BIOMETRICS 58, 823–831
December 2002

Seamlessly Expanding a Randomized Phase II Trial to Phase III

Lurdes Y. T. Inoue,[1,*] Peter F. Thall,[2] and Donald A. Berry[2]

[1]Department of Biostatistics, University of Washington,
Box 357232, Seattle, Washington 98195, U.S.A.
[2]Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center,
Box 447, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
**email:* linoue@u.washington.edu

SUMMARY. A sequential Bayesian phase II/III design is proposed for comparative clinical trials. The design is based on both survival time and discrete early events that may be related to survival and assumes a parametric mixture model. Phase II involves a small number of centers. Patients are randomized between treatments throughout, and sequential decisions are based on predictive probabilities of concluding superiority of the experimental treatment. Whether to stop early, continue, or shift into phase III is assessed repeatedly in phase II. Phase III begins when additional institutions are incorporated into the ongoing phase II trial. Simulation studies in the context of a non–small-cell lung cancer trial indicate that the proposed method maintains overall size and power while usually requiring substantially smaller sample size and shorter trial duration when compared with conventional group-sequential phase III designs.

KEY WORDS: Bayesian sequential design; Clinical trials; Markov chain Monte Carlo; Mixture models.

1. Introduction

The randomized comparative phase III clinical trial is the established scientific standard for determining whether an experimental treatment, E, is effective compared with a standard treatment, S. The usual method for deciding whether E is sufficiently promising to warrant phase III evaluation is to first conduct a phase II trial. In oncology trials of the sort considered here, this is typically a small, single-arm study of E, and the phase II data are compared with historical data on S. Many phase II trials are based on a k-nary variable, Y, recording early outcomes. For example, Y may be a binary indicator of >50% tumor shrinkage in oncology, an ordinal variable recording degree of lipid lowering in cardiology, or, more generally, a categorical variable recording possible combinations of desirable and adverse events. Except in trials of rapidly fatal diseases, few uncensored values of patient survival time, T, are available at the completion of phase II. Thus, the Y data are usually the basis for deciding whether to go to phase III. This widespread practice is motivated by the belief that it is not feasible in phase II to wait to observe the patients' survival times and the assumption that Y is a reasonable surrogate for T. Interim and final inferences comparing E with S in phase III typically are based only on the survival data from phase III while ignoring both the phase III data on Y and all of the phase II data.

Scientifically, this conventional approach suffers from several defects. The E-versus-S treatment effects in phase II are confounded with latent trial effects because the data on E and S arise from separate trials. Thus, the decision of whether to proceed with phase III is based on a confounded treatment-trial effect estimate. Once phase III has begun, however, the phase II data typically are discarded to avoid introducing bias into the confirmatory phase III comparison. Even when patients are randomized between E and S in phase II, the use of Y alone for treatment evaluation relies implicitly on its surrogacy for T, a generally tenuous assumption (Fleming and DeMets, 1996). Finally, given the assumption that T is related to Y, ignoring the Y data in phase III wastes information.

A typical complication is that Y is not observed immediately but rather is defined over a period of length t_0. If U is the right-censoring time of T and $T^* = \min\{T, U\}$, then Y is observed only if $T^* \geq t_0$. Thus, in addition to the indicator $d = I[T < U]$ that T is not censored, we will require $W = I(T^* \geq t_0)$, the indicator that Y is observed. The patient's outcome data thus consist of W and either (Y, T^*, d) if $W = 1$ or (T^*, d) if $W = 0$.

In this article, we propose a Bayesian phase II/III treatment evaluation strategy. Patients are randomized between E and S throughout. In phase II, the decision of whether to stop early, continue phase II, or proceed to phase III is made repeatedly during a time interval rather than at one time point. Phase III begins when additional institutions join the trial, with the phase II trial expanded into phase III without interruption. Decisions and inferences are based on predictive probabilities that E will be found to be superior given the observed (Y, T^*, d, W) data and possibly patient covariates. We do not assume that Y is a surrogate for T. Instead, we specify parametric models for $P(T \mid Y)$ and $P(Y)$ and assume that Y may affect T through the mixture model $P(T) = \Sigma_{y=1}^{k} P(T \mid Y = y)P(Y = y)$.

Our approach has several practical advantages. Randomizing from the start allows all current data, including the phase II data, to be utilized in each decision. Moreover, there is no interim suspension of accrual while waiting to evaluate Y for the most recent patients, which may be required using conventional phase II designs. If it is decided to proceed with phase III, then phase II is continued and data are accumulated while the phase III portion of the trial is being organized. In our application, these advantages result in substantial savings in time and resources compared with the conventional approach.

In Section 2, we describe the non–small-cell lung cancer (NSCLC) trial that motivated this research. The probability model is presented in Section 3. Six hypotheses under which we evaluate the design are described in Section 4. In Section 5, we apply the method to design the NSCLC trial, and we describe a simulation study in Section 6. Robustness is discussed in Section 7, and we conclude with a discussion in Section 8.

2. The Lung Cancer Trial

Our illustrative application is a trial of unresectable stage II or III non–small-cell lung cancer (NSCLC). Patients are randomized to chemotherapy + radiation either with (E) or without (S) an adjuvant adenovirus, Ad-p53, that carries a gene thought to restore programmed cell death, apoptosis, while also sensitizing cancer cells to the chemoradiation. The Ad-p53 is injected directly into the patient's tumor. Because patients with local control (LC) have better survival (Thomas et al., 1999), the rationale is that, through apoptosis and sensitization, Ad-p53 will increase the LC rate and thus prolong survival. Patient outcome consists of survival time and the binary indicator, Y, of whether a fine-needle aspiration biopsy of the patient's primary tumor at 5 months is negative, known as local control (LC), with $Y = 1$ if LC is achieved and $Y = 2$ if not. Thus, $t_0 = 5$ and $W = I(T^* \geq 5)$. Our probability model and trial design will provide a basis for evaluating the effect of Ad-p53 on LC, the effect of LC on survival, and a possible direct Ad-p53 effect on survival not mediated by LC.

The following conventional designs (CDs), which were considered initially, will provide a basis for evaluating the Bayesian design. Denote the fixed probability of LC with E under a frequentist model by p_E. For phase II, the null LC probability $p_E = .16$ was based on a previous study of chemoradiation in NSCLC (Le Chevalier et al., 1991). A Simon (1989) optimal two-stage phase II design with size .05 and power .90 to detect $p_E = .36$ stops and accepts H_0 after 21 patients have been evaluated if four or fewer have LC. Otherwise, 30 more patients are treated, with final acceptance or rejection of H_0 if the total number of LCs among the 51 patients is ≤ 12 or ≥ 13. That LC is evaluated 5 months after the start of therapy has the logistical implications that any two-stage phase II trial may require up to 10 months of waiting to observe the LC data needed to make the interim and final decisions.

Given phase II results sufficiently promising to warrant a phase III trial, initially phase III was planned using a conventional group-sequential design (Pocock, 1977; Lan and DeMets, 1983; Kim and DeMets, 1987; Jennison and Turnbull, 2000) with both inner and outer O'Brien–Fleming boundaries (1979) for a symmetric log-rank test comparing median survival times. At up to four successive times, the design rejects or accepts the null so that the trial may be stopped early due to either a significant treatment difference or futility. The overall size is .05 and, assuming a null median, $t_{.50}$, of 15.5 months, the power to detect a 25% increase in $t_{.50}$ with Ad-p53 is .80. Given a maximum of 900 patients and an accrual of 30 patients per month, the design requires up to 30 months of accrual plus 24 additional months of follow-up, with tests conducted at 175, 350, 524, and 699 deaths. For all designs considered here, the randomization is stratified by disease stage because stage III NSCLC reduces median survival by approximately 5% compared with stage II.

The NSCLC trial is a registration trial. Thus, the statistical design must be approved in advance by regulatory agencies, including the U.S. Food and Drug Administration (FDA). Because the Bayesian approach is not yet fully embraced by the FDA, the extent to which our design can be Bayesian is limited. We thus base all decisions on predictive probabilities rather than using a fully Bayesian decision theoretic approach. In virtually all practical applications, the overall Type I and Type II error rates both must be controlled. A major criticism of Bayesian methods is that the Type I rate is not controlled (Jennison and Turnbull, 2000, Chapter 18). We thus calibrate the Bayesian design's parameters to ensure an overall Type I error $\leq .05$, and we evaluate its frequentist properties.

3. Probability Model

3.1 *Mixture Model*

We assume that the probability density function (p.d.f.) f of T takes the general form

$$f(t) = f(t \mid T < t_0)\Pr(T < t_0) + \sum_{y=1}^{k} f(t \mid T \geq t_0, Y = y)\pi_y \Pr(T \geq t_0), \qquad t \geq 0, \tag{1}$$

where $\pi_y = \Pr(Y = y \mid T \geq t_0)$ varies with y but not t_0 because Y is observed only if $T \geq t_0$. The mixture in (1) may be generalized to allow continuous Y, although we do not consider this case here. We assume that $T = T_0(1-W) + (T_1 + t_0)W$, where T_0 and T_1 are latent survival times with T_0 following p.d.f. f_0 not depending on Y and T_1 following the mixture p.d.f. $f^{(\pi)} = \Sigma_{y=1}^{k} f_y \pi_y$. Expression (1) thus may be written

$$\begin{aligned} f(t) &= \{f_0(t)\}^{1-W}\{f^{(\pi)}(t-t_0)\mathcal{F}_0(t_0)\}^{W} \\ &= \begin{cases} f_0(t) & \text{if } t < t_0 \\ \mathcal{F}_0(t_0)\sum_{y=1}^{k} f_y(t-t_0)\pi_y & \text{if } t \geq t_0, \end{cases} \end{aligned} \tag{2}$$

where $\mathcal{F}_0(t) = \Pr(T_0 > t)$ is the survival function corresponding to f_0.

Introducing treatment, Z, we denote the survival and hazard functions of $[T_0 \mid Z]$ by $\mathcal{F}_{0,Z}(t)$ and $h_{0,Z}(t) = -\mathcal{F}'_{0,Z}(t)/\mathcal{F}_{0,Z}(t)$ and, for $y = 1, \ldots, k$, those of $[T_1 \mid Z, Y = y]$ by $\mathcal{F}_{y,Z}(t)$ and $h_{y,Z}(t)$. For a sample of n patients, we denote $Y = (Y_1, \ldots, Y_n)$, $Z = (Z_1, \ldots, Z_n)$, and so on for T, U, T^*, d, and W. Under (2), the likelihood is the product of two components, the first from patients who die or are censored before Y can be observed, $T^* < t_0$, and the second from patients for whom $T^* \geq t_0$ and hence Y is observed, i.e.,

$\mathcal{L}(\theta; T^*, d, Y, W, Z)$

$$= \prod_{i=1}^{n}\left[\left\{h_{0,Z_i}(T_i^*)\right\}^{d_i}\mathcal{F}_{0,Z_i}(T_i^*)]\right]^{1-W_i} \times \left[\prod_{y=1}^{k}\left[\left\{h_{y,Z_i}(T_i^*-t_0)\right\}^{d_i} \times \mathcal{F}_{y,Z_i}(T_i^*-t_0)\pi_{y,Z_i}\right]^{I[Y_i=y]} \mathcal{F}_{0,Z_i}(t_0)\right]^{W_i}. \tag{3}$$

We will use the following version of (3) in our application. Let $\lambda_{0,Z}$ be the death rate of $[T \mid T < t_0, Z]$ on $[0, t_0)$ and $\lambda_{y,Z}$ the death rate of $[T \mid T \geq t_0, Y = y, Z]$ on $[t_0, \infty)$. Let $exp(\lambda)$ denote the exponential distribution with mean $1/\lambda$ and $X = (X_1, \ldots, X_p)$ a vector of non-negative-valued covariates and denote $\beta^X = \Pi_j \beta_j^{X_j}$. We assume that $f_{0,Z}$ is given by

$$[T \mid T < t_0, Z, X, \theta] \sim \exp\left(\lambda_{0,Z}\beta^X\right) \tag{4}$$

and that, for each $y = 1, \ldots, k$, the component $f_{y,Z}$ of $f_Z^{(\pi)}$ is given by

$$[T \mid T \geq t_0, Y = y, Z, X, \theta] \sim \exp\left(\lambda_{y,Z}\beta^X\right). \tag{5}$$

Thus, $P(T \mid Y, Z)$ is piecewise exponential on $[0, t_0)$ and $[t_0, \infty)$, with parameters for the effects of treatments, covariates, whether Y is observed and, if so, the value of Y. Our basis for assuming an exponential model in this setting is empirical survival distributions reported by Le Chevalier et al. (1991) and Schaake-Koning et al. (1992). When Y is not observed, we characterize the death rates in terms of $\lambda_0 = \lambda_{0,S}$ and the multiplicative E-versus-S treatment effect $\eta_0 = \lambda_{0,E}/\lambda_0$. For $T \geq t_0$, we define $\eta_y = \lambda_{y,E}/\lambda_{y,S}$ for each $y = 1, \ldots, k$. Thus, the treatment effect on the death rate may vary as a function of the early outcome. Denoting $\lambda_1 = \lambda_{1,S}$, the comparative effects among different values of Y relative to the baseline outcome $Y = 1$ are $\gamma_y = \lambda_{y,S}/\lambda_1$. Thus, $\lambda_{y,S} = \lambda_1\gamma_y$, $\lambda_{y,E} = \lambda_1\gamma_y\eta_y$, and $\gamma_1 = 1$. For brevity, we will denote $\lambda = (\lambda_0, \lambda_1)$, $\gamma = (\gamma_2, \ldots, \gamma_k)$, and $\eta = (\eta_0, \eta_1, \ldots, \eta_k)$. Finally, $[Y \mid Z]$ is k-nary multinomial in $\pi_Z = (\pi_{1,Z}, \ldots, \pi_{k,Z})$, where $\pi_{y,Z} = \mathrm{P}(Y = y \mid Z)$.

3.2 *Prior Distributions*

We denote the gamma distribution with mean uv^{-1} and variance uv^{-2} by gam(u, v). *A priori,* we assume

$$\lambda_0, \lambda_1 \overset{\text{i.i.d.}}{\sim} \text{gam}(u_\lambda, v_\lambda) \tag{6}$$

$$\gamma_2, \ldots, \gamma_k \overset{\text{i.i.d.}}{\sim} \text{gam}(u_\gamma, v_\gamma) \tag{7}$$

$$\eta_0, \eta_1, \ldots, \eta_k \overset{\text{i.i.d.}}{\sim} \text{gam}(u_\eta, v_\eta) \tag{8}$$

$$\beta_1, \ldots, \beta_p \overset{\text{i.i.d.}}{\sim} \text{gam}(u_\beta, v_\beta). \tag{9}$$

Although Y may affect the death rate parameters, *a priori* there is no bias in either direction because λ_0 and λ_1 have the same prior, and similarly for the γ_j's and η_j's. Finally, we assume that the LC probabilities π_S and π_E are independent with common Dirichlet prior.

3.3 *Posterior Distributions*

We denote the parameter vector by $\theta = (\lambda_0, \lambda_1, \eta_0, \ldots, \eta_k, \gamma_2, \ldots, \gamma_k, \pi_S, \pi_E, \beta)'$ and the subvector of θ without λ_0 by $\theta_{-\lambda_0}$, with other subvectors denoted similarly. The posteriors of π_S and π_E are each Dirichlet with parameters depending on the data through the values of $\{(Y_i, W_i, Z_i), i = 1, \ldots, n\}$. Because the posteriors of parameters in $\theta_{-(\pi_E,\pi_S)}$ are not available in closed form, we use Gibbs sampling (Gelfand and Smith, 1990) to draw samples of $\theta_{-\pi_S,\pi_E}$ from the posterior $p(\theta_{-\pi_S,\pi_E} \mid \text{data})$. This is facilitated by the fact that the full-conditional distributions of all parameters other than π_E and π_S are gammas. Samples from the posterior of $p(\theta_{-(\pi_S,\pi_E)} \mid \text{data})$ are obtained by iteratively sampling θ from the full conditionals of the other parameters after initial burn-in iterations.

The decision rules to be presented in Section 5 will be based on $\Pr\{\Delta > 0\}$, where Δ is the difference between the mean survival times with E and S. We estimate this probability, using Monte Carlo integration, as the proportion of posterior samples for which $\Delta > 0$. We also simulate predictive survival times at future times, given the current data, by obtaining posterior samples of θ at time t and then, conditional on these samples, simulating survival times under the piecewise exponential model with censoring at the future time.

4. Application to the NSCLC Trial

In the NSCLC trial $k = 2$, the baseline death rates are $\lambda = (\lambda_0, \lambda_1)$, the early outcome probability vectors are simply the single values $\pi_{1,S} = \pi_S$ and $\pi_{1,E} = \pi_E$ for LC (Y=1), and the effect of LC on survival is $\gamma \equiv \gamma_2$. The direct treatment effects on survival not mediated by LC are η_0 if W =0, η_1 if W =1 and there is no LC (Y=2), and η_2 if W =1 and there is LC, so $\eta = (\eta_0, \eta_1, \eta_2)$. There is one covariate, the indicator X that the patient has stage III disease, which has multiplicative effect β on the death rate. Thus, for $t \geq t_0 = 5$,

$$\Pr[T > t \mid X, Z] = e^{-5\lambda_0\eta_0^{I[Z=E]}} \times \left\{\pi_Z e^{-(t-5)\lambda_1\gamma\eta_2^{I[Z=E]}\beta^X} + (1-\pi_Z)e^{-(t-5)\lambda_1\eta_1^{I[Z=E]}\beta^X}\right\}. \tag{10}$$

Figure 1 illustrates the manner in which π, γ, and λ determine $\mathrm{P}[T > t]$ under this version of the mixture model, assuming for simplicity that $\eta_0 = \eta_1 = \eta_2 = \beta = 1$.

The possible paths whereby Ad-p53 may affect survival are illustrated in Figure 2. Ad-p53 (1) improves or has no effect on the LC rate when $\pi_E - \pi_S > 0$ or $= 0$, (2) LC improves or has no effect on survival when $\gamma < 1$ or $= 1$, and (3) Ad-p53 has or does not have a direct effect improving survival, not mediated by LC, when $\eta_j < 1$ or $= 1$ for j = 0, 1, or 2. Thus, (π, γ) characterize the treatment effect for the pathway that is mediated through LC, while η is the direct effect. The global null hypothesis is H_0: $\pi_E = \pi_S, \gamma = 1$, and all $\eta_j = 1$. The first alternative hypothesis, which formalizes the investigators' motivation, is H_1: $\pi_E > \pi_S, \gamma < 1$, and all $\eta_j = 1$. If neither of the two effects under H_1 is present, then Ad-p53 does not improve survival. These cases are formalized by the hypotheses H_0^*: $\pi_E > \pi_S, \gamma = 1$, all $\eta_j = 1$, and H_0^{**}: $\pi_E = \pi_S, \gamma < 1$, and all $\eta_j = 1$, which may be considered two additional null hypotheses. The most optimistic hypothesis is H_1^*: $\pi_E > \pi_S, \gamma < 1$, all $\eta_j < 1$, under which Ad-p53 improves survival via both the direct

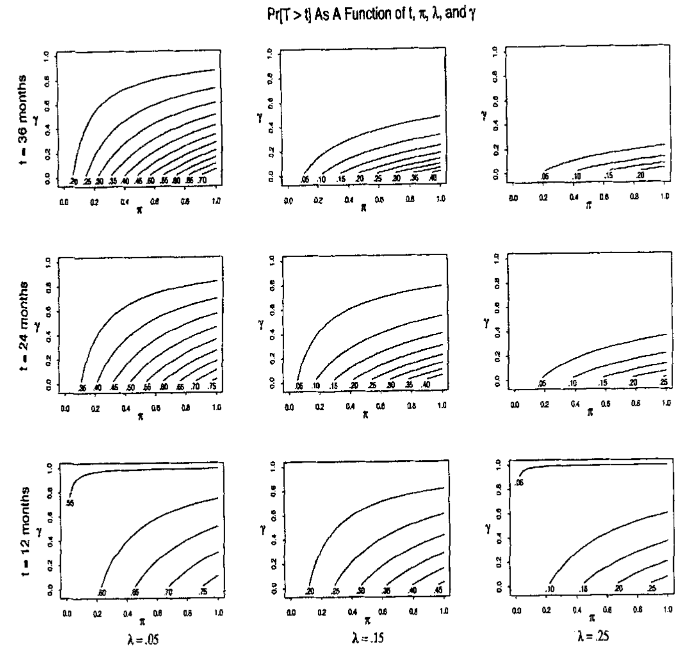

Figure 1. Contour plots of $P[T > t]$ as a function of (π, γ) for t = 12, 24, 36 and λ = .05, .15, .25.

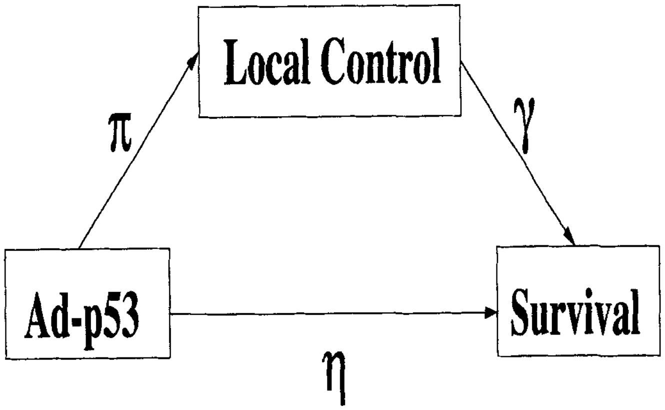

Figure 2. Possible pathways for effects of Ad-p53 on survival.

effects of the η_j's and by improving the probability of LC, which in turn improves survival. Finally, H_1^{**}: $\pi_E = \pi_S, \gamma = 1$, $\eta_0 = \eta_1 = \eta_2 = \eta < 1$ yields the simple exponential model in which Y is neither affected by treatment nor has any effect on survival but the direct treatment effect η is present.

For the NSCLC trial, we assume *a priori* that $\lambda_0, \lambda_1, \eta_0, \eta_1, \eta_2, \gamma, \beta \overset{\text{i.i.d.}}{\sim} \exp(1)$, and $\pi_S, \pi_E \overset{\text{i.i.d.}}{\sim} \text{beta}(1, 1)$. Because the anticipated phase II accrual rate is 20 patients per month; the expected sample size at each monitoring look is large. Consequently, these priors are nearly noninformative in the sense that, even at the earliest monitoring time at month 8, the prior distribution plays essentially no role in the decision.

5. A Bayesian Phase II/III Design

We present the Bayesian design in the context of the NSCLC trial. To facilitate comparison with the CD described earlier, we calibrate the Bayesian design parameters to obtain a false positive rate $\leq$ .05 under H_0 and power $\geq$ .80 under H_1. The two designs are not strictly comparable, however, because they have different maximum durations, utilize the available data differently, and accrue patients at different rates. The last difference is because, with the Bayesian strategy, the time t^* when it is decided whether to organize phase III or stop the trial early depends on the data and hence is random. If it is decided to organize phase III at t^*, then accrual increases thereafter as new institutions join the expanding trial.

Our Bayesian design requires specifying a maximum number of patients, N, and a maximum duration, D. All decisions are based on predictive probabilities involving future data available either 1 year from the present or at the maximum study duration. Specifically, at any given time t during the trial, we let $\mathcal{X}_1(t)$ denote the data available at future time $t + 12$ if accrual is terminated at t and patients are followed for 12 more months and $\mathcal{X}_2(t)$ denote the data available at D if all N patients are accrued and follow-up continues to D. Decisions are based on probabilities

$$\phi_j(t) = \Pr[\Delta > 0 \mid \chi_j(t)], \qquad j = 1, 2. \tag{11}$$

We denote the conclusion that E is superior to S with regard to survival by $E \succ S$ and its complement by $E \preceq S$. A large value of the criterion probability $\phi_1(t)$ provides evidence that, if no additional patients are accrued after t and the patients are followed for 12 more months, then it would be likely that $E \succ S$. A large value of $\phi_2(t)$ says that, if the maximum allowed future resources were expended, then it would be likely that $E \succ S$.

In the NSCLC trial, N = 900, as with the conventional phase III design, *but with* D = 72 *months rather than* 54. The decision criteria used throughout the trial are based on predictive probabilities of concluding superiority of the experimental treatment through

$$p_1(t) = \min\{\Pr[\phi_1(t) > .98], \Pr[\phi_2(t) > .98]\} \tag{12}$$

and

$$p_2(t) = \max\{\Pr[\phi_1(t) > .98], \Pr[\phi_2(t) > .98]\}. \tag{13}$$

The NSCLC trial is conducted as follows.

Stopping rules. At any time t = 8, 10, 12 and 16, 20, ..., 72 if

(i) $p_1(t) \geq .98$, then stop and conclude $E \succ S$;
(ii) $p_2(t) \leq .01$ or $p_1(72) < .98$, then stop and conclude $E \prec S$;
(iii) $p_1(t) < .98$ and $p_2(t) > .01$, then continue.

If neither (i) nor (ii) above is the case for $8 \leq t \leq 12$ during phase II, then the following criteria are used to decide whether to expand the trial from phase II to phase III.

Phase II to phase III. At t = 8, 10, and 12 months, if

(i) $.01 < \Pr[\phi_2(t) > .98] < .80$, then continue phase II;
(ii) $\Pr[\phi_2(t) > .98] \geq .80$, then organize phase III;
(iii) $\Pr[\phi_2(12) > .98] < .80$, then stop and conclude $E \prec S$.

If it is decided at time t^* = 8, 10, or 12 months to proceed with phase III, then it will take some time, t_{ORG}, to organize the phase III trial. The "phase II" portion of the trial will continue to accrue patients during the period from t^* to the time $t^* + t_{\mathrm{ORG}}$ when phase III begins. This provides an important logistical advantage over the conventional approach, under which no patients are accrued during the period of length t_{ORG} between the two phases.

Our decision rules are based on several numerical probability cut-offs that may appear somewhat arbitrary. We obtained these numerical values via trial and error, based on a series of preliminary simulations, to obtain a design with good frequentist operating characteristics (OCs). Similarly, the qualitative forms of $p_1(t)$ and $p_2(t)$ and the decision rules were constructed to obtain a practical design with desirable frequentist properties. In this sense, the decision rules are *ad hoc*. We anticipate that, in future applications of this methodology to other trials, different numerical probability cut-offs and qualitatively different rules may be used.

6. Simulation Study

6.1 *Simulation Parameters*

To evaluate and compare the Bayesian and CDs, we simulated the trial 10,000 times under each of the six hypotheses described in Section 4.1 using each design. We assumed $\pi_S = .16$, a null median survival, med(T), of 15.5 months, and set $\lambda_0 = \lambda_1 = .05$. The value $\beta = 1.05$ corresponds to stage III patients having .95 times the med(T) of stage II patients. The remaining parameters had null values $\pi_E = .16$, $\gamma = 1$, and all $\eta_j = 1$. The alternative values $\gamma = .40$ and $\pi_E = .53$ were set to correspond with an improvement of 25% over the null med(T), which was set by the physicians planning the trial. The values $\eta_0 = \eta_1 = \eta_2 = .80$ were set to increase the mean survival by 25% under the simple exponential model of H_1^{**}, where $\pi_E = \pi_S = .16$ and $\gamma = 1$.

To facilitate comparison, we used the same maximum number of patients (900) and maximum trial duration (72 months) for all simulated trials. Two CDs were considered, the first having up to four tests at equally spaced information time intervals, the second having up to 18 tests at the same times used by the Bayesian design. For the Bayesian design, in the case where it is decided at time t^* during phase II to proceed with phase III, we assumed that it would take $t_{\mathrm{ORG}} = 9$ months to organize phase III, with phase III starting and accrual increasing 20–30 patients per month at time $t^* + 9$.

6.2 *Computing*

We used Markov chain Monte Carlo (MCMC) methods to compute the predictive probabilities. For each of the 10,000 simulated trials under the Bayesian design, after an initial burn-in of 100 iterations, every 10th sample obtained in 1000 iterations of the Gibbs sampling procedure was retained for computing $\Pr(\Delta > 0 \mid \chi_j(t))$. To estimate each $\Pr(\phi_j(t) > p)$, we obtained 50 data sets of the form $\mathcal{X}_j(t)$ by first sampling 50 values of θ from the posterior $\mathrm{P}(\theta \mid data_t)$ and generating one such future data set from $\mathrm{P}(T, Y \mid \theta)$ for each θ.

The numerical design parameters were determined in an initial simulation study to obtain a design with size $\leq .05$ under H_0 and power $\geq .80$ under H_1. We used the following empirical rule. For each of the 50 pairs of simulated future data sets $\mathcal{X}_1(t)$ and $\mathcal{X}_2(t)$ noted above, $\phi_1(t)$ and $\phi_2(t)$ and the binary indicators $I_1 = I[\phi_1(t) > .98]$ and $I_2 = I[\phi_2(t) > .98]$ were computed. If $\Sigma_{i=1}^{50} I_{1,i} = 50$ or $\Sigma_{i=1}^{50} I_{2,i} = 50$, then stopping rule (i) was applied. If $\Sigma_{i=1}^{50} I_{1,i} = 0$ or $\Sigma_{i=1}^{50} I_{2,i} = 0$, then stopping rule (ii) was applied, and so on. Alternative designs may be obtained by modifying the design parameters to obtain other size and power figures.

While the nominal MCMC simulation sample size of 100 might be considered small, preliminary simulations assessing the convergence of the Gibbs sampler indicated that the design parameters we chose were adequate. There is a tradeoff between computing time and the use of a larger simulation sample size to improve the precision of the posterior estimates. While the simulation sample size during the design stage is limited by computing time, in the actual trial, we will base all inferences on a much larger posterior sample size to gain precision and ensure convergence of the Gibbs sampler.

6.3 *Results*

Table 1 summarizes the simulation results. The designs all have essentially the same Type I error under all three null hypotheses, H_0, H_0^*, and H_0^{**}. The most striking result is that the Bayesian design has much shorter mean duration under H_0, H_0^*, H_0^{**}, and H_1^* and much smaller mean sample size under all hypotheses while maintaining good overall significance level and power. Figure 3 gives box plots of the trial duration and sample size under H_0 and H_1 for each of the three designs. The plots show that the achieved sample size distributions of the Bayesian design are both more variable and likely to be much smaller than those of the CDs. The advantage of the Bayesian design in terms of trial duration is also substantial. Under the CD with ≤ 4 tests, there is virtually no difference in mean trial duration under H_0 and H_1. In contrast, under the Bayesian design, the trial duration is much more variable and is, on average, 10 months shorter under H_0 compared with H_1, with the longer mean duration under H_1 about the same as that of the CD. The only case where the Bayesian design has substantially smaller power is the simple exponential model under the alternative H_1^{**}. Much of the Bayesian design's advantage over the CD is because the model underlying the Bayesian design accounts for LC. Because the biological motivation for the NSCLC trial is based on H_1, the investigators consider H_1^{**} a medically untenable hypothesis.

We compared the Bayesian design with the CD with up to 18 decisions to determine whether the Bayesian design's advantage may be due to the fact that it makes decisions more frequently. Increasing the maximum number of tests from 4 to 18 under the CD has the effect of greatly increasing its duration under all hypotheses but H_1^*. This is due to the fact that, if additional interim tests are conducted, the earlier boundaries of the conventional group sequential test must be more extreme in order to maintain the same overall Type I error.

Under H_0, accounting for accrual and the time required to evaluate Y, the Simon (1989) two-stage phase II design described earlier has a mean duration of 7 months, much less than the 20.4 months under the Bayesian phase II/III design. The situation is completely reversed under H_0^*, where $\pi_E > \pi_S$ but E has no survival advantage over S. In this case, the Simon design has probability .90 of correctly concluding that $\pi_E > \pi_S$ and thus incorrectly leading to a phase III trial because this design ignores survival time. If $\pi_E > \pi_S$, the conventional approach requires about 12.5 months to complete a single-arm, two-stage phase II trial plus an additional 9 months to organize phase III, during which no patients are accrued. Thus, under H_0^*, on average, the total

Table 1

Operating characteristics of the Bayesian and conventional designs under the six mixture model-based hypotheses. The conventional designs allow up to either 4 or 18 tests.

Hypothesis	Design	Duration	No. of patients	No. of LCs	No. of deaths	Conclude $E \succ S$
H_0	Bayesian	20.4	425	53	180	.03
	Conventional (4)	28.1	842	105	459	.05
	Conventional (18)	40.2	884	109	583	.05
H_1	Bayesian	30.7	640	189	306	.85
	Conventional (4)	29.5	884	262	512	.83
	Conventional (18)	40.7	888	263	534	.91
H_0^*	Bayesian	21.6	453	134	197	.04
	Conventional (4)	28.1	842	249	460	.05
	Conventional (18)	40.2	884	262	584	.04
H_0^{**}	Bayesian	21.7	452	56	193	.03
	Conventional (4)	28.5	854	106	461	.05
	Conventional (18)	40.4	884	110	558	.05
H_1^*	Bayesian	23.2	525	162	181	.97
	Conventional (4)	28.7	861	267	365	>.99
	Conventional (18)	27.9	799	246	322	>.99
H_1^{**}	Bayesian	29.2	576	74	296	.56
	Conventional (4)	29.1	873	111	512	.79
	Conventional (18)	42.2	879	112	573	.86

time required under the conventional approach to correctly conclude that $E \preceq S$ is 37.5 months, compared with 21.6 months under the Bayesian design.

Under H_1, H_1^*, or H_1^{**}, Table 1 and the box plots in Figure 3 are conservative with respect to the trial duration comparison for the Bayesian design and CDs. If we account for accrual and follow-up time in phase II and the 9 months required to organize phase III using the conventional approach, a somewhat different comparison emerges. The Simon two-stage design described earlier requires 6 months if it stops after 21 patients and 12.5 months if it stops after 51 patients. If it is decided to proceed with phase III, the entire phase II/III process takes 21.5 months plus the duration of phase III. This yields total mean durations of 9.1, 45.4, and 46.7 months under H_0, H_0^*, and H_1, respectively, under the conventional approach, compared with 20.4, 21.6, and 30.7 months using the Bayesian design. Thus, in terms of actual time invested, the Bayesian design risks spending an additional 11.3 months in the null case but saves on average 23.8 months under H_0^* and 16 months under H_1. Some of the advantage of the Bayesian design under H_0^* and H_1 is attributable to the mixture model and some to the fact that it does not delay accrual between phase II and phase III. To make a somewhat more fair comparison, one may assume that there is no delay between phase II and phase III with the conventional design. Under this assumption, simply subtracting 9 months from the trial durations with the conventional design yields mean durations of 36.4 months under H_0^* and 37.7 under H_1, still much larger than the corresponding values 23.8 and 16 with the Bayesian design.

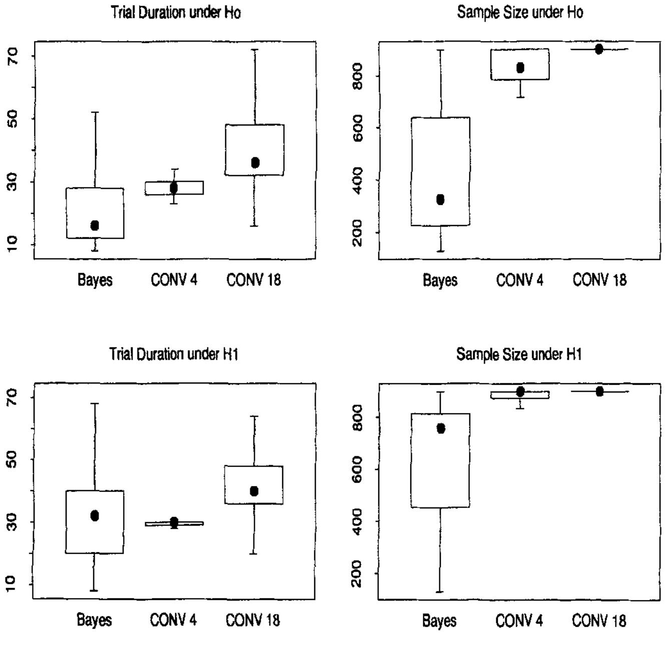

Figure 3. Sample size and trial duration distributions under H_0 and H_1 for the Bayesian and conventional designs. Each box runs from p_{25} to p_{75}, the dot denotes p_{50}, and the whiskers extend to the nearest value not beyond $1.5 \times (p_{75} - p_{25})$ from the box, where p_q is the qth percentile.

If this advantage is disregarded, however, and the Bayesian design is compared with the CD with $D = 54$, a natural question is whether proceeding beyond 54 months is worthwhile. Under H_1, only 3.5% of the trials simulated under the Bayesian design exceeded 54 months. Among these, the Bayesian design has power .76 of correctly concluding $E \succ S$. In comparison, among the simulated conventional trials that

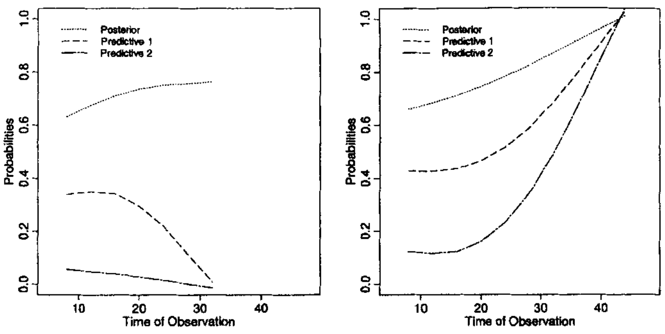

Figure 4. Posterior and predictive probabilities for two typical simulated trials. **Left panel.** Based on a trial simulated under H_0 that concludes $E \preceq S$ at 32 months. **Right panel.** Based on a trial simulated under H_1 that concludes $E \succ S$ at 44 months. In both panels, "Posterior" refers to $\Pr(\Delta > 0 \mid \text{data}_t)$, "Predictive 1" is $\Pr(\phi_1(t, t+12) > .98)$, and "Predictive 2" is $\Pr(\phi_2(t, 72) > .98)$.

went beyond 54 months, a log-rank test based on the 54-month data would have rejected the null with mean probability .41, whereas a log-rank test based on the data at the end of the study concluded $E > S$ with probability .63. This illustrates an important advantage of the Bayesian approach namely, that it is likely to continue in cases where the results at $t = 54$ favor E but are not statistically significant. In contrast, using conventional methods, this situation is typically dealt with at the regulatory level by requiring that a second phase III trial be conducted.

Figure 4 illustrates how the decision criteria operate during the trial. The left panel is based on a trial simulated under H_0. Although the posterior probabilities $\Pr[\Delta > 0 \mid data_t]$ indicate some improvement with the experimental therapy E over the conventional S, the predictive probabilities indicate that $E \succ S$ is not likely. In contrast, in the right panel, a trial simulated under H_1 indicates $E \succ S$ with both posterior and predictive probabilities increasing in t. The figure illustrates that $\phi_1(t, s)$ and $\phi_2(t, s)$ are more sensitive to t than $\Pr[\Delta > 0 \mid data_t]$ and thus together provide a more flexible basis for decision making.

The Bayesian mixture model uses the LC data to update the distributions of π_S and π_E while also accounting for the effects of (π_S, π_E) on survival. To assess the extent to which this contributes to the method's performance, we reran the simulations with the LC probabilities fixed rather than random under the Bayesian model. Assuming that $\Pr(\pi_S = \pi_E = .16) = 1$, the OCs are essentially the same under H_0 but mean (duration, number of patients, power) change from (30.7, 640, .85) in Table 1 to (11.5, 234, .01) under H_1. Thus, stopping the model from learning about π_S and π_E may destroy the method's power. The opposite action of successively increasing one of the prior variances of $\{\lambda_0, \lambda_1\}$, $\{\eta_0, \eta_1, \eta_2\}$, γ, or β from 1 to 100 only has a substantive effect on the design's OCs for $\text{var}(\eta_j) = 100$, where the mean (duration, number of patients, power) become (14.1, 295, .42) under H_0 and (14.6, 297, .83) under H_1. This indicates that this prior variance should be carefully controlled.

7. Robustness

To examine the design's behavior under nonexponentially distributed survival times, we consider the Weibull with hazard function $h(t) = c\lambda^c t^{c-1}$. Because $\text{med}(T)$ varies with c if the other model parameters are fixed, to evaluate sensitivity to c, we first fixed $\gamma = .40$, $\eta = \beta = 1$, and $\pi_S = \pi_E = .16$ and solved for (c, λ) pairs with $.75 \leq c \leq 1.25$ yielding the null $\text{med}(T) = 15.5$. For H_1, we used the same (c, λ) pairs and solved for (β, π_E) pairs yielding $\text{med}(T) = 19.375$. The results are summarized in Table 2. While both designs show negligible changes under H_0, their properties change substantively with $(c, \lambda, \beta, \pi_E)$ under H_1, with the only exception that the conventional design's sample size is quite stable under all cases. While the power of both designs is sensitive to $(c, \lambda, \beta, \pi_E)$ under H_1, the Bayesian design shows a smaller loss of power in the most extreme case considered.

8. Discussion

We have shown, via simulation in the context of a particular trial, that the use of mixture model-based predictive probabilities as decision criteria may provide substantial savings in time and sample size compared with conventional group-sequential designs having similar overall significance level and

Table 2

Operating characteristics of the Bayesian design (B) and conventional design with up to 18 *tests (C18) under Weibull survival time distributions* $\Pr(T > t) = \exp\{-(\lambda t)^c\}$. *We fixed* $\eta = 1$ *and* $\pi_S = .16$.

		(c, λ)	(.75, .034)		(1.00, .048)		(1.25, .059)	
H_0	$\gamma = 1$	(β, π_E)	(1.0, .16)		(1.0, .16)		(1.0, .16)	
			B	C18	B	C18	B	C18
		Duration	19.7	40.7	19.5	40.5	19.3	40.3
		Number of patients	410	884	415	884	414	884
		Pr(conclude $E \succ S$)	.03	.05	.03	.05	.03	.04
H_1	$\gamma = .4$	(β, π_E)	(1.047, .525)		(1.050, .531)		(1.053, .536)	
			B	C18	B	C18	B	C18
		Duration	33.7	46.5	29.9	42.9	27.0	37.9
		Number of patients	623	888	616	888	593	886
		Pr(conclude $E \succ S$)	.62	.52	.86	.82	.94	.94

power. Our use of the mixture model is motivated by the desire to use more of the available information, specifically both Y and T. We do not assume that Y is a surrogate for T but only that $\mathrm{P}(T \mid Y = y)$ varies with y. Randomizing from the start allows the conventional sharp division between phase II and phase III to be replaced by a process of repeatedly deciding whether to stop the trial or expand it by adding new institutions without wasting the phase II data. Additional flexibility is obtained by allowing the confirmatory decision of whether $E \succ S$ to be made in phase II.

The idea of combining phase II and phase III within the same trial is not new. Thall, Simon, and Ellenberg (1988) propose a two-stage design for trials with binary outcomes in which a selection stage is followed by a second-stage comparison of the selected experimental treatment to a standard. Schaid, Wieand, and Therneau (1990) provide a similar design for time-to-event outcomes. A common goal of these designs is to control the overall Type I and Type II error rates of the entire procedure. We also do this, but we use (Y, T) as the outcome rather than only one of the two and we allow decisions to be made much more frequently.

There is an extensive literature on the use of auxiliary variables to improve inferences. Lagakos (1976, 1977) utilizes the time to a nonfatal event to improve survival time estimation. Our formulation is similar in that Y plays the role of Lagakos' time-to-event variable, although here the observation of Y at t_0 and its discreteness lead to a rather different model formulation. In a particular case, Cox (1983) quantifies the amount of information lost due to censoring that is recovered by an auxiliary variable. Pepe (1992) factors the likelihood into the components $P(T \mid Z)P(Y \mid T, Z)$ on the set where both T and Y are observed and $P(Y \mid Z)$ on the set where Y but not T is observed. Fleming et al. (1994) give a general method for incorporating auxiliary variables by augmenting the estimating equations of the Cox regression model (1972). Finkelstein and Schoenfeld (1994) consider time-dependent auxiliary variables on disease progression for improving nonparametric survival estimates. Hogan and Laird (1997) model the joint distribution of (Y, T) by considering a mixture in which the components are $P(Y \mid T)$ and $P(T)$, with Y denoting repeated measurements possibly subject to missing data. Nam and Zelen (2001) derive statistical tests to verify whether a clinical intermediate endpoint induces a change in the survival distribution. In the same spirit, we take a Bayesian approach, model $P(Y \mid Z)$ and $P(T \mid Y, Z, W)$, and factor the likelihood in terms of whether Y is observed. As shown by the simulations with nonrandom π_S and π_E, much of our method's advantage over conventional designs may be attributed to the Bayesian mixture model.

Bayesian methods for clinical trials have been proposed for many years, dating at least to Anscombe (1963). Differences between the Bayesian and frequentist approaches and impediments to using the former are discussed in Berry (1993). In the present article, we use the Bayesian approach as a tool for deriving a design having good frequentist properties. In a similar application, Berry et al. (2001) consider whether to shift from dose-finding in phase II to a confirmatory phase III trial on the basis of a decision analysis, with the goal to maximize expected profit. The same idea could be used in the present context by considering the costs associated with extending accrual to additional centers.

A fully Bayesian approach would make explicit use of decision theory (Berry, Wolff, and Sack, 1994; Berry and Stangl, 1996). This would replace the problem of obtaining the decision rules and their numerical cutoffs, as we have done, with that of formulating a loss function and specifying its numerical parameter values. As shown by Stallard, Thall, and Whitehead (1999), even in the context of a single-arm phase II trial, this is a nontrivial process. We have not taken a decision theoretic approach here for practical reasons. The NSCLC trial is a registration trial, and current negotiations indicate that the U.S. regulatory agencies involved look favorably on the proposed design, despite its novelty. Our goal is to bring about the actual use of this new statistical methodology to conduct the NSCLC trial. Decision theory, while scientifically ideal, would introduce a level of innovation that would make the design unlikely to be approved for actual use. We feel that the design described here, once approved at the regulatory level and actually used in clinical trial conduct, will provide a basis for the future use of Bayesian methods in phase III trials, including those based on decision theory.

Résumé

Nous proposons pour l'extension en continu à des essais cliniques comparatifs un dispositif séquentiel bayésien adapté aux essais de phase II/III. Le dispositif suppose un mélange de modèles paramétriques, il est basé à la fois sur des durées de survie et la survenue d'événements discrets précoces affectant éventuellement la survie. Les patients sont randomisés entre les groupes traitements dans le petit nombre de centre impliqués dans l'essai de phase II. Les probabilités prédictives de conclure à la supériorité du traitement à l'essai servent de base aux décisions, soit d'arrêter précocement l'essai, soit de continuer la phase II, soit de passer à la phase III en incorporant des nouveaux centres à l'étude. Des études de simulation dans le contexte de cancers du poumon non à petites cellules indiquent que la méthode proposée requiert des échantillons substantiellement plus petits et des essais de durée moindre, tout en conservant seuil et puissance globales des tests conventionnels des dispositifs d'essais séquentiels de phase III.

References

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association* **58,** 365–383.

Berry, D. A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine* **12,** 1377–1404.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health-related research. In *Bayesian Biostatistics,* D. Berry and D. Stangl (eds), 1–66. New York: Marcel Dekker.

Berry, D. A., Wolff, M. C., and Sack, D. (1994). Decision making during a phase III randomized controlled trial. *Controlled Clinical Trials* **15,** 360–379.

Berry, D. A., Mueller, P., Grieve, A. P., Smith, M., Parke, T., Blazek, R., Mitchard, N., and Krams, M. (2001). Adaptive Bayesian designs for dose-ranging drug trials. In *Case Studies in Bayesian Statistics,* Volume 5, C. Gastonis, B. Carlin, and A. Carriquiry (eds), 99–181. New York: Springer-Verlag.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. (1983). A remark on censoring and surrogate response variables. *Journal of the Royal Statistical Society, Series B* **45,** 391–393.

Finkelstein, D. M. and Schoenfeld, D. A. (1994). Analyzing survival in the presence of an auxiliary variable. *Statistics in Medicine* **13,** 1747–1754.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine* **125,** 605–613.

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13,** 955–968.

Gelfand, A. E. and Smith, A. M. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 383–409.

Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event time. *Statistics in Medicine* **16,** 239–257.

Jennison, C. J. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman and Hall.

Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the Type I error spending rate function. *Biometrika* **74,** 149–154.

Lagakos, S. W. (1976). A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics* **32,** 551-559.

Lagakos, S. W. (1977). Using auxiliary variables for improved estimates of survival time. *Biometrics* **33,** 399–404.

Lan, G. and DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Le Chevalier, T., Arriagada, R., Quoix, E., et al. (1991). Radiotherapy alone versus combined chemotherapy and radiotherapy in nonresectable non–small-cell lung cancer: First analysis of a randomized trial in 353 patients. *Journal of the National Cancer Institute* **83,** 417–423.

Nam, C. M. and Zelen, M. (2001). Comparing the survival of two groups with an intermediate clinical event. *Lifetime Data Analysis* **7,** 5–19.

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

Pepe, M. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79,** 355–365.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.

Schaake-Koning, C., van der Bogaert, W., Dalesio, O., et al. (1992). Effects of concomitant cisplatin and radiotherapy on inoperable non-small cell lung cancer. *New England Journal of Medicine* **326,** 524-530.

Schaid, D. J., Wieand, S., and Therneau, T. M. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* **77,** 507–513.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10,** 1–10.

Stallard, N., Thall, P. F., and Whitehead, J. (1999). Decision-theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **55,** 971–977.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75,** 303–310.

Thomas, M., Rube, C., Semik, M., von Eiff, M., Freitag, L., Macha, H. N., Wagner, W., Klinke, E., Scheld, H. H., Willich, N., Berdel, W. E., and Junker, K. (1999). Impact of preoperative bimodality induction including twice-daily radiation on tumor regression and survival in stage III non-small cell lung cancer. *Journal of Clinical Oncology* **17,** 1185–1193.

Received May 2001. *Revised May* 2002.
Accepted July 2002.