there is likely to be more agreement about the desired range of the posterior error probabilities. Because a positive finding from a clinical trial is likely to influence clinical practice, nearly all clinical investigators would favor a high posterior error probability associated with a true positive outcome. We favor values of $P_2$ to be in the neighborhood of 0.95 or even higher. Alternatively there may be more flexibility in choosing a posterior negative probability. We favor that the value of $P_1$ should be at least 0.90. If the clinical trial concludes that the treatments are comparable, when in reality one of the treatments is more beneficial, less harm is likely to be done as both treatments are likely to be used in practice.

An issue arises in the reporting of a clinical trial. Suppose the data generated a significance level ($p$-value) of $\alpha_0$. If $\alpha_0 \leq \alpha$, then the results would be reported as being "statistically significant" and the posterior probability $P_1$ serves as a lower bound to the actual posterior probability. However, suppose the value of $\alpha_0$ is within the interval $\alpha < \alpha_0 \leq 0.05$. How should such results be reported?

We suggest that the posterior probability $P_2$ be recalculated. This can be done by adopting $\alpha_0$ as the significance level and recalculating a new type II error, denoted by $\beta_0$. The recalculated type II error is carried out by using the actual sample size of the trial, the significance level $\alpha_0$ and the same noncentrality parameter $\delta$ as in the initial calculation; consequently, $z_{\beta_0} = z_{\alpha/2} + z_\beta - z_{\alpha_0/2}$. Using (2), the new posterior probability, denoted by $P_2^*$ is readily calculated. Thus if the investigators declare "statistical significance," the recalculated posterior probability reflects the probability of the certainty of the conclusions. For example, suppose the initial sample size was calculated on the basis of $P_1 = 0.92$ and $P_2 = 0.93$ with $\alpha = 0.025$, $\beta = 0.20$ and $\theta = 0.30$. Suppose the actual trial generated a significance level of $\alpha_0 = 0.05$. Hence recalculating the type II error results in $\beta_0 = 0.16$ and the recalculation of the posterior probability yields $P_2^* = 0.88$. Hence the interpretation would be that if the clinical trial outcome is regarded as "statistically significant," there is a posterior probability of $P_2^* = 0.88$ of the conclusions being correct.

The example from the ECOG illustrates that objective prior probabilities can be found for some classes of clinical trials. In the U.S., many cancer clinical trials are carried out with cooperative clinical trial groups of which ECOG is representative. In general, for any disease there will be trial data available from past studies that can be used to calculate prior probabilities.

The cancer cooperative clinical trials groups are a special situation, as data on past trials is readily accessible. Such data may not exist for other diseases. In such cases, the probability $\Pr(C = +)$ may be estimated by using MEDLINE to search for the recent history of clinical trials for the specific disease. If both negative and positive trials are published, then $\Pr(C = +)$ may be appropriately estimated. Otherwise subjective modifications have to be made, if there is an underreporting of negative trials.

In practice, there may be a modification of one-sided hypothesis as described by Lan and Friedman (1986). This arises when the trial is evaluating a new treatment compared to a standard treatment and the experimental plan allows for early termination. Operationally the trial will terminate early if (a) the new treatment generates data showing superiority or (b) the new and standard therapies are comparable and it is unlikely that the new treatment will be shown superior to the standard treatment. This latter action is taken by calculating the conditional power of concluding that the new treatment is superior if the trial was to continue. Hence the trial may have been planned with a two-sided alternative, motivated by ethical considerations. However, the trial may be aborted if the trial is unlikely to show the superiority of the new treatment. Hence the statistical hypotheses in practice are $H_0$: $\delta \leq 0$ versus $H_1$: $\delta > 0$ where $\delta > 0$ indicates superiority of the new treatments.

In summary, we believe that planning clinical trials by prespecifying the posterior probabilities is the appropriate way to plan all phase III clinical trials.

The traditional frequentist approach does not seem relevant in the planning stages. Even though the planning of the clinical trial uses a Bayesian formulation, we do not necessarily advocate the use of Bayesian methods of statistical analysis. Our development leads to assigned type I and type II errors as well as the required sample size. The assigned type I error should be used to assess statistical significance. Thus we have a sample size formulation which is Bayesian, but the analysis may proceed in the usual frequentist mode.

### ACKNOWLEDGMENTS

We thank the Editors and reviewers for many helpful comments. Work supported in part by research grants from the National Cancer Institute, National Institutes of Health.

### REFERENCES

BERGER, J. O., BOUKAI, B. and WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statist. Sci.* **12** 133–160.

BERGER, J. O., BOUKAI, B. and WANG, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika* **86** 79–92.

JOSEPH, L. and BELISLE, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *Statistician* **44** 209–226.

JOSEPH, L., WOLFSON, D. B. and DU BERGER, R. (1995). Sample size calculations for binomial proportions via highest posterior density interval. *Statistician* **44** 167–171.

LAN, K. K. G. and FRIEDMAN, L. (1986). Monitoring boundaries for adverse effects in long-term clinical trials. *Controlled Clinical Trials* **7** 1–7.

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

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., MCPHERSON, K., PETO, J. and SMITH, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *British Journal of Cancer* **34** 585–612.

PHAM-GIA, T. (1997). On Bayesian analysis, Bayesian decision theory and the sample size problem. *Statistician* **46** 139–144.

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* **5** 1–13.

# Comment

**Richard Simon**

The paper by Lee and Zelen (L&Z) provides a nice framework for thinking about important aspects of the planning of clinical trials and the interpretation of results of such trials. The interaction of frequentist and Bayesian concepts in the paper also provides an opportunity to highlight the contrasts and similarities of these approaches.

Determination of sample size is an important aspect of planning a clinical trial. 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. As a result, in some fields there is a glut of small clinical trials with inadequate power for detecting treatment effects that might realistically be expected to exist. In such a setting, many of the "positive" trials reporting statistically significant differences are likely to be false positives. This phenomenon was also previously noted by Staquet, Rozencweig, Von Hoff and Muggia (1979) and Simon (1982), using the sensitivity–specificity derivation employed by Lee and Zelen in their current paper. I have previously referred to this phenomenon as the "thermodynamics of clinical trials" (Simon, 1982).

In the current paper, L&Z propose to use this same approach to establish the sample size of a clinical trial. The accept–reject formulation employed by L&Z is not adequate, however, for ensuring that frequentist interpretations of clinical trial results are associated with strong Bayesian support for the acceptance or rejection of hypotheses about treatment differences. I would like to present a simple alternative development of the ideas raised by L&Z which I believe is more appropriate for aligning

*Richard Simon, D.Sc., is Chief, Biometric Research Branch, National Cancer Institute, Building EPN, Room 739, Bethesda, Maryland 20892* (*e-mail*: *rsimon@nih.gov*).

Bayesian and frequentist analyses and for planning sample size.

Let $\delta$ denote the true treatment effect, as in L&Z. Let $\hat{\delta}$ denote the maximum likelihood estimate of $\delta$. We will assume that $\hat{\delta}$ is sufficient for $\delta$ and that $\hat{\delta} \mid \delta \sim N(\delta, s^2)$, where the experimental variance $s^2$ depends on the sample size but is otherwise known. In practice, $s^2$ will be estimated, but we will ignore this additional variability. In order to present the concepts involved in a clear manner that avoids technical complexities, we will assume that $\delta$ has a two-point prior distribution which assigns probability $1 - \theta$ to the null hypothesis that $\delta = 0$ and probability $\theta$ to an alternative hypothesis that $\delta = \delta_1$. More complex prior densities are easily accommodated but the development is more complex and will not be reported here. L&Z argue that the prior distribution must be symmetric about zero for the clinical trial to be "ethical." Not all biomedical ethicists agree with this position. The statistics of the outcomes of large numbers of randomized clinical trials are not likely to be symmetric about zero and in that sense a symmetric prior is inappropriate. Clinical trials have multiple endpoints. Most major clinical trials compare a new treatment to a standard treatment. Frequently for trials of cancer treatments, the new treatment is more toxic and will not be adopted unless it is better by a non-negligible amount compared to the control regimen. Often the new treatment is expensive and will only be adopted if it is superior to the standard. Often the new regimen is not approved for marketing and is only available in a clinical trial. There is also the issue of whose prior should be used for planning the trial. Different audiences have different a priori degrees of skepticism or enthusiasm for the effectiveness of the new treatment relative to the control (Spiegelhalter, 1994). Hence, it seems inadequate to assume that the prior for the primary endpoint of an "ethical" clinical trial must be symmetric about zero.

L&Z claim to present a Bayesian analysis, but they do not specify a prior on specific values other than $\delta = 0$. They also do not utilize a proper likelihood. A likelihood specifies the probability density of the data for a specified value of the parameters. The sensitivity–specificity derivation used by L&Z specifies the probability of an infinite interval $\{\hat{\delta}/s > k_{1-\alpha/2}\}$, but this is not a likelihood function.

The posterior probability of the null hypothesis having observed $\hat{\delta}$ is easily shown to be

$$\left\{1 + \left(\frac{\theta}{1-\theta}\right)\frac{\phi((\hat{\delta} - \delta_1)/s)}{\phi(\hat{\delta}/s)}\right\}^{-1}, \tag{1}$$

where $\phi$ denotes the standard normal density function. The ratio of normal densities is the Bayes factor

$$\mathrm{BF} = \frac{\phi((\hat{\delta} - \delta_1)/s)}{\phi(\hat{\delta}/s)}.$$

In order to have the posterior probability of the null hypothesis given the data be less than 0.1, expression (1) implies that

$$\frac{\theta}{1-\theta}\mathrm{BF} \geq 9. \tag{2}$$

Most major clinical trials are planned to have 80% or 90% power for rejecting the null hypothesis at a two-sided 5% significance level when the alternative hypothesis is true. Using the usual sample size formula, this implies

$$\delta_1/s = k_{1-\alpha/2} + k_{1-\beta}.$$

Hence for most major clinical trials, $\delta_1/s \approx 3$.

The probability of obtaining a "positive" result favoring the new treatment and statistically significant at the two-sided 5% level is approximately

$$\Pr(\hat{\delta}/s \geq 2) = (1-\theta)\Phi(-2s/s) + \theta\Phi\left(\frac{\delta_1 - 2s}{s}\right),$$

where $\Phi$ denotes the standard normal distribution function. For $\delta_1/s \approx 3$, we obtain

$$\Pr(\hat{\delta}/s \geq 2) = (1-\theta)\Phi(-2) + \theta\Phi(1). \tag{3}$$

L&Z report that about 30% of phase III trials of the Eastern Cooperative Oncology Group are statistically significant. Assuming that the vast majority of these are significantly in favor of the new treatment, equating (3) to 0.30 and solving for $\theta$ gives approximately $\theta = 0.33$.

For $\theta = 0.33$, it follows from (2) that in order to have the posterior probability of the null hypothesis 0.1, we require BF = 17.7. If the sample size is planned in the frequentist manner as described above for most trials, then $\delta_1/s \approx 3$ and BF = 17.7 corresponds to $\hat{\delta}/s = 2.46$. Hence, in order to have any "statistically significant" result favoring the new treatment be associated with a posterior probability of the null hypothesis of no greater than 0.1, the critical value of 2.46 should be used for statistical significance. This corresponds to a two-sided significance level of 0.014. This is somewhat different from the claim of L&Z that a value of $\alpha$ in the range of 0.025–0.030 is appropriate.

L&Z have proposed two requirements for planning sample size. The first is that the finding of "statistical significance" be associated with a small posterior probability for the null hypothesis. The preceding paragraphs indicate that this leads to the requirement that the significance level should be no

greater than a two-sided 0.014. L&Z also proposed that the lack of finding of statistical significance should be associated with a large posterior probability for the null hypothesis. This is not uniformly possible because an outcome that is almost statistically significant carries approximately the same posterior probability as one that is just barely statistically significant. It is an inherent flaw in the Neyman–Pearson theory of hypothesis testing to sharply distinguish between falling just barely on one side of the rejection region boundary compared to falling just barely on the other side. It is an embarrassment to many biostatisticians to see biomedical investigators infer that since $p = 0.06$, the results are not statistically significant and the null hypothesis should be accepted. The embarrassment should not be for statistical naivete of the investigator, but rather for the inadequacy of the inferential framework that the field of statistics has provided for interpreting data. We should be careful not to force this defect onto Bayesian methods.

There are some outcomes that result in a high posterior probability for the null hypothesis. What is the largest value of the outcome $\hat{\delta}$ that results in a posterior probability of the null hypothesis of 0.9? With $\theta = 0.33$, we obtain from (2) that the outcome should correspond to a BF of 0.22 or less. This is less evidence against $\delta_1$ than was required for rejecting the null hypothesis (i.e., $1/0.22 = 4.6 < 17.7$) because the prior probabilities favor the null hypothesis. For a trial designed in the conventional way with $\delta_1/s \approx 3$, BF $= 0.22$ corresponds to $\hat{\delta}/s \approx 1$. So an outcome corresponding to a "$z$ value" no greater than 1 provides strong support against the alternative hypothesis used to design the trial when one considers the prior probabilities.

It follows from the above, that for a conventionally designed clinical trial, an outcome with a "$z$ value" $z = \hat{\delta}/s$ greater than 2.46 provides adequate support for rejecting the null hypothesis, and a $z$ value less than 1.0 provides adequate support for rejecting the alternative hypothesis. Whether the conventionally defined sample size is adequate may be addressed by computing the probability that the clinical trial provides a result that represents strong support for rejecting either the null or alternative hypothesis. As noted above, an inconclusive result corresponds to $1 \leq \hat{\delta}/s \leq 2.46$. We compute the probability of an inconclusive result with regard to the prior probability distribution, and find it to equal 0.197. If this is deemed too large, one can select a smaller value of $s$, corresponding to a larger sample size, recalibrate the upper and lower limits of $\hat{\delta}/s$ that correspond to strong posterior support for either the null or alternative hypothesis and then recompute the probability of an inconclusive result. One can automate this process to obtain any desired probability of an inconclusive result.

The above analysis provides a consistent Bayesian approach to planning the interpretation of results and planning sample size. The inference is based on the posterior probability of the null hypothesis given the data, as is required by Bayes theorem, not given that the test statistic was at an unspecified location in a semiinfinite interval. The approach is also Bayesian because the sample size is determined based on a figure of merit, the probability of obtaining conclusive results, which is an average with regard to the prior distribution. For the calculations above, this results in a frequentist power of only 0.71, but power is a non-Bayesian notion. The approach of L&Z uses the frequentist approach of establishing sample size to achieve a specified power under the alternative hypothesis.

The conclusion of the analysis presented here is that clinical trials whose sample size is based on the frequentist approach with $\delta_1/s \approx 3$ provide about an 80% probability of providing strong enough evidence to reject either the null or alternative hypothesis, where the evidence is based on Bayesian analysis. Although the conventional sample size planning approach appears adequate, our analysis indicates that the usual frequentist interpretations of the data are not adequate. Our analysis also shows that a critical value for significance should be about 2.46 and that only $z$ values less than 1 represent sufficient support for rejecting the alternative hypothesis in favor of the null. Of course, whether the approach to sample size planning is sensible depends on whether a sensible value of the alternative hypothesis is specified. This value should represent the smallest treatment difference which is of medical significance, given the costs and toxicities of the new treatment. For example, in the comparison of survival distributions with proportional hazards, $\delta$ may represent the natural logarithm of the hazard ratio and a $\delta_1 = \ln(1.33)$, representing a 25% reduction in the hazard rate is often used and considered reasonable. In this case, if $\delta_1/s \approx 3$, then $s = 0.095$ and this corresponds to observing approximately 444 events, using the approximation $s^2 = 4/(\#\text{ events})$. The conclusions derived here are based on the simple two-point prior distribution used. This approach to sample size planning and results interpretation can be carried out with more general prior distributions. The simple two-point model was used here only to clarify the concepts involved.

# Comment

**John Bryant and Roger Day**

Lee and Zelen propose an interesting approach to the design of Phase III clinical trials, and make a number of intriguing points. They correctly argue that the usual selection of type I and type II error rates is rarely predicated on rational consideration of losses or prior experience, but rather is usually based on tradition. Because the consequences of incorrect decisions are conditional on the trial results, they argue that $\alpha$ and $\beta$ are not the appropriate error rates to (directly) control. Instead, they propose to design the trial in such a way that posterior false negative and false positive error probabilities are controlled. Their analysis suggests that using the "traditional" 0.05 type I error rate may result in an excessive number of false positives, leading to the introduction of ineffective therapies into clinical practice. Based on similar arguments, Simon [(1982, 1994); see also Staquet, Rozencweig, Von Hoff and Mugia (1979)] has noted that, in studies of common diseases for which the success rate of clinical trials has been historically low, use of adequately powered 0.01-level tests might be preferable. Berger and Sellke (1987), from a different perspective, have also argued for using smaller type I error rates in many circumstances, based on computing bounds for Bayes factors. Below, we will discuss ways to incorporate the need for eventual clinical consensus into the trial design process, and show that a simple approach based on the notion of prior robustness also leads to similar conclusions.

Lee and Zelen's development treats both the null and alternative hypotheses as point hypotheses. This is both a strength and a weakness. In reality, the alternative is composite (and indeed the null hypothesis is also). By ignoring this, Lee and Zelen greatly simplify the method and make it more practical to perform computations. This approach also facilitates the specification of prior information, making possible a rough assessment such as their estimate of $\theta$ based on the ECOG experience. On the other hand, unless one believes that the true prior is actually tightly concentrated about points representing the null and alternative hypotheses, difficulties of interpretation and implementation may occur, for two reasons: (i) For composite hypotheses, type I and type II error rates as defined in Section 2, which are conditional on the truth of $H_0$ and $H_A$, respectively, differ from the "usual" definitions used in the frequentist formulation of the hypothesis test, which are conditional on specific values of $\delta$, and (ii) If the prior places significant probability on a region in which $\delta$ is nonzero but of negligible clinical import, then one would not want to equate the acceptance of conclusions that are only "marginally" false with acceptance of those which are egregiously false. A complete Bayesian analysis would take this into acount by the specification of an appropriate loss function. The authors note that one might adopt the notion of "indecisive" regions separating $H_0$ and $H_A$, which would be more in keeping with the philosophy of frequentist significance testing, but this idea is not explicitly developed in the paper. The discontinuous loss function that their method implicitly uses, when coupled with a smooth prior, may lead to much larger sample sizes than would be practical. For these reasons, when the prior is believed to be smooth rather than sharply bi- or tri-modal, it may not always be apparent how their method should be applied.

It may be of concern that the posterior probabilities which are controlled by Lee and Zelen's method are not the ones which would be of interest after the trial is completed. Assuming that the test statistic is sufficient, the posterior probability that $T = +$ (i.e., that $H_A$ is true) is $\Pr\{T = +|Z = z\}$, where $z$ is the observed value of the standardized test statistic $Z$. The probabilities $\Pr\{T = +|C = -\}$ and $\Pr\{T = -|C = +\}$ are relevant only if one knew only that the hypothesis test was "significant" or "not significant." These controlled probabilities are averages of the appropriate posterior probabilities, weighted by the predictive distribution $g(\cdot)$ of $Z$; for

*John Bryant is Associate Professor, Departments of Statistics and Biostatistics, and Director of the National Surgical Adjuvant Breast and Bowel Project Biostatistical Center at the University of Pittsburgh, Pittsburgh, Pennsylvania 15213 (e-mail: bryant@nsabp.pitt.edu). Roger Day is Associate Professor, Department of Biostatistics, and Director of the University of Pittsburgh Cancer Center Biostatistical Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.*

example,

$$\Pr\{T = -|C = +\} = \int_{|z|>z_{\alpha/2}} \Pr\{T = -|Z = z\} g(z)\, dz \Big/ \int_{|z|>z_{\alpha/2}} g(z)\, dz.$$

Thus posterior use of the controlled probabilities $\Pr\{T = +|C = -\}$ and $\Pr\{T = -|C = +\}$ implies a disregard of much relevant available information, namely the value of $Z$. For example, suppose we test $H_0$: $\delta = 0$ against $H_A$: $|\delta| = \delta_{\text{ALT}}$, assuming that $\Pr\{\delta = 0\} = 1-\theta = 0.75$ and $\Pr\{\delta = \delta_{\text{ALT}}\} = \Pr\{\delta = -\delta_{\text{ALT}}\} = \theta/2 = 0.125$. To control the posterior error rates $\beta^* = 0.05$ and $\alpha^* = 0.10$, we choose type I and type II error rates $\alpha = 0.0118$ and $\beta = 0.3294$ from Table 1 in Lee and Zelen. However, if the trial results are only marginally significant ($p$-value just slightly less than 0.0118; $|Z|$ slightly greater than 2.52), computations show that the posterior probability that $H_0$ is true given the value of $z$ is about 0.22, not 0.05. (For a one-sided test, the posterior probability for $H_0$ would be about 0.20).

We suggest that directly bounding relevant posterior probabilities may be preferable to controlling average posterior probabilities. Consider a one-sided hypothesis test $H_0$: $\delta \leq 0$ versus $H_A$: $\delta > 0$. We focus on the one-sided case primarily for notational convenience. However, as has been noted by the authors, in the common situation of a trial comparing an experimental regimen B to an accepted standard A, the hypotheses are generally altered in practice to be one-sided, by adoption of asymmetric early stopping rules that would terminate the trial prior to accumulating sufficient evidence that $\delta < 0$, if early results were to strongly favor A. Let $\delta_1$ be a "minimally clinically significant" treatment effect, perhaps equal to the "planning $\delta$" used by the frequentist to power the hypothesis test. In lieu of bounds on $\Pr\{T = +|C = -\}$ and $\Pr\{T = -|C = +\}$, one may determine $\alpha$ and $n$ in such a way that quantities such as

$$M_1 = \max_{z \leq z_\alpha} \Pr\{\delta > \delta_1 | Z = z\}$$

and

$$M_0 = \max_{z > z_\alpha} \Pr\{\delta \leq 0 | Z = z\}$$

are suitably small. In this case, if $H_0$ were rejected, we would be assured that the posterior probability of at least some effect was acceptably high, whereas if $H_0$ were accepted, we would be assured that the posterior probability of a "material" effect (defined as $\delta > \delta_1$) was acceptably low. This bounding of posterior probabilities more closely conforms with the basic intent of frequentist significance testing (which essentially bounds these probabilities under a noninformative prior).

It may be advisable to use two different priors to compute these bounds: $M_1$ might be determined under a prior favoring the hypothesis of a material treatment effect, whereas $M_0$ might be computed under a prior favoring the null. This is because the trial will have limited practical consequence unless a broad consensus exists among the clinical community regarding its findings. Thus its design should take into account the goal of achieving consensus. A good design should have sufficient sample size to overcome prior differences of opinion, so that effective consensus can be achieved.

To elaborate, one should consider the study data in light of prior distributions representing the beliefs of both a very skeptical observer, who a priori assigns little probability to the possibility that the new treatment will be materially more effective than the currently available standard, and a very optimistic observer, who believes this possibility to be rather likely. Substantive posterior agreement between these two observers, either to the effect that the new treatment is highly likely to be superior to the control, or that it is highly unlikely to be materially superior to the control, would give reason to conclude that the trial results will compel consensus. The use of "skeptical" and "optimistic" priors in clinical trial design and monitoring is discussed by Kass and Greenhouse (1989), Freedman, Spiegelhalter and Parmar (1994) and Spiegelhalter, Freedman and Parmar (1994); Their application to an assessment of the need for confirmatory trials is addressed by Parmar, Ungerleider and Simon (1996). These references also address the construction of more-or-less canonical skeptical and optimistic priors that may be useful in practice. Freedman, Spiegelhalter and Parmar (1994) suggest that an optimistic prior may be approximated by a normal distribution centered at value $\delta_1$, the planning value or alternative hypothesis at which the trial is powered. The optimist believes that the new treatment is quite likely to show at least some benefit relative to the control, so that according to his or her prior $\Pr\{\delta > 0\}$ is large, perhaps 0.95. In this case, the prior standard deviation is found be to $\delta_1/1.645$. On the other hand, the skeptic believes that the new treatment is unlikely to offer material improvement over the control. Thus his or her normal prior is centered at $\delta = 0$, and assuming a prior standard deviation of $\delta_1/1.645$, according to this prior $\Pr\{\delta < \delta_1\} = 0.95$.

Therefore, in designing a hypothesis test of $H_0$: $\delta \leq 0$ versus $H_A$: $\delta > 0$, the following requirements operationally define the notion of posterior consensus:

(i) Given any test result $Z = z$ leading to the rejection of $H_0$, both the skeptic and the optimist must agree that the new treatment is superior to the control, in the sense that both must assign a posterior probability $\geq 1 - \beta^*$ to the event $\delta > 0$.

(ii) Given any test result $Z = z$ leading to the acceptance of $H_0$, both the skeptic and the optimist must agree that the new treatment is not materially superior to the control, in the sense that both must assign a posterior probability greater than or equal to $1 - \alpha^*$ to the event $\delta \leq \delta_1$.

We now show how to design a test meeting these requirements, using as an example a one-sided comparison of an experimental treatment B to an established standard A. We suppose that efficacy is measured by survival time, and that the final analysis of results will take place after $d$ deaths have been observed. The efficacy of the proposed treatment relative to control is summarized by the logged hazard ratio $\delta$, assumed to be time independent. $\delta$ is defined so that values of $\delta > 0$ imply that treatment B is superior to A, while values of $\delta < 0$ favor treatment A. The test statistic is the maximum partial likelihood estimator of $\delta$ (asymptotically, this is equivalent to using the logrank test). The variance of this estimator $\cong 4/d$, assuming equal randomized allocation of patients to either treatment arm and $|\delta|$ moderately close to 0. In the notation of Lee and Zelen, $d = 2n$, so that $\sigma^2 = 1$. In this situation, an application of Bayes theorem leads to the required sample size and type I and II error rates, as follows:

(i) Compute the "prior sample size" $d_0 = (2 \cdot 1.645/\delta_1)^2$. (This is the number of deaths which would be required to generate information equivalent to that reflected by the priors of either the skeptic or the optimist.)

(ii) Compute the number of required events $d$ by solving the quadratic equation

$$d^2 - (d + d_0) \cdot 4(z_{\alpha^*} + z_{\beta^*})^2/\delta_1^2 = 0.$$

[The formula for $d$ can be written as $d = (1 + d_0/d) \cdot 4(z_{\alpha^*} + z_{\beta^*})^2/\delta_1^2$, which is identical to the "usual" formula (5) (with $d = 2n$ and $\sigma = 1$) except for the factor $(1 + d_0/d)$.]

(iii) The type I and type II error rates for the hypothesis test are found from

$$z_\alpha = [1 + d_0/d]^{1/2} z_{\beta^*}, \qquad z_\beta = [1 + d_0/d]^{1/2} z_{\alpha^*}.$$

As an example, suppose $\alpha^* = 0.10$, $\beta^* = 0.05$ and $\delta_1 = |\log(0.75)| = 0.2877$, that is, a 25% reduction in hazard rate is considered to be a "material" or clinically significant treatment effect. Then $d_0 = 131$, $d = (1 + 131/d) \cdot 414 => d = 519$, $\alpha = 0.033$, and $\beta = 0.076$. The required number of deaths is about 25% larger than called for by the usual frequentist calculation. The prior sample size $d_0$ is a measure of prior discordance of opinion, which influences the required sample size $d$. For example, if both the skeptic and the optimist were a priori willing to admit as much as a 10% chance of the correctness of the other's position, then $d_0 = (2 \cdot 1.282/\delta_1)^2 = 79$. In this case the usual sample size is increased by only 16%, to $d = 482$.

Consistent with the recommendations of Lee and Zelen, this approach leads to a more stringent Type I error requirement than the traditional 0.05 level. In contrast, the power requirements are also more stringent. The method we have described utilizes priors to model the notion of discordance, and Bayes theorem provides a model for convergence of opinion. The method can also be thought of as requiring a robustness of conclusions over a range of priors. This is in contrast to traditional Bayesian experimental design, and to Lee and Zelen's method, which are intended to allow a particular prior to influence the study design and the inference.

# Rejoinder

**Sandra J. Lee and Marvin Zelen**

We thank Drs. Bryant, Day and Simon for their comments on our paper. It is worth noting that all agree that it might be more appropriate to plan clinical trials on humans in a Bayesian context. Furthermore the discussants have concluded that the standard 5% significance level is too liberal and should be made more stringent.

Our basic philosophy is that the special circumstances associated with clinical trials require ethical concerns as well as providing preliminary informa-

tion which cannot be ignored. As we have shown, it is straightforward to use this information in the planning of trials. The strength of our formulation of utilizing the posterior probabilities, $P_1$ and $P_2$, lies in its simple relationship to $\alpha$ and $\beta$ for a specified prior $\theta$. This relationship leads to an easy way of calculating the sample size for specified $P_1$ and $P_2$.

One point that we wish to elaborate further relates to the prior distribution. Although the prior distribution is formally defined to be $\theta = \Pr\{|\delta| > 0\}$, when one chooses the pair $(\delta, n)$ from (5), it is essentially recasting the alternative to be $H_1$: $|\delta| = \delta_1$ with $\theta = \Pr\{|\delta| = \delta_1\}$. Therefore, we are assuming a three-point prior distribution; that is, $1 - \theta = \Pr\{\delta = 0\}$ under $H_0$ and $\theta/2 = \Pr\{\delta = \delta_1\}$ or $\theta/2 = \Pr\{\delta = -\delta_1\}$ under $H_1$ in a two-sided testing setting. As a result, different values of $n$ will generate different values of $\delta_1$, but the prior probability is fixed at $\theta$.

Both discussants note the earlier references to Staquet, Rozencweig, von Hoff and Mugia (1979) and Simon (1982). It is of some interest that, in 1979, the National Cancer Institute conducted a review of the Cancer Clinical Trials Cooperative Group Program in which one of us (MZ) made a presentation which gave tables of the ratios of posterior false positive to posterior true positive probabilities; see Zelen, Gehan and Glidewell (1980).

Bryant and Day present a somewhat different approach to the formulation of the sample size problem. We are at a loss to comprehend the remark in which with $\alpha = 0.0118$ and $\beta = 0.3294$, they claim that if the $p$-value is slightly less than 0.0118, the posterior probability that $H_0$ is true is about 0.22, not 0.05. Our calculations show that the posterior probability is in the neighborhood of 0.05; that is, for $0.005 \leq \alpha \leq 0.01$ the value of $\alpha^* = 0.05$.

We view with amusement that employment of a "skeptical" and "optimistic" observer. It appears to be a game. There will always be one more extreme skeptic or optimist, which may change how one proceeds. However, the sample size calculations in Tables 2 and 3 show that for the range of prior probabilities from $\theta = 0.25$ to 0.75, the calculated sample sizes do not change very much. However, skeptics and optimists having prior probabilities close to zero or unity will greatly affect the sample size calculations.

Bryant and Day suggest using two different priors in computing posterior probabilities. They illustrate their methods with a numerical example which results in a $d = 519$ (number of deaths) to detect a 25% reduction in failure rates among two populations, with a one-sided test using posterior error rates $\alpha^* = 0.10$ and $\beta^* = 0.05$. We have also used our proposed methodology to estimate the required sample size for the same parameters, but varying the prior probability. Over the range $(0.25 \leq \theta \leq 0.75)$, the value of $d$ does not change very much $(335 \leq d \leq 415)$. It is only with an extreme value of $\theta$ that the value of $d$ is in the neighborhood of 500. According to our calculations, a value of $\theta = 0.14$ generates $d = 510$ and a value of $\theta = 0.93$ results in $d = 533$. These values increase rapidly as $\theta$ approaches the boundary values $1 - P_1 < \theta < P_2$. Hence Bryant and Day's calculation corresponds to having a low prior probability that $\delta \neq 0$. We do not consider $\theta = 0.93$ to be realistic as it places too small prior probability on $\{\theta = 0\}$ for a trial to be initiated. Also we could conjecture that a prior probability of $\theta = 0.14$ is too small to generate a clinical trial.

Simon's contributions mainly relate to the analysis of the trial. Our view is that any suitable method is appropriate for analysis whether it be frequentist, Bayesian or likelihood. Whichever analysis method is utilized does not affect our recommendation on sample size. Simon comments on the inadequacy of small trials and the illogic of adopting a hypothesis testing point of view. We concur with his criticisms. The significance level of a test is not sacrosanct. It would be difficult to ignore a $p$-value of $p = 0.03$ if the predesigned level was $\alpha = 0.025$. Our recommendation is to recalculate the posterior error probabilities with the observed $\alpha$.

Simon takes issue with our formulation that the prior distributions must be symmetric about $\delta = 0$ for the clinical trial to be ethical. If this is not so, then there is an advantage to the patients to be assigned to the treatment with the highest probability of benefit. We find this position difficult to defend and are surprised that some biomedical ethicists may agree with this position. Not that we deny the fact that in many clinical trials, the prior distributions are asymmetrical. Of course, there may be issues of defining benefit, in that overwhelming toxicity may negate positive advantages. It is possible that a series of informal studies, pilot studies and phase II trials may have built up a body of evidence that a therapy is beneficial. Then it may be impossible to carry out a randomized phase III trial comparing new therapy with standard therapy as the prior distribution on benefit is so high.

Simon criticizes our proposal on two points; that is, there is no prior on specific values other then $\delta = 0$ and we did not use a proper likelihood. We have answered the first criticism earlier by pointing out that the choice of $(\delta, n)$ in (5) essentially reformulates the problem by having prior probabilities on the chosen $|\delta| = \delta_1$. We also note that the

calculation of the test may use a statistic based on the likelihood function.

Simon presents a Bayesian analysis of a trial from which a sample size can be calculated. In the context of ECOG trials, Simon computes $\theta$ to be 0.33 based on assumptions of two-sided $\alpha$ level of 0.05, 80–90% power and $\delta_1/s = 3$. His formulation results in a two-sided test with $\alpha = 0.014$ and $\beta = 0.29$. This contrasts to our method of initially choosing $P_1 = 0.95$ and $P_2 = 0.90$ (based on the two-sided $\alpha = 0.05$ and $\beta = 0.10$). Since Simon focuses on "positive" results in the two-sided setting, he is essentially formulating the problem to be one-sided with an $\alpha$ level of 0.025 in his analysis.

In summary, we propose that sample size for all clinical trials be estimated according to our methods. Not to do so may continue to introduce an excessive number of false positive therapies in the practice of medicine. Error rates should be focused on posterior probabilities which are important in extending the conclusions from a trial to the practicing physician. Our method is easy to implement and makes use of prior information. It serves to rationalize the choice of significance levels and power in planning clinical trials. Calculations have shown that the sample sizes are reasonably robust with regard to prior probabilities in the middle range ($0.25 \leq \theta \leq 0.75$). The nonrobustness occurs at the extreme ranges of the prior probabilities which are unlikely to generate a clinical trial to test hypothesis for which there is very strong belief or disbelief.

In order to be more focused, we have not elaborated on the principal idea. As we have noted, one can use indifference regions or even attempt to construct a complete prior distribution on $\delta$. These generalizations are straightforward. Realistically, the more one attempts to make use of detailed prior distributions the more likely the conclusions will be nonrobust due to the issues in specifying detailed prior distributions.

There are some statistical applications which can only be carried out using Bayesian ideas. In our opinion, this is one of them. It is not an issue of faith or adherence to Bayesian ideas. It is simply doing the right thing.

## ADDITIONAL REFERENCES


BERGER, J. and SELLKE, T. (1987). Testing of a point null hypothesis: the irreconcilability of significance levels and evidence. *J Amer. Statist. Assoc.* **82** 112–139.

FREEDMAN, L. S., SPIEGELHALTER, D. J. and PARMAR, M. K. B. (1994). The what, why and how of Bayesian clinical trials monitoring. *Statist. in Med.* **13** 1371–1383.

KASS, R. and GREENHOUSE, J. (1989). Comment on "Investigating therapies of potentially great benefit: ECMO" by J. Ware. *Statist. Sci.* **4** 310–317.

PARMAR, M. K., UNGERLEIDER, R. S. and SIMON, R. (1996). Assessing whether to perform a confirmatory randomized clinical trial. *J. Nat. Cancer Inst.* **88** 1645–1651.

SIMON, R. (1982). Randomized clinical trials and research strategy. *Cancer Treatment Reports* **66** 1083–1087.

SIMON, R. (1994). Some practical aspects of the interim monitoring of clinical trials. *Statist. in Med.* **13** 1401–1409.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials (with discussion). *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

STAQUET, M. J., ROZENCWEIG, M., VON HOFF, D. D. and MUGIA, F. M. (1979). The delta and epsilon errors in the assessment of clinical trials. *Cancer Treatment Reports* **63** 1917–1921.

ZELEN, M., GEHAN, E. and GLIDEWELL, O. (1980). Biostatistics. In *Cancer Research: Impact of the Cooperative Groups* (B. Hoogstraten, ed.) Chapter 15. Masson Publishing, New York.