[Clinical Trials and Sample Size Considerations: Another Perspective]: Comment

Richard Simon

Statistical Science, Vol. 15, No. 2. (May, 2000), pp. 103-105.

Stable URL:
http://links.jstor.org/sici?sici=0883-4237%28200005%2915%3A2%3C103%3A%5BTASSC%3E2.0.CO%3B2-G

Statistical Science is currently published by Institute of Mathematical Statistics.

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 δ denote the true treatment effect, as in L&Z. Let $\hat{\delta}$ denote the maximum likelihood estimate of δ. We will assume that $\hat{\delta}$ is sufficient for δ 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 δ has a two-point prior distribution which assigns probability $1 - \theta$ to the null hypothesis that $\delta = 0$ and probability θ 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 ϕ 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 Φ 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 θ 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 α 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 δ_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, δ 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.