Monitoring Clinical Trials: Conditional or Predictive Power?

David J. Spiegelhalter, Laurence S. Freedman, and Patrick R. Blackburn

MRC Biostatistics Unit (D.J.S.; P.R.B.) and MRC Cancer Trials Office, Clinical Oncology and Therapeutics Unit (L.S.F.), Cambridge, England

ABSTRACT: At an interim point in a clinical trial, trial organisers may wish to use the data on the initial series of patients to judge the likely consequences of further patient accrual. Halperin and colleagues (*Controlled Clin Trials* 3:311–323, 1982) have suggested calculating the power of a continued trial, *conditional* on the data observed so far and the null and alternative hypothesis specified at the start of the trial. Here we argue that this idea should be extended to obtain the *predictive* power of the trial, derived by averaging the conditional power with respect to the current belief about the unknown parameters. Although numerical methods are generally required for evaluating the necessary integrals, the results may be presented graphically and enable the statistician to answer the question: "With the data so far, what is the chance that the trial will end up showing a conclusive result?"

KEY WORDS: *posterior distribution, predictive distribution*

INTRODUCTION

When faced with interim results from a long-term clinical trial, a collaborative group will weigh many factors in determining whether early termination of the experiment is appropriate. Halperin and colleagues [1] have recently suggested a statistical aid to this difficult decision based on the framework of "stochastically curtailed tests." Specifically, they recommend estimating two conditional probabilities of the null hypothesis being rejected were the trial to continue to its planned extent; the first probability is conditional on the data observed so far and the null hypothesis being true, whereas the second is conditional on the current data and the truth of an alternative hypothesis specified as the "expected" improvement at the start of the trial. These two probabilities may be considered as values of a *conditional power* function. Andersen [2] has also recently suggested this approach to interim analyses in clinical trials with survival time as the response variable.

Address reprint requests to: David J. Spiegelhalter, Ph.D., MRC Biostatistics Unit, 5 Shaftesbury Road, Cambridge CB2 2BW, England

Received February 25, 1985; revised December 8, 1985.

0197-2456/86/$3.50

Controlled Clinical Trials 7:8–17 (1986)

52 Vanderbilt Ave., New York, New York 10017

In this article we argue that these suggestions fall short of being an intuitive, rational aid. In particular, simply conditioning on selected hypotheses that were specified at the start of the trial ignores the considerable knowledge concerning the treatment effects that has accumulated by the time of interim analysis. In particular, there seems little point in being concerned about power conditional on hypotheses that are no longer fairly plausible. Hence we suggest extending the "conditional" analysis to an "unconditional" prediction of the consequences were the trial to continue. Since this essentially involves averaging the conditional power function with respect to the current opinion about the treatment effect, we are led to a Bayesian approach but one that need not necessarily be based on subjective judgments, but on the results of the first part of the trial.

We should emphasize that we are not concerned with formal stopping criteria corresponding to prespecified sequential designs with fixed Type I and Type II error [3]. Such designs only have their stated properties if the decision to stop is based purely on a formal statistical argument, which we believe is seldom appropriate. Rather, our approach is a response to the question we find is often posed to the statistician in a collaborative group: "What, given the data so far, are the chances of getting a conclusive result if we complete the trial?" It is unreasonable that such an intuitive request is currently unanswered by conventional methodology.

We first describe the analytic steps in a general notation, and then in the following section we illustrate our argument using the particular case of two-group binomial sampling. Let X_f denote the future data to be observed in the remaining part of the trial, and x_0 the data observed so far. It is assumed that X_f and x_0 are both random samples from a distribution conditional on the unknown parameters θ that describe both the treatment effect of interest and any other nuisance parameters. We define a critical region R of future observations, so that $X_f \in R$ implies that the overall data (X_f, x_0) will lead to a "firm conclusion."

Clearly the definition of "firm conclusion" depends on the aims of the study, and possibly on the statistical ideology of the investigators. Although fully aware of the limitations of hypothesis tests as summaries of a trial, we assume here, for the sake of familiarity, that the trial participants are concerned with a significance test of a null hypothesis defined on θ, where a "firm conclusion" represents rejection of the null hypothesis at the $100\alpha\%$ level. The region R will depend on x_0, the final sample size, and the significance level α: we assume no adjustment is made in the significance level for the possibility that the trial may have been terminated early. Other, more preferable forms of "firm conclusions," such as confidently asserting treatment equivalence within a certain range, may also be treated within this framework.

Firstly, the conditional power function $p(X_f \in R|\theta)$ should be calculated and plotted. Secondly the current belief concerning θ should be plotted as a *posterior distribution* $p(\theta|x_0)$, derived via Bayes's theorem from a pretrial prior distribution $p(\theta)$; $p(\theta)$ will generally be "noninformative" [4], so that our predictions can be considered as being based only on the data observed so far. Thirdly the conditional power may be averaged with respect to the posterior distribution to produce an unconditional, predictive probability of rejecting the null hypothesis:

$$p(X_f \in R) \mid x_0) = \int p(X_f \in R|\theta)p(\theta| x_0)d\theta.$$

Finally, we note that it may be more "communicative" to calculate the predictive judgments in an alternative manner. The rejection region R may be explicitly calculated and superimposed on a plot of the *predictive distribution* of the future data conditional on the observed data so far:

$$p(X_f \mid x_0) = \int p(X_f \mid \theta)p(\theta|x_0)d\theta.$$

The predictive probability of rejecting the null hypothesis is then obtained by integrating the predictive distribution over R. A general discussion of the predictive approach in a biometrical context is provided by Geisser [5].

The above description may appear rather technical, and numerical methods will usually be required to evaluate the necessary integrals. However, the essential results may be presented graphically so as to preserve the intuitive appeal of the approach. In the next section, we present a reanalysis of Halperin's example, in which simplifying assumptions are adopted in order to emphasize the conceptual rather than the technical issues.

AN EXAMPLE IN BINOMIAL SAMPLING

Suppose a trial is being conducted in which a binomial response is observed in two groups labeled "control," (C) and "treatment" (T), in which p_C and p_T are the unknown underlying probabilities of the event of interest occurring to a patient in the respective groups. At an interim point in the trial, suppose m_C control patients and m_T treatment patients have been entered and r_C and r_T events observed in the two groups. The trial organizers are considering an extension of the trial to include a further n_C control and n_T treatment patients and wish to assess the likely consequences, expressed as the chance that this extension will lead to a rejection of the null hypothesis $p_C = p_T$ by a one-sided test at the $100\alpha\%$ level.

Let S_C and S_T denote the random number of events to be observed in the two future groups, making the necessary assumption that the remainder of the trial is carried out in the same circumstances as the initial part. Thus, in the notation of the previous section, $X_f = (S_C, S_T)$, $x_0 = (r_C, r_T)$ and $\theta = (p_C, p_T)$. Then the critical region R consists of those future observations that will lead to rejection of the null hypothesis, i.e.,

$$R = [S_C, S_T|Z > k_\alpha],$$

where k_α is the upper $100(1 - \alpha)\%$ point of a standard normal distribution, and Z is the approximate standardized normal statistic with continuity correction calculated on the basis of comparing observed proportions $(r_C + S_C)/(m_C + n_C)$ and $(r_T + S_T)/(m_T + n_T)$. The conditional power function is defined as

$$p(S_C, S_T \in R|r_C, m_C, n_C, p_C, r_T, m_T, n_T, p_T)$$

and a good approximate formula for this is provided in Appendix 1.

Halperin and colleagues introduce an example that they analyze using the conditional power argument. They consider 1249 patients in each of a control

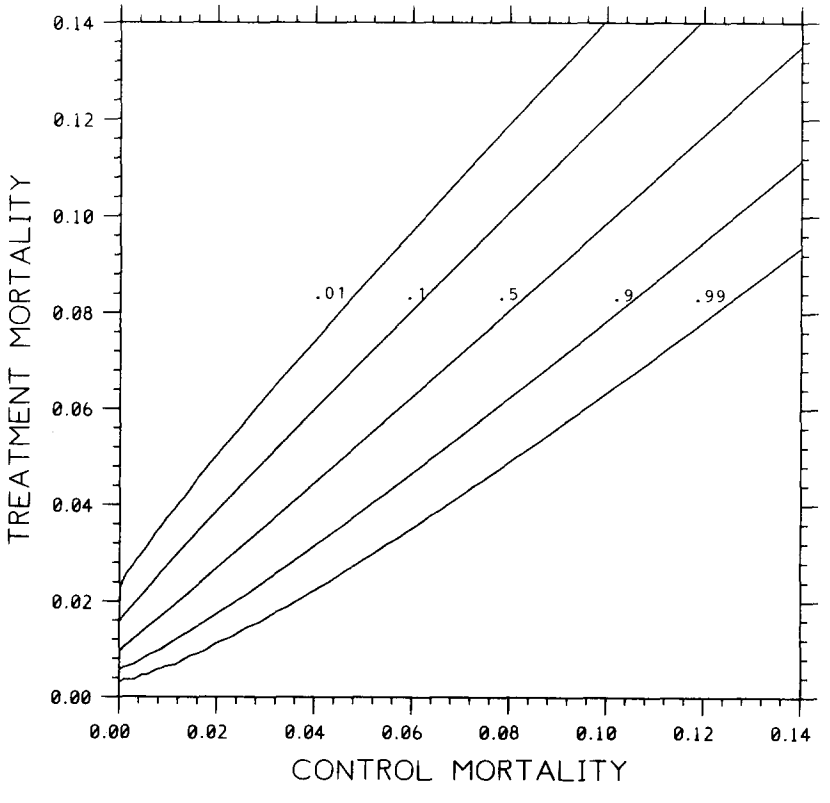
and treatment group, who are all to be observed for 5 years. After 3 years, 67 of the control group group and 43 of the treatment group have died. They envisage the prospect of testing the hypothesis $p_C = p_T$, where these denote the respective 5-year death rates, after observing for a further 2 years.

Our proposals are primarily intended for trials with staggered accrual and short-term outcome, whereas Halperin's example has all patients entered at the outset with prolonged follow-up. Nevertheless, in order to compare the alternative conceptual approaches, we can reformulate Halperin's study within our terms without too much distortion.

Specifically, if we assume death occurs at a low constant hazard rate, then, when put in terms of "patient-five-years" the study may be roughly considered as being equivalent to having observed $3/5 \times 1249 = 887 = m_C = m_T$ patients each with probability p_C or p_T of death, with the prospect of a further $2/5 \times 1249 = 592$ patients to come in each group, and $r_T = 43$, $r_C = 67$. Figure 1 shows conditional power functions derived as shown in Appendix 1 for $\alpha = 0.05$, plotted as contours.

We note that the contour heights decline as the parameters tend to those likely to lead to future data that would cancel out the current superiority

Figure 1. "Conditional power" contours showing the probability of rejecting $H_0 : p_C = p_T$ in favor of $p_C > p_T$ given the data so far, as a function of true mortality rates p_C and p_T.



shown by the treatment group. Halperin et al. calculate, using a similar approximate formula, that the conditional power when $p_C = p_T = 0.03$ is 0.54 compared to our 0.48, and under the alternative hypothesis that $p_T = 0.7 \times p_C = 0.021$ their conditional power is 0.97 compared with our 0.96; the agreement is reasonable considering our reformulation.

In the previous section we described how we would extend their argument. Essentially we should calculate not only the conditional power function, but also the current plausibility of the various values of p_C and p_T. This "belief" is expressed as posterior distributions for p_C and p_T, derived from the data currently available, as described in Appendix 2 and displayed in Figures 2a and 2b. Making the simplifying assumptions of independent pretrial opinions about control and treatment responses, the posterior beliefs are independent beta distributions from which a joint distribution may be drawn as a contour plot (Fig. 2c).

By superimposing Figures 1 and 2c it can be seen that the bulk of belief given the data so far is concentrated on values of the parameters that, if true, would almost certainly lead to rejection of the null hypothesis, were the trial to continue. Indeed, the observed data already provide substantial evidence against H_0. The current Z statistic, based on comparing proportions 43/887 and 67/887, is 2.26 with a one-tailed p value of 0.012; and the posterior probability that p_T is greater than p_C obtained by numerically integrating the posterior distribution over the upper triangle of Figure 2c is equal to 0.009. Thus both classical and Bayesian philosophical approaches would infer substantial evidence for p_T being less than p_C on the data so far, although neither approach would necessarily recommend stopping at this point: the classical statistician may wish to adjust the overall p value for the interim analyses that have been planned whereas the Bayesian may demand more certainty because of the cost of drawing a false conclusion.

This example shows that it may well seem reasonable to terminate the trial immediately, even though were the trial to continue Figure 1 shows there is still about a 50–50 chance of rejecting H_0 (control mortality = treatment mortality) if it is actually true: this latter point does not seem to provide genuine grounds for concern, given the unlikeliness that H_0 actually *is* true.

As a summary to help in the decision whether to terminate the trial, it is useful to assess an overall predictive probability of rejecting H_0 were the trial to continue. This is obtained by numerically integrating the conditional power function of Figure 1 with respect to the joint posterior distribution in Figure 2c. We may do this separately for the values of p_T and p_C in the lower and upper triangle to obtain the summary shown in Table 1.

We note that overall there is about a 95% chance of rejecting the null hypothesis at the 5% level were the trial to continue. This may seem surprisingly low in view of the fact that in a classical analysis the null hypothesis could already possibly be rejected with the data in hand, depending on the predefined stopping criteria; however, the predictions take into account the possibility that future observations may serve to cancel out the effect observed so far.

There is an alternative, and possibly more intuitive, way of deriving and displaying the above conclusions, in which the critical region R is explicitly calculated by numerical solution of the equation $Z = k_\alpha$. The joint *predictive*

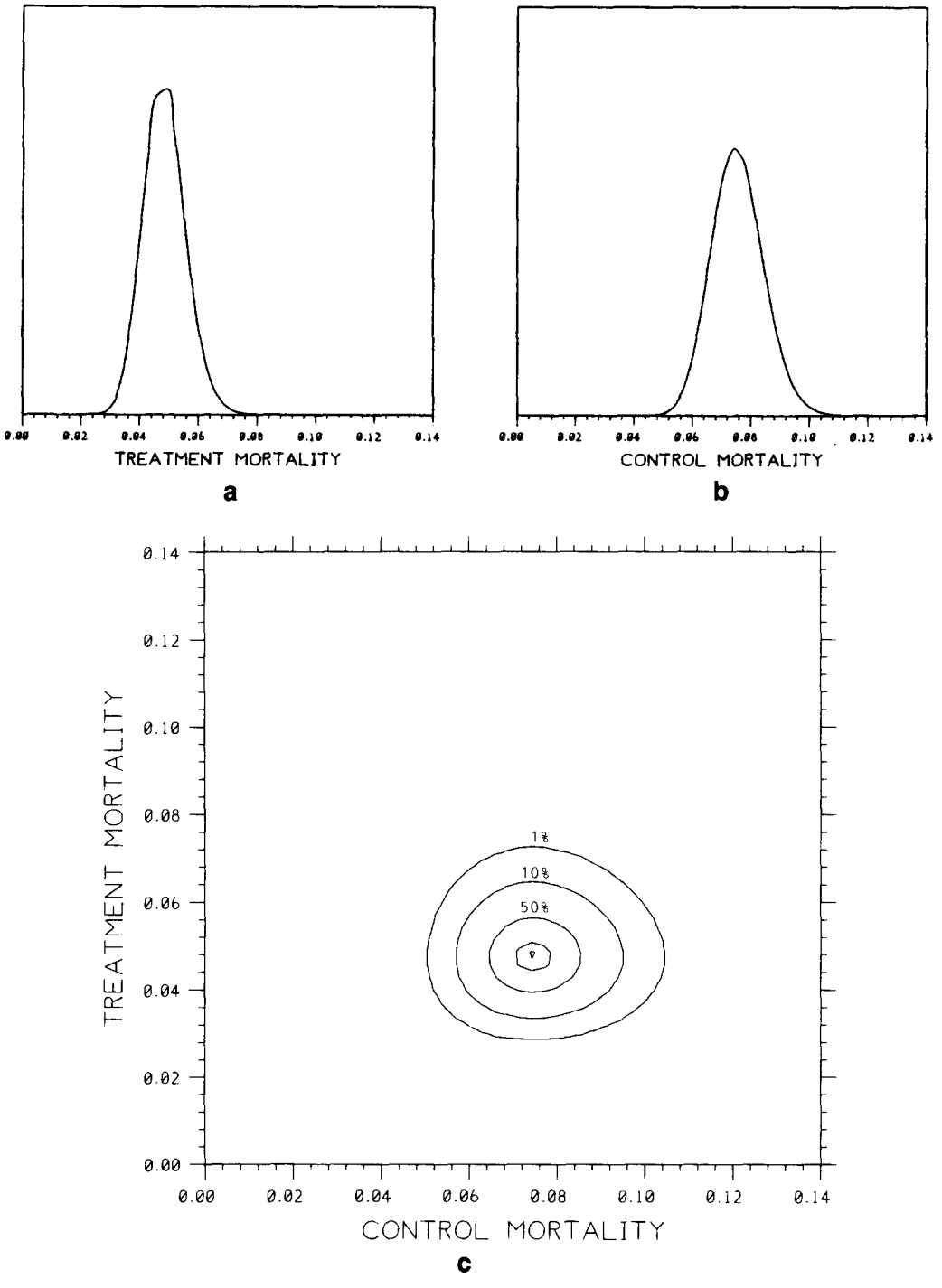


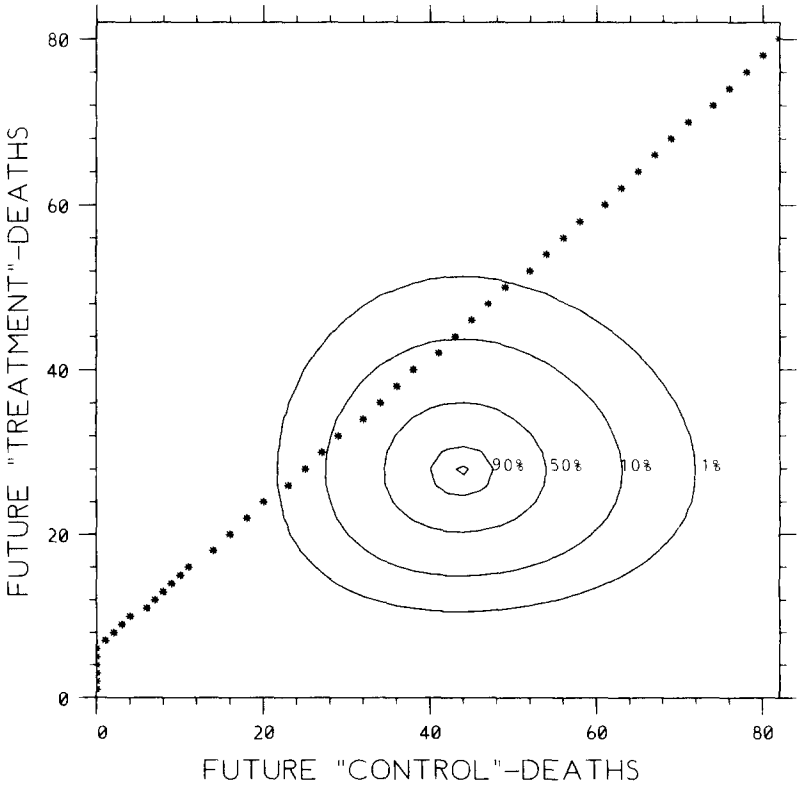
Figure 2(a). Current marginal belief concerning p_T. 2(b). Current marginal belief concerning p_C. 2(c). Current joint belief concerning p_C and p_T, expressed as contours of 1%, 10%, 50%, 90%, and 99% of the maximum ordinate, assuming pretrial independence of p_C and p_T.

Table 1. Predictive Probabilities of Eventual Conclusion Related to True Hypothesis

Conclusion at End of Trial	True Parameter Order $p_C < p_T$ (Control Superior)	$p_C > p_T$ (Treatment Superior)	
Reject H_0 in favor of $p_0 > p_T$	0.004	0.946	0.950
Do not reject H_0	0.005	0.045	0.050
	0.009	0.991	1.000

distribution of S_C and S_T, derived in Appendix 3, may then be superimposed on R to produce the display in Figure 3. It is clear that the bulk of the plausible future values of S_C and S_T lie inside R, and numerical integration over R reveals an independent check of the predicted power of 0.950. A picture such as Figure 3 with the numbers contained in Table 1 seems to provide the

Figure 3. Joint predictive distribution of future number of deaths, superimposed on critical region R (lying to the right of the asterisks), where R includes values leading to the rejection of H_0 at the 5% level. Contours shown are 1%, 10%, 50%, 90%, and 99% of the maximum ordinate.



precise, quantitative expression of the considerations usually required to aid decision making during the progress of a clinical trial.

DISCUSSION

We have found this technique particularly useful when the early part of a trial shows some evidence against a null hypothesis, although not as strong as in our example above, and the participants wish to know whether it is worthwhile continuing. Of course, a predicted probability of rejecting the null hypothesis, were a further $n_T + n_C$ patients enrolled, does not provide the answer to whether to carry on; the decision may depend on the current direction of the departure from the null hypothesis, the cost of the proposed extension, and the many other criteria that influence whether to terminate a trial prematurely. We emphasize that, although it provides a valuable aid to the team's decision, "predicted power" should not be treated as a formal criterion for stopping.

The calculations, in general, require numerical solution. However, we have found a simple approach based on a fairly fine, equally spaced, 60 × 60-grid followed by simple summation quite adequate, as is reflected in the agreement obtained in the two means of obtaining the overall power of continuing the study: by integrating the conditional power function with respect to the current posterior distribution of the parameters, and by integrating the predictive distributions of future events over the critical region. If the samples are large and p_C and p_T are neither near 0 nor 1, it may be possible to work simply in terms of the difference in observed proportions, for which a normal sampling distribution with mean $p_C - p_T$ is assumed. This is equivalent to assuming normal posterior and predictive distributions and linear rejection boundaries: the predicted powers can then be obtained from the distribution function of a bivariate normal distribution. However, estimates of $p_C(1 - p_C)$ and $p_T(1 - p_T)$ are necessary for the variance of the observed difference in proportions; using the current data for estimates in the Halperin example, this approach yields a predictive power of 0.958, which is close to that obtained by the "exact" method that assumed prior independence.

A number of other studies have used predictions based on parameter distributions to investigate proposed trial designs: Herson and colleagues [6,7] consider single-sample binomial sampling in Phase II studies, McPherson [8] suggests using subjective prior opinion to predict the effect of multiple interim analyses, and we have advocated a similar approach [9]. However, each of the above has remained within the classical Neyman–Pearson framework of guaranteeing fixed type I and type II error relative to a prespecified hypothesis. We now feel this inappropriate and, essentially, the type II error should be averaged with respect to our belief about the unknown parameters; we believe this is a natural extension of viewing the alternative hypothesis as an *expected* improvement, as specifically stated by Halperin and colleagues [1] and by many other authors. We have also explored this approach in the context of making predictions before a trial even starts, by using the subjective opinions of clinicians to provide an initial distribution for the parameter [10].

APPENDIX 1: CONDITIONAL POWER FUNCTION

The test statistic Z, to be calculated at the end of the trial, may be written

$$Z = \frac{\hat{p}_C - \hat{p}_T - r/2}{[\hat{p}(1 - \hat{p})r]^{1/2}}$$

where $r = (m_C + n_C)^{-1} + (m_T + n_T)^{-1}$, $\hat{p}_C = (r_C + S_C)/(m_C + n_C)$ is the potential estimate of p_C, $\hat{p}_T$ is defined in parallel to $\hat{p}_C$, and $\hat{p} = (r_C + S_C + r_T + S_T)/(m_C + n_C + m_T + n_T)$ is the mortality rate that would be estimated under the null hypothesis.

Following previous authors [11,12] we approximate $\hat{p}$ by its conditional expectation $p' = (r_C + n_C p_C + r_T + n_T p_T)/n$, where n is the total sample size $m_C + n_C + m_T + n_T$. Then the criterion for rejecting H_0 is

$$\hat{p}_C - \hat{p}_T > k_\alpha[p'(1 - p')] + r/2.$$

Now define

$$E = E(\hat{p}_C - \hat{p}_T|p_C,p_T) = (r_C + n_C p_C)/(m_C + n_C) - (r_T + n_T p_T)/(m_T + n_T),$$
$$V = V(\hat{p}_C - \hat{p}_T|p_C,p_T) = n_C p_C(1 - p_C)/(m_C + n_C)^2 + n_T p_T(1 - p_T)/m_T + n_T)^2,$$

and then the conditional power is approximately

$$1 - \Phi\left\{\frac{k_\alpha[p'(1 - p')r] + r/2 - E}{V^{1/2}}\right\},$$

where $\Phi(\,)$ is the standard normal distribution function. The contours in Figure 1 are derived from this formula.

APPENDIX 2: POSTERIOR DISTRIBUTIONS

Consider the control group, in which r_C events are observed in m_C trials. We assume our pretrial opinion about p_C is expressed by a beta prior distribution with parameters (a,a). Then it is a standard result that the posterior distribution is beta with parameters $(a + r_C, a + m_C - r_C)$ [4].

Different criteria for "noninformative" priors set $a = 1$, 1/2, or 0 [13], although this choice has little effect on the results of the analysis. We adopt the convention that $a = 0$, and hence the posterior distribution plotted in Figure 2a is

$$p(p_C|r_C,m_C) = \frac{(m_C - 1)!}{(r_C - 1)!(m_C - r_C - 1)!} p_C^{r_C - 1}(1 - p_C)^{m_C - r_C - 1},$$

which has mean r_C/m_C—the maximum likelihood estimate of p_C. The marginal distribution of p_T is similarly derived. It would be quite reasonable to wish to express a pretrial correlation between p_C and p_T, and it would be possible for this to be introduced, at the cost of some additional computational complexity. It would, perhaps, be more natural to represent this correlation by reparameterizing in terms of a prior on a parameter of treatment difference, plus an independent prior on the control mortality. For the purposes of this simple example, however, we take the default position of assuming prior independence between p_C and p_T. Since this implies the posterior distributions

are independent, the contours in Figure 2c are obtained from the joint distribution $p(p_C|r_C,m_C)p(p_T|r_T,m_T)$.

APPENDIX 3: PREDICTIVE DISTRIBUTIONS

Given the posterior distribution derived in Appendix 2, the predictive distribution of the number of events S_C in the next n_C patients in the control group is given by

$$\begin{aligned} p(S_C|n_C,r_C,m_C) &= \int p(S_C|n_C,p_C)p(p_C|r_C,m_C)dp_C \\ &= \int \frac{n_C!}{(n_C - S_C)\,!S_C!} \cdot p_C{}^{S_C}(1 - p_C)^{n_C - S_C} \cdot p(p_C|r_C,m_C)dp_C \\ &= \frac{n_C!(m_C - 1)\,!(r_C + S_C - 1)\,!(n_C - S_C + m_C - r_C - 1)!}{(n_C - S_C)!\,S_C!\,(r_C - 1)\,!(m_C - r_C - 1)\,!(n_C + m_C - 1)!}, \end{aligned}$$

which is a beta-binomial distribution [14]. The contours in Figure 3 re obtained from the joint predictive distribution $p(S_C|n_C,r_C,m_C)p(S_T|n_T,r_T,m_T)$.

REFERENCES

1. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL: An aid to data monitoring in long-term clinical trials. Controlled Clin Trials 3:311–323, 1982
2. Andersen PK: Basing the decision to continue a clinical trial on conditional power calculations. Controlled Clin Trials in press.
3. Whitehead J: The Design and Analysis of Sequential Clinical Trials. Chichester; Ellis Horwood, 1983
4. Lindley DV: Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference. Cambridge: Cambridge University Press, 1970
5. Geisser S: Aspects of the predictive and estimative approaches in the determination of probabilities. Biometrics 38 (Suppl): 75–85, 1982
6. Herson J: Predictive probability early termination plans for phase II clinical trials. Biometrics 35:775–783, 1979
7. Atkinson EN, Brown BW, Herson J: KSTAGE: An interactive computer program for designing phase II clinical trials using predictive probability. Comp Biomed Res 15:220–227, 1982
8. McPherson K: On choosing the number of interim analyses in clinical trials. Stat Med 1:25–36, 1982
9. Freedman LS, Spiegelhalter DJ: The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. The Statist 33:153–160, 1983
10. Spiegelhalter DJ, Freedman LS: A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Stat Med (in press)
11. Casagrande JT, Pike MC, Smith PG: An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics 34:483–486, 1978
12. Fleiss JL, Tytun A, Ury HK: A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36:343–346, 1980
13. Box GEP, Tiao GC: Bayesian Inference in Statistical Analysis Reading MA: Addison-Wesley
14. Aitchison J, Dunsmore IR: Statistical Prediction Analysis. Cambridge: Cambridge University Press, 1975