Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials

Peter F. THALL, Hsi-Guang SUNG, and Elihu H. ESTEY

Therapy of rapidly fatal diseases often requires multiple courses of treatment. In each course, the treatment may achieve the desired clinical goal ("response"), the patient may survive without response ("failure"), or the patient may die. When treatment fails in a given course, it is common medical practice to switch to a different treatment for the next course. Most statistical approaches to such settings simply ignore the multicourse structure. They characterize patient outcome as a single binary variable, combine death and failure, and identify only one treatment for each patient. Such approaches waste important information. We provide a statistical framework, including a family of generalized logistic regression models and an approximate Bayesian method, that incorporates historical data while accommodating multiple treatment courses, a trinary outcome in each course, and patient prognostic covariates. The framework serves as a basis for data analysis, treatment evaluation, and clinical trial design. In contrast with the usual approach of evaluating individual treatments, our methodology evaluates outcome-adaptive, multicourse treatment strategies that specify, within prognostic subgroups, which treatment to give in each course. We describe a general approach for constructing clinical trial designs that may be tailored to different multicourse settings. For each prognostic subgroup, based on a real-valued function of the covariate-adjusted probabilities of response and death, the design drops inferior treatment strategies during the trial and selects the best strategy at the end. The methodology is illustrated in the context of designing a randomized two-course, three-treatment acute leukemia trial with two prognostic covariates. To validate the model and develop a prior, we first fit the model to a historical dataset. We describe a simulation study of the design under several clinical scenarios. The simulations show that the method can reliably identify treatment–subgroup interactions based on moderate sample sizes.

KEY WORDS: Bayes information criterion; Generalized logistic model; Leukemia; Markov chain Monte Carlo; Simulation.

1. INTRODUCTION

Therapy of rapidly fatal diseases often requires multiple courses of treatment. The clinical goal is to achieve a "response," such as complete remission (CR) of leukemia, 50% shrinkage of a solid tumor, or resolution of infection. Such responses are presumed to predict longer survival. The other therapeutic outcomes are death during treatment and "failure," in which the patient survives therapy but does not respond. Death during therapy is an unavoidable risk in oncology trials involving acute or advanced disease where only very aggressive, life-threatening treatments have any substantive antidisease effect. Thus, in general, each treatment course results in one of three possible outcomes: response, death, or failure. When treatment fails after a given course, it is common medical practice to switch to a different treatment for the next course. We consider settings where it is reasonable to define outcome as a discrete variable observed within a sufficiently short time period such that interim monitoring is feasible. Most statistical approaches to this or similar settings characterize patient outcome as a single binary variable by collapsing the multicourse structure and combining death and failure, and typically evaluate only one treatment for each patient. Such approaches waste important information, because each patient may receive several different treatments over successive courses, these treatments may have interactive effects, and the distinction between death and treatment failure is very important clinically.

In this article we provide a statistical framework for treatment evaluation and adaptive clinical trial design and conduct in multicourse settings. We take a Bayesian approach, because it provides a natural basis for incorporating historical data and making inferences sequentially during the trial and on its completion. The methodology is presented in the context of the two-course chemotherapy trial that motivated this research. The trial involves acute myelogenous leukemia (AML) patients who previously achieved a CR but subsequently relapsed in less than 24 months. For these patients, the outcome probabilities vary with age and the length of first CR. Each patient receives either one or two courses of chemotherapy. The three possible outcomes for each course, determined within 1 month from initiation of that course's treatment, are CR, death, and failure (when the patient is alive but has not achieved CR). The occurrence of either death or CR, or the completion of two courses that are both failures, marks the end of the patient's therapy. Patients with two courses of treatment failure are given subsequent palliative care. This definition of therapeutic outcome is motivated by the poor overall survival time in AML patients who do not achieve CR and the necessity of achieving CR as a precursor to long-term survival. Hence, although no discrete early outcome has been shown to be a perfect surrogate for survival time, CR is a valuable and universally accepted early endpoint in AML therapeutics. Moreover, once failure has occurred in each of two courses, the probability of a subsequent CR is very low. The trial includes the standard chemotherapy combination of idarubicin + high-dose cytosine arabinoside (IDA), and two experimental treatments, IDA + mylotarg (M) and IDA + topotecan (T). For the first course, all patients are randomized fairly among the three treatments. A patient for whom IDA fails in the first course is randomized between IDA + M and IDA + T for the second course. A patient for whom either IDA + M or IDA + T fails in the first course must

Peter F. Thall is Professor and Deputy Chairman, Department of Biostatistics, Houston, TX 77030 (E-mail: *rex@mdanderson.org*); Hsi-Guang Sung is Statistical Analyst, Department of Biostatistics (E-mail: *hgs@odin.mdacc.tmc.edu*); and Elihu H. Estey is Professor, Department of Leukemia, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030. Research for this article was partially supported by National Institutes of Health grant CA83932.

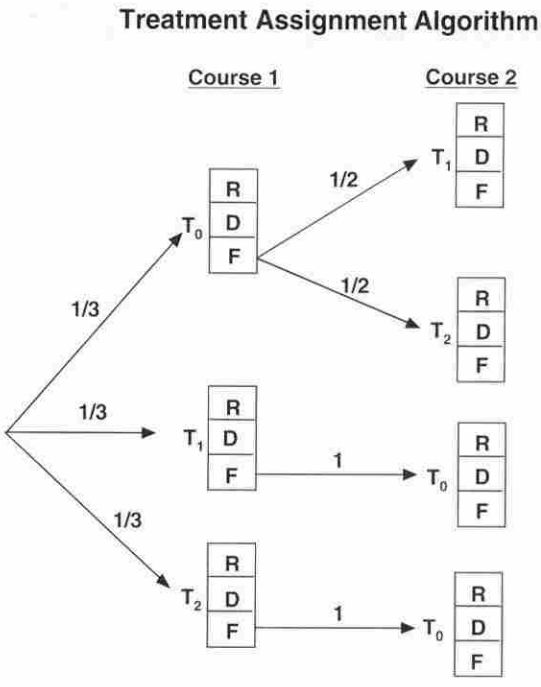

Figure 1. Schematic for Conduct of the AML Trial. R = response; D = death; F = failure to respond but alive; T_0 = idarubicin + ara-C (IDA); T_1 = IDA + mylotarg; T_2 = IDA + topotecan.

receive IDA in the second course, however, because it is considered unacceptable to give a patient experimental treatments in both courses. Figure 1 illustrates this treatment assignment algorithm.

The primary scientific goal of the trial is to select the best two-course treatment strategy within each prognostic subgroup based on the probabilities of CR and death. Trial conduct is outcome-adaptive in that if interim data show a particular treatment strategy to be inferior to the others within a subgroup, then that strategy is dropped within that subgroup for the remainder of the trial. The design thus comprises an algorithm for assigning a treatment to each patient in each course, interim safety monitoring rules, and, at the end of the trial, treatment strategy selection within prognostic subgroups. The method requires a real-valued objective function of the probabilities of CR and death that quantifies the clinically acceptable trade-off between these two outcomes. This function is used as a basis for interim decision making and inferences at the conclusion of the trial. In practice, the trade-off function is elicited from the physicians planning the trial; we illustrate how this may be done using contour plots as a graphical aid.

We use a generalized logistic regression model to characterize the probabilities of CR and death in each course as functions of the patient's treatments and prognostic covariates. The model also allows for pairwise interactions between treatment strategy, course, and covariates. Because the probabilities of CR and death may vary with patient prognosis, different prognostic subgroups may have different optimal treatment strategies. In addition to its application in the context of trial design and conduct, the regression model is also a useful analytic tool for evaluating covariate and treatment strategy effects on the probabilities of response and death based on existing data.

Table 1. Outcome Counts for Each Course and Treatment Combination in the Historical AML Data

Treatments		Outcome		
t_1	t_2	CR	Death	# Patients
Course 1				
0		84 (.27)	66 (.21)	316
1		50 (.56)	18 (.20)	89
2		13 (.04)	41 (.13)	309
Course 2				
0	0	14 (.17)	24 (.29)	82
0	1	5 (.36)	5 (.36)	14
0	2	0 (.00)	5 (.22)	23
1	0	1 (.14)	5 (.71)	7
1	1	1 (.50)	0 (.00)	2
1	2	0 (.00)	0 (.00)	3
2	0	4 (.11)	12 (.34)	35
2	1	3 (.33)	3 (.33)	9
2	2	4 (.02)	26 (.16)	159

NOTE: Row probabilities are given in parentheses. 0 = high dose ara-C, 1 = allogeneic bone marrow transplant, and 2 = chemotherapy without ara-C.

As the first step in developing a design, we fit the model to historical data from 714 AML patients treated at M.D. Anderson Cancer Center between 1990 and 1999. This analysis served to validate the model, obtain informative distributions of model parameters unrelated to treatment, and also obtain reasonable numerical values of parameters for use in a simulation study of the design. Like the patients in the trial being planned, each historical patient previously achieved CR but later relapsed and then received salvage therapy in an attempt to reinduce remission. The salvage treatments were allogeneic bone marrow transplant, combination chemotherapy containing high-dose ara-C, or chemotherapy not including ara-C. The data for each patient consisted of prognostic covariates and the treatment and outcome in each of one or two courses. A summary of the empirical outcome probabilities in each course, ignoring prognostic covariates, is given in Table 1.

In Section 2 we describe the probability model and approximate Bayesian method that serve as the basis for treatment evaluation and trial design. We describe a general strategy for constructing trial designs in Section 3. In Section 4 we summarize our analysis of the historical data. We describe the AML trial design in Section 5, and summarize a simulation study of the design in Section 6. We close with a discussion in Section 7.

2. PROBABILITY MODELS

2.1 A Two-Course Model

Denote by (s, t) the two-course treatment strategy wherein the patient receives treatment T_s in the first course and, if the first course results in failure, receives T_t in the second course. Denoting IDA, IDA + M, and IDA + T by T_0, T_1, and T_2, the AML trial design allows for the four two-course strategies $\mathcal{S} = \{(1, 0), (2, 0), (0, 1), (0, 2)\}$. Although strategies (1, 2) and (2, 1) are not permitted in the AML trial, in general the methodology allows for $\mathcal{S}$ to contain any two-course combination, including strategies of the form (s, s) that give the same

treatment in both courses. Each two-course strategy (s, t) has five possible outcomes. Therapy may end in the first course with either response or death with T_s, or failure with T_s in the first course may be followed by response, death, or failure with T_t in the second course.

For each patient baseline prognostic covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_q)$, the goal is to select the best two-course treatment strategy (s, t) from $\mathcal{S}$ based on the probabilities $\xi_R(s, t, \mathbf{Z})$ of achieving CR and $\xi_D(s, t, \mathbf{Z})$ of death. For course $c = 1$ or 2, let τ_c denote the treatment index 0, 1, or 2, and let Y_{Rc} and Y_{Dc} denote the indicators of response and death, so that $Y_{Fc} = 1 - Y_{Rc} - Y_{Dc}$ indicates failure. Because there is no second course if $Y_{F1} = 0$, for completeness we define $Y_{R2} = Y_{D2} = 0$ and $\tau_2 = 0$ in this case. Denote the probability of outcome $k = R, D$, or F with T_s in course 1 by

$$\pi_{k1}(s, \mathbf{Z}) = \Pr[Y_{k1} = 1 \mid \mathbf{Z}, \tau_1 = s], \tag{1}$$

and the probability of outcome k with T_t in course 2 after a failure with T_s in course 1 by

$$\pi_{k2}(s, t, \mathbf{Z}) = \Pr[Y_{k2} = 1 \mid \mathbf{Z}, \tau_1 = s, Y_{F1} = 1, \tau_2 = t]. \tag{2}$$

Aside from covariates, π_{k1} is a function of τ_1 alone, whereas π_{k2} is a function of both τ_1 and τ_2. Because one of R, D, or F must occur in each course and the occurrence of either R or D, or two treatment failures, marks the end of the patient's therapy, for any strategy (s, t),

$$\pi_{R1}(s, \mathbf{Z}) + \pi_{D1}(s, \mathbf{Z}) + \pi_{F1}(s, \mathbf{Z}) \sum_{k=R,D,F} \pi_{k2}(s, t, \mathbf{Z}) = 1. \tag{3}$$

The likelihood function of the ith patient thus takes the form

$$\mathcal{L}_i = \prod_{k=R,D,F} \{\pi_{k1}(\tau_{i1}, \mathbf{Z}_i)\}^{Y_{i,k1}} \times \left\{ \prod_{r=R,D,F} [\pi_{r2}(\tau_{i1}, \tau_{i2}, \mathbf{Z}_i)]^{Y_{i,r2}} \right\}^{Y_{i,F1}}, \tag{4}$$

with $\mathcal{L} = \prod_{i=1}^{n} \mathcal{L}_i$ the likelihood of a sample of n patients.

2.2 A Generalized Logistic Model

The following generalized logistic model (cf. Agresti 1990, chap 9.2) accounts for trinary outcomes, the two-course treatment structure, and prognostic covariates. The formulation also accommodates an arbitrary number of treatments and any collection of two-course treatment sequences formed from them. In addition to its use as a basis for clinical trial design and conduct, this regression model is also very useful per se when the primary goal is to analyze existing data consisting of trinary outcomes with covariates.

For outcome $k = R$ or D, treatment strategy (s, t), and covariates $\mathbf{Z}$, denote the linear components corresponding to courses 1 and 2 by

$$\eta_{k1}(s, \mathbf{Z}) = \mu_k + \alpha_k(s) + \sum_{j=1}^{q} \{\gamma_{kj} + \zeta_{kj}(s)\} Z_j, \tag{5}$$

and

$$\eta_{k2}(s, t, \mathbf{Z}) = \mu_k + \alpha_k(t) + \beta_k(s, t) + \sum_{j=1}^{q} \{\gamma_{kj} + \zeta_{kj}(t) + \delta_{kj}\} Z_j, \tag{6}$$

subject to the $2(q+1)$ constraints

$$\sum_s \alpha_R(s) = \sum_s \alpha_D(s) = 0$$

and

$$\sum_s \zeta_{Rj}(s) = \sum_s \zeta_{Dj}(s) = 0, \quad j = 1, \ldots, q. \tag{7}$$

Our application includes $q = 2$ covariates, and hence 6 constraints, so we set $\alpha_k(0) = 0$ and $\zeta_{kj}(0) = 0$ for $k = R, D$ and $j = 1, 2$. That is, we use $s = 0$ as the baseline treatment group.

We characterize the regression of the outcomes $\mathbf{Y}_1 = (Y_{R1}, Y_{D1})$ and $\mathbf{Y}_2 = (Y_{R2}, Y_{D2})$ on treatment strategy (s, t) and covariates $\mathbf{Z}$ by the probability functions

$$\pi_{k1}(s, \mathbf{Z}) = \frac{\exp\{\eta_{k1}(s, \mathbf{Z})\}}{1 + \exp\{\eta_{R1}(s, \mathbf{Z})\} + \exp\{\eta_{D1}(s, \mathbf{Z})\}} \tag{8}$$

and

$$\pi_{k2}(s, t, \mathbf{Z}) = \frac{\exp\{\eta_{k2}(s, t, \mathbf{Z})\}}{1 + \exp\{\eta_{R2}(s, t, \mathbf{Z})\} + \exp\{\eta_{D2}(s, t, \mathbf{Z})\}}, \tag{9}$$

where $k = R$ or D, with each $\pi_{Fc} = 1 - (\pi_{Rc} + \pi_{Dc}) = 1/[1 + \exp(\eta_{Rc}) + \exp(\eta_{Dc})]$. Under this generalized logistic model, for each $k = R$ or D, the intercept of η_{kc} is decomposed into the baseline mean μ_k, the main effect $\alpha_k(s)$ of treatment s, and, for course 2, the additional effect $\beta_k(s, t)$ of t as a salvage treatment following failure with s. Similarly, the coefficient of Z_j for outcome Y_k is decomposed into the baseline parameter γ_{kj}, the treatment effect $\zeta_{kj}(s)$, and the course 2 effect δ_{kj}. Viewing (π_{k1}, π_{k2}) as a function of treatment, course, and covariates, if we write $\beta_k(s, t) = \beta_k + \beta_k^*(s, t)$ with $\sum_{(s,t)} \beta_k^*(s, t) = 0$ for each k, then $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the treatment, course, and covariate main effects and $\boldsymbol{\beta}^*$, $\boldsymbol{\zeta}$, and $\boldsymbol{\delta}$ are the [treatment × course], [treatment × covariate], and [covariate × course] interactions. If there are m treatments and r_m two-course strategies, then, subject to the constraints, the vector $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta}^*)$ of intercept parameters has dimension $2(m + r_m)$ and the vector $(\boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\delta})$ of covariate effect parameters has dimension $2q(m+1)$, so that the overall model dimension is $p = 2(q + m + qm + r_m)$. For the AML trial, $p = 30$, because $q = 2$, $m = 3$, and $r_m = 4$.

The probability of overall outcome $k = R$ or D, in either one or two courses, with the treatment strategy (s, t) for a patient with covariates $\mathbf{Z}$ is

$$\xi_k(s, t, \mathbf{Z}) = \pi_{k1}(s, \mathbf{Z}) + \pi_{F1}(s, \mathbf{Z}) \pi_{k2}(s, t, \mathbf{Z}). \tag{10}$$

We use the overall probabilities $\boldsymbol{\xi}(s, t, \mathbf{Z}) = (\xi_R(s, t, \mathbf{Z}), \xi_D(s, t, \mathbf{Z}))$ of response and death as the basis for both interim safety monitoring and treatment strategy selection, because these are what matter clinically. Because $\boldsymbol{\xi}(s, t, \mathbf{Z})$ is two-

dimensional, using this criterion to compare treatment strategies is problematic. We thus define an objective function (see Sec. 2.4) that reduces $\boldsymbol{\xi}(s, t, \mathbf{Z})$ to a single real number by quantifying the trade-off between the probability of response and the risk of death.

2.3 An Approximate Bayesian Method

In addition to providing a practical framework for formulating interim monitoring and treatment strategy selection criteria, the following Bayesian formulation also facilitates interpretation and explanation to medical colleagues, many of whom think and behave like Bayesians. To meet the practical demands of evaluating and explaining the design to a broad audience, we also compute the design's frequentist OCs. Thus our approach may be regarded as a hybrid of Bayesian and frequentist methods.

We use the following computational approximation. Assume a priori that $\boldsymbol{\theta}$ is multivariate normal, denoted by $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_\theta, \Omega)$. Under the usual frequentist large-sample theory, the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is approximately multivariate normal, denoted $\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \dot{\sim} N(\boldsymbol{\theta}, \Sigma)$. It follows from Bayes's theorem that, a posteriori, $\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}} \dot{\sim} N(B\mathbf{b}, B)$, where $B = (\Sigma^{-1} + \Omega^{-1})^{-1}$ and $\mathbf{b} = \Sigma^{-1}\hat{\boldsymbol{\theta}} + \Omega^{-1}\boldsymbol{\mu}_\theta$ (Lindley and Smith 1972). This approach has been used by many authors, including Dixon and Simon (1991) in the context of Bayesian subset selection and Faraggi and Simon (1997) in proportional hazards regression. The method is straightforward, because it relies on multivariate normal distributions. The necessary computations include deriving the MLE, computing an estimator $\hat{\Sigma}$ of the covariance matrix, and generating multivariate normal posterior samples using a Cholesky decomposition. It may be implemented with standard statistical software and provides a practical alternative to more computationally intensive methods, such as Markov chain Monte Carlo (MCMC).

2.4 An Objective Function

The following function of $\xi_R(s, t, \mathbf{Z})$ and $\xi_D(s, t, \mathbf{Z})$ quantifies the trade-off between these two probabilities. We use it as a basis for both interim monitoring and treatment selection. Temporarily suppress the argument $(s, t, \mathbf{Z})$. The function ϕ is constructed so that all pairs (ξ_R, ξ_D) for which $\phi(\xi_R, \xi_D)$ equals a given constant are equally desirable. The process of eliciting ϕ from the physicians planning the trial may be facilitated by interactively modifying (ξ_R, ξ_D) while viewing $\phi(\xi_R, \xi_D)$ on a computer screen.

For the AML trial, we began with the family of linear objective functions $\phi = a\xi_R + b\xi_D$ in the triangular two-dimensional domain of (ξ_R, ξ_D) over a range of (a, b) values with $a > 0 > b$. We determined ϕ by specifying two equations and solving for a and b. The null value $(\xi_R, \xi_D) = (.40, .40)$ corresponding to all patients in the historical data was assigned $\phi = 0$, and the desirable goal $(\xi_R, \xi_D) = (.50, .15)$ was assigned $\phi = 1$. The values 0 and 1 for ϕ in these two cases were chosen purely for numerical convenience. After examining plots of the resulting linear contours, we decided that ϕ should increase more rapidly in ξ_R for smaller values of ξ_D, especially for ξ_D near 0. We thus considered functions of the more general form

$$\phi(\xi_R, \xi_D) = a\xi_R + b\xi_D^c, \tag{11}$$

with $a > 0 > b$ and $c > 0$. Given the foregoing two constraints, a third equation to determine c was given by the value of ξ_R that would be required to still have $\phi = 1$ if there were no fatalities, that is, the value of ξ_R such that $\phi(\xi_R, 0) = \phi(.50, .15) = 1$. After examining contour plots corresponding to several different values of ξ_R using the three-parameter version of ϕ, we specified this to be $\xi_R = .30$. A contour plot of the resulting ϕ, which is characterized by $a = 3.333$, $b = -2.548$ and $c = .707$, is given in Figure 2.

Other functional forms for ϕ could be used, provided that ϕ increases in ξ_R and decreases in ξ_D. The shape of its contours should provide a reasonably flexible graphical representation of the trade-off between ξ_R and ξ_D that reflects the physicians' goals and opinions. The particular shape of our trade-off function contours is one of several geometries in the two-dimensional parameter plane that have been proposed to characterize the trade-off between safety and efficacy. To define hypotheses for tests based on bivariate outcomes, Willan and Pater (1985) used two parallel lines that partition the plane into three hypotheses, Jennison and Turnbull (1993) and Bryant and Day (1995) used various rectangular regions, and Thall and Cheng (1999) proposed polygonal regions.

The probability model, probabilities of response and death in each course, overall probabilities of response and death, and objective function constitute a parametric hierarchy. The mapping $\boldsymbol{\theta} \to \boldsymbol{\pi}(s, t, \mathbf{Z}) = (\pi_{R1}(s, \mathbf{Z}), \pi_{D1}(s, \mathbf{Z}), \pi_{R2}(s, t, \mathbf{Z}), \pi_{D2}(s, t, \mathbf{Z}))$ reduces the parameter vector to the probabilities of response and death in each course using the strategy (s, t) in prognostic group $\mathbf{Z}$. Next, mapping $\boldsymbol{\pi}(s, t, \mathbf{Z}) \to \boldsymbol{\xi}(s, t, \mathbf{Z})$ into the two-dimensional triangular region illustrated in Figure 2 limits attention to the overall two-course probabilities of response and death. The final real-valued mapping $\boldsymbol{\xi}(s, t, \mathbf{Z}) \to \phi(\boldsymbol{\xi}(s, t, \mathbf{Z}))$ induces an ordering among the strategies, thus providing a basis for comparison and selection.

3. A GENERAL DESIGN STRATEGY

Our overall strategy for trial design and conduct is as follows. The first step is to formulate a generalized logistic model that accommodates the particular multicourse structure of the trial at hand, including the k-nary outcome and the maximum number of courses. Denote the subvector of $\boldsymbol{\theta}$ comprising parameters corresponding to specific treatments by $\boldsymbol{\theta}_T = (\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\zeta})$ and the vector of baseline parameters not specific to any treatments by $\boldsymbol{\theta}_B = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta})$. Let $\boldsymbol{\theta}_{T(H)}$ denote the treatment-specific parameter vector corresponding to the treatments in the historical data χ_H. We make a sharp distinction between $\boldsymbol{\theta}_{T(H)}$ and the effects $\boldsymbol{\theta}_T$ of the treatments to be studied in the trial being planned.

Starting with a reasonably noninformative prior on $\boldsymbol{\theta} = (\boldsymbol{\theta}_B, \boldsymbol{\theta}_{T(H)})$, the model is fit to χ_H and, based on the MLE $(\hat{\boldsymbol{\theta}}_B, \hat{\boldsymbol{\theta}}_{T(H)})$, the approximate Bayesian method is used to obtain the marginal posterior $f(\boldsymbol{\theta}_B \mid \chi_H)$. On specifying a suitable prior $f(\boldsymbol{\theta}_T)$ on the vector $\boldsymbol{\theta}_T$ of new treatment parameters, the prior of $\boldsymbol{\theta} = (\boldsymbol{\theta}_B, \boldsymbol{\theta}_T)$ at the start of the trial is $f(\boldsymbol{\theta}_B \mid \chi_H) f(\boldsymbol{\theta}_T)$. All inferences during the trial and at its conclusion are based on posteriors obtained from maximum likelihood estimates (MLEs) of $(\boldsymbol{\theta}_B, \boldsymbol{\theta}_T)$ from the trial data, again using the approximate Bayesian method. Thus all of the

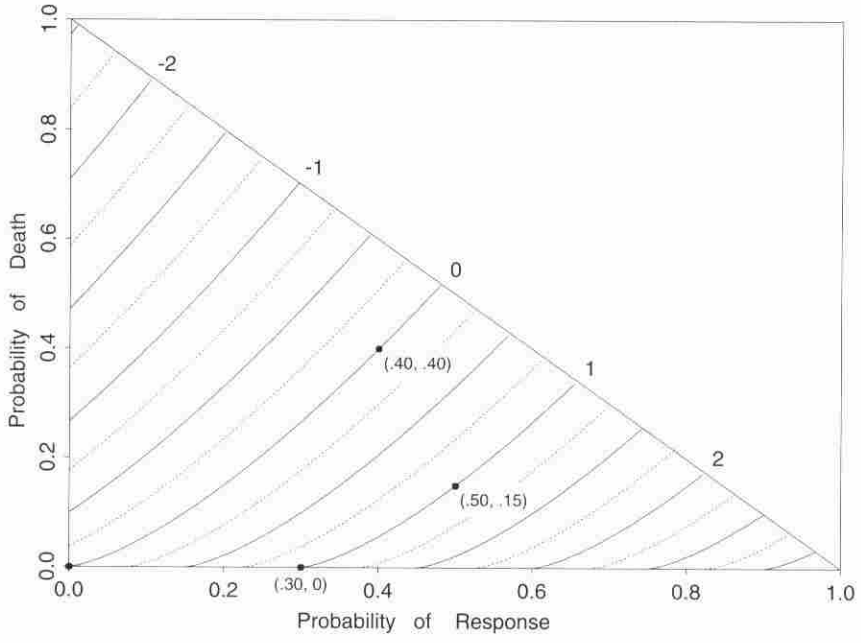

Figure 2. Contour Plot of the Objective Function $\phi(\xi_R, \xi_D) = 3.333\xi_R - 2.548\xi_D^{.707}$.

information from the historical data used in the trial design and conduct is contained in $f(\boldsymbol{\theta}_B \mid \chi_H)$, and in particular $\boldsymbol{\theta}_B$ has dimension $4(q+1)$.

Denoting a given J-course treatment strategy by $\boldsymbol{\tau} = (t_1, \ldots, t_J)$ and the vector of probabilities of the possible outcomes with $\boldsymbol{\tau}$ over J courses by $\boldsymbol{\xi}(\boldsymbol{\tau})$, comparison of different multicourse treatment strategies is based on a real-valued objective function $\phi(\boldsymbol{\xi}(\boldsymbol{\tau}))$ elicited from the physician(s) planning the trial. Interim decisions to drop comparatively inferior strategies and selection of a best strategy at the end of the trial may be based on posterior probabilities such as $\Pr[\phi(\boldsymbol{\xi}(\boldsymbol{\tau}_1)) < \phi(\boldsymbol{\xi}(\boldsymbol{\tau}_2)) \mid \chi]$, on posterior means, or on predictive probabilities. Inferences may be made for prognostic subgroups so that, for example, strategy $\boldsymbol{\tau}_1$ may be best for one subgroup, whereas strategy $\boldsymbol{\tau}_2$ is best for another. Given a maximum sample size and criteria for interim decisions, the design's operating conditions (OCs) may be evaluated via simulation, with design parameters calibrated on that basis. The OCs may include the sample size distribution, probabilities of dropping treatment strategies during the trial, and final strategy selection probabilities. We have found it useful to compute the design's OCs under several different clinical scenarios, each characterized by fixed probabilities of the possible outcomes.

4. ANALYSIS OF THE HISTORICAL DATA

For all model fits reported here, both in analysis of the historical data and in fitting models to simulated datasets, a priori all parameters in each model were assumed to be iid normal random variables with mean 0 and variance 10. In settings where the parameters are known to be dependent, incorporating this into the prior might improve the design's efficiency.

The prognostic covariates used in the model-based analysis of the historical data were the binary indicators of whether the patient's age was under 50 years and whether the patient's initial remission duration prior to entering the trial was at least 1 year. Thus $q = 2$, and there were four prognostic subgroups. For example, the group having worst prognosis comprised the older patients with short initial CR duration, whereas the best prognostic group comprised the younger patients with long initial CR duration. There were $m = 3$ treatments, and the treatment effects in the model correspond to allogeneic bone marrow transplant ($s = 1$) and chemotherapy not including ara-C ($s = 2$) relative to the baseline treatment group comprised of chemotherapy containing high-dose ara-C ($s = 0$). Because there were $r_m = 9$ different two-course treatment combinations, the full model has a total of $p = 40$ parameters.

Starting with the full model and including $\boldsymbol{\theta}_B$ throughout, we obtained a more parsimonious model by successively eliminating entries from the parameter vector $\boldsymbol{\theta}_{T(H)}$ pertaining to treatments in the historical data. We considered only hierarchical models. Because the two elements of each pair $\boldsymbol{\zeta}_j(s) = (\zeta_{R,j}(s), \zeta_{D,j}(s))$ act together, we either included or deleted both entries. For model comparison, we used the maximized log-likelihood, the posterior parameter variances, and the *Bayes information criterion* (BIC) (Kass and Raftery 1995),

$$\mathrm{BIC}(\mathcal{M}) = \log \mathcal{L}(\hat{\boldsymbol{\theta}}_{\mathcal{M}}) - 1/2\ p_{\mathcal{M}} \log(n), \qquad (12)$$

where $p_{\mathcal{M}}$ is the number of parameters in model $\mathcal{M}$. In particular, the BIC penalizes the log-likelihood for larger $p_{\mathcal{M}}$.

The fitted models that we considered are summarized in Table 2. We first eliminated the vector $\boldsymbol{\beta}^*$ of treatment–course interactions (model 3), because this increased the BIC greatly, much more than the increase obtained by eliminating $\boldsymbol{\zeta}$ (model 2). Moreover, under the full model, the absolute value of the posterior mean of each entry of $\boldsymbol{\beta}^*$ was small compared to its standard deviation. Next, focusing on the four pairs of treatment–covariate interactions, $\{\boldsymbol{\zeta}_j(s), j = \mathrm{AGE}, \mathrm{DUR}, s = 1, 2\}$, we successively eliminated pairs in a stepdown manner. We stopped with model 4, our final model, because

Table 2. Summary of Models Fit to the Historical Data

Model	Parameters dropped	p	$\log \mathcal{L}(\hat{\boldsymbol{\theta}})$	BIC
1	None (full model)	40	−796	−928
2	$\boldsymbol{\zeta}$	32	−801	−906
3	$\boldsymbol{\beta}^*$	24	−801	−881
4	$\boldsymbol{\beta}^*, \zeta_{\text{AGE}}(1), \zeta_{\text{AGE}}(2), \zeta_{\text{DUR}}(2)$	18	−803	−862
5	$\boldsymbol{\beta}^*, \boldsymbol{\zeta}$	16	−807	−859

$\Pr\{\zeta_{D,\text{DUR}}(1) > 0 \mid \text{data}\} = .95$; hence it was appropriate to retain the pair $\boldsymbol{\zeta}_{\text{DUR}}(1)$.

Posterior means and corresponding MLEs of the parameters in the final model are given in Table 3. The signs of the estimates $\{\hat{\alpha}_k(s), k = R, D, s = 1, 2\}$ of the main treatment effects show that, relative to high-dose ara-C, transplant had higher rates of both CR and death, whereas nonara-C chemo had lower rates of both events. The well-known fact that the CR rate decreases and the death rate increases in a second course of treatment after failure in a previous course is borne out by the relationship $\hat{\beta}_R < 0 < \hat{\beta}_D$.

The signs of the remaining parameter estimates in Table 3 should be interpreted in the context of the generalized logistic model's algebraic structure, which differs from that of the usual logistic model. For example, although the fact that $\hat{\gamma}_{D,\text{DUR}} > 0$ considered per se might seem to imply the model predicts a higher probability of death for patients with a longer initial CR duration, this is not the case. The effect of a given covariate on π_{Dc} is determined by all of that covariate's coefficients, including those indexed by both R and D. The numerical values of $\hat{\gamma}_{R,\text{DUR}}$, $\hat{\gamma}_{D,\text{DUR}}$, $\hat{\zeta}_{R,\text{DUR}}(1)$, and $\hat{\zeta}_{D,\text{DUR}}(1)$ act together so that $\hat{\pi}_{D1}$ decreases and $\hat{\pi}_{R1}$ increases with longer initial CR duration, as should be the case on medical grounds. This illustrates the fact that these four parameters act together algebraically for each treatment to determine (π_{R1}, π_{D1}). Similarly, these parameters and $(\delta_{R,2}, \delta_{D,2})$ act together to determine the effect of Z_2 on (π_{R2}, π_{D2}). Table 4, which gives the predicted and empirical overall CR and death probabilities within each prognostic group for patients who received high dose ara-C in both courses, shows that the fitted model gives predictions that make sense for the four prognostic subgroups. Table 4 illustrates the importance of accounting for prognosis, because $\hat{\xi}_R$ increases and $\hat{\xi}_D$ decreases with increasing CR duration and with younger age, and these changes are quite large. Moreover, the good agreement between the model-based estimates and the corresponding empirical values provides a further validation of the model.

Table 3. MLEs and Posterior Means of the Generalized Logistic Model Parameters for the Historical Data

	Estimates	
Parameter	MLE	Bayesian
μ_R	$-1.35_{.25}$	$-1.34_{.24}$
$\alpha_R(1)$	$1.74_{.30}$	$1.72_{.30}$
$\alpha_R(2)$	$-2.14_{.28}$	$-2.13_{.28}$
μ_D	$-.68_{.20}$	$-.68_{.20}$
$\alpha_D(1)$	$.56_{.36}$	$.56_{.35}$
$\alpha_D(2)$	$-1.06_{.19}$	$-1.06_{.19}$
β_R	$-.46_{.60}$	$-.47_{.58}$
β_D	$.47_{.28}$	$.46_{.28}$
$\gamma_{R,\text{DUR}}$	$1.57_{.27}$	$1.54_{.27}$
$\gamma_{R,\text{AGE}}$	$.22_{.28}$	$.22_{.27}$
$\gamma_{D,\text{DUR}}$	$.00_{.32}$	$-.01_{.31}$
$\gamma_{D,\text{AGE}}$	$-.44_{.25}$	$-.44_{.25}$
$\zeta_{R,\text{DUR}}(1)$	$-.26_{.75}$	$-.28_{.71}$
$\zeta_{D,\text{DUR}}(1)$	$1.36_{.85}$	$1.31_{.81}$
$\delta_{R,\text{DUR}}$	$-.64_{.56}$	$-.60_{.55}$
$\delta_{R,\text{AGE}}$	$.08_{.66}$	$.08_{.64}$
$\delta_{D,\text{DUR}}$	$-.99_{.54}$	$-.95_{.53}$
$\delta_{D,\text{AGE}}$	$.14_{.40}$	$.14_{.39}$

NOTE: Standard deviations are given as subscripts.

5. TRIAL CONDUCT

Aside from technical details related to accounting for two courses, the trinary outcome, and four prognostic subgroups, in principle the conduct of the AML trial is straightforward. Patients are randomized fairly among the acceptable treatments at each of two stages, using dynamic allocation to balance on the two covariates. Halfway through the trial, a safety monitoring rule is applied within each prognostic subgroup to drop any treatment strategy that is comparatively inferior. The stage 2 randomization thus accounts for all strategies that are dropped within each prognostic subgroup after stage 1. At the end of the trial, the best strategy for each prognostic subgroup is selected. Formally, the trial is conducted as follows.

Stage 1. Randomize $n/2$ patients fairly among the three treatments for their first course of therapy, using the Pocock and Simon (1975) algorithm to balance on Z_1 and Z_2. Patients who fail with T_0 in their first course are randomized between T_1 and T_2 for their second course. All patients who fail with

Table 4. Estimated Two-Course Probabilities of Response and Death, and Objective Function Values, by Prognostic Group, for Historical Patients Treated with High-Dose ara-C in Both Courses

		Model-Based			Empirical			
CR duration	Age	$\hat{\xi}_R$	$\hat{\xi}_D$	$\hat{\phi}$	$\hat{\xi}_R$	$\hat{\xi}_D$	$\hat{\phi}$	n
Short	Old	$.19_{.05}$	$.52_{.07}$	$-.95_{.31}$	$.16_{.04}$	$.44_{.05}$	$-.90_{.17}$	29
Short	Young	$.27_{.10}$	$.40_{.10}$	$-.42_{.54}$	$.31_{.05}$	$.40_{.06}$	$-.31_{.22}$	30
Long	Old	$.54_{.12}$	$.25_{.09}$	$.85_{.60}$	$.63_{.07}$	$.27_{.06}$	$1.10_{.26}$	11
Long	Young	$.65_{.13}$	$.16_{.08}$	$1.48_{.65}$	$.62_{.06}$	$.16_{.05}$	$1.40_{.25}$	12

NOTE: Standard deviations are given as subscripts.

either T_1 or T_2 in course 1 are treated with T_0 in course 2. If

$$\Pr[\phi(s,t,\mathbf{Z}) > \phi(u,v,\mathbf{Z}) \mid \text{data}] > .95 \qquad (13)$$

for distinct strategies (s,t) and (u,v), then drop strategy (u,v) in patient subgroup $\mathbf{Z}$.

Stage 2. Randomize $n/2$ additional patients among the treatments in each course as in stage 1, subject to the constraints imposed by dropping any treatment strategies. Once n patients have been treated and evaluated, for each $\mathbf{Z}$ select the two-course strategy, among those not dropped in that subgroup, for which the posterior mean of $\phi(s,t,\mathbf{Z})$ is largest.

6. SIMULATION STUDY

The simulations were designed to provide a reasonable reflection of actual trial conduct. Although the purpose of the trial is to learn about the treatment-related parameters $\boldsymbol{\theta}_T = (\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\zeta})$, we use the historical data to obtain preliminary knowledge about the nontreatment-related parameters $\boldsymbol{\theta}_B = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta})$. An important point is that the particular treatments in the historical data and thus the corresponding parameters $\boldsymbol{\theta}_{T(H)}$ are different from the treatments and parameters $\boldsymbol{\theta}_T$ to be studied in the trial being planned. The values of $\boldsymbol{\theta}_{T(H)}$ per se are not relevant to inferences about either $\boldsymbol{\theta}_T$ or the treatment strategies in the trial. In fitting each simulated dataset, we applied the approximate Bayesian method using the posterior $f(\boldsymbol{\theta}_B \mid \chi_H)$ from the fitted historical data, under the model summarized in Table 3, as the prior of the nontreatment-related parameters $\boldsymbol{\theta}_B$ and noninformative iid $N(0, 10)$ priors for the treatment effect parameters in $\boldsymbol{\theta}_T$.

6.1 Clinical Scenarios

Because the two-course, trinary outcome setting considered here is more complex than a single-course selection trial based on a univariate outcome, our design and criteria for evaluating its performance necessarily also are more complex. To provide a conceptual framework for what follows, we first briefly review the analogous single-course setting with a univariate outcome where the goal is to select the best among k treatments based on estimates of their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$. Without loss of generality, assume that $\mu_1 \leq \cdots \leq \mu_k$. For a randomized trial to select a single best treatment, let μ_0 be a null value and let $\mu_0 + \delta$ be a desirable target, where δ is a clinically significant improvement over μ_0. The *null configuration* $\boldsymbol{\mu}^0$ is the k-vector having all $\mu_j = \mu_0$, whereas the *least favorable configuration* (LFC) $\boldsymbol{\mu}^*$ has $\mu_1 = \cdots = \mu_{k-1} = \mu_0$ and $\mu_k = \mu_0 + \delta$ (cf. Gibbons, Olkin, and Sobel 1977, Chapter 1.3). It can easily be shown that among the set of $\boldsymbol{\mu}$ having no entries between μ_0 and $\mu_0 + \delta$ and at least one entry $\geq \mu_0 + \delta$, the LFC minimizes the probability of correct selection (PCS) of treatment k. Because the PCS under $\boldsymbol{\mu}^*$ increases with sample size, n, given $\boldsymbol{\mu}_0$ and δ one may determine n to achieve a given PCS. In the present setting, one may regard the two contours on which $\phi(\xi_E, \xi_D) = 0$ and 1 as two-dimensional generalizations of the points μ_0 and $\mu_0 + \delta$ in the one-dimensional case.

The two clinical scenarios given in Table 5 may be regarded as multidimensional generalizations of the LFC in the one-dimensional case. We use these scenarios, along with one more complex scenario that is not tabled, as a basis for evaluating the selection design and for determining sample size. Because we account for trinary outcomes, two treatment courses, and four patient prognostic groups, our parametric characterizations of clinical settings are necessarily more complex than those in the univariate single-course case. Consequently, there are more qualitatively different cases than the two, described above, that typically are considered in the one-dimensional case. The three scenarios under which we evaluate the design's OCs here were chosen to cover a reasonable range of cases that may actually obtain in practice, and they should illustrate the design's essential properties.

We determined each scenario in Table 5 by first specifying values for the 14 probabilities $\{\pi_{k1}(s,\mathbf{0}), \pi_{k2}(s,t,\mathbf{0})\}$ corresponding to $\mathbf{Z} = \mathbf{0}$ and then using these values to determine the 14 parameters $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ via a one-to-one transformation. These parameters in turn determine the linear components $\eta_{k1}(s,\mathbf{0})$ and $\eta_{k2}(s,t,\mathbf{0})$. The probabilities $\{\pi_{k1}(s,\mathbf{Z}), \pi_{k2}(s,t,\mathbf{Z})\}$ for $\mathbf{Z} \neq \mathbf{0}$ were obtained by adding the covariate adjustment terms $(\boldsymbol{\gamma} + \boldsymbol{\delta} I[c=2])'\mathbf{Z}$ to the linear components $\eta_{k1}(s,\mathbf{0})$ and $\eta_{k2}(s,t,\mathbf{0})$ using the posterior means of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ from the historical data. To obtain covari-

Table 5. Operating Characteristics of the Design Under Scenarios A and B for the Prognostic Subgroup With Short CR Duration and Younger Age

Scenario	Treatment strategy	ξ_B	ξ_D	ϕ	Decision probabilities: Selected	Decision probabilities: Dropped early	Number of patients
A	(0, 1)	.46	.25	.55	[.76]	.03	7.7
	(0, 2)	.22	.32	−.39	.04	[.32]	5.3
	(1, 0)	.22	.32	−.39	.08	[.34]	10.4
	(2, 0)	.22	.32	−.39	.11	[.29]	10.9
B	(0, 1)	.35	.50	−.40	.09	[.36]	5.2
	(0, 2)	.22	.32	−.39	.06	[.31]	5.2
	(1, 0)	.54	.30	.72	[.77]	.04	11.6
	(2, 0)	.22	.32	−.39	.08	[.30]	10.4

NOTE: Correct decision probabilities are enclosed in boxes.

ate adjusted probabilities in the simulations, the value of $\mathbf{Z}$ for each simulated patient was chosen randomly using the historical frequencies of the four prognostic groups, which were .42 for (CR duration, Age) = (Short, Old), .35 for (Short, Young), .11 for (Long, Old), and .12 for (Long, Young).

Because the probabilities for each scenario vary with $\mathbf{Z}$, to conserve space we present numerical values corresponding to the prognostic group with short CR duration and younger age, $\mathbf{Z} = (0, 1)$, because this is a reasonably representative subgroup. For each scenario, the corresponding probabilities and values of ϕ for the other three prognostic subgroups vary in a manner analogous to the estimates in Table 4. Suppressing the argument $\mathbf{Z} = (0, 1)$ in π_{kc} for brevity, the null scenario (not tabled) corresponds to $(\pi_{R1}, \pi_{D1}) = (.16, .22)$ and $(\pi_{R2}, \pi_{D2}) = (.09, .17)$; hence $\xi_R = .22$, $\xi_D = .32$, and $\phi = -.39$, regardless of treatments, based on the historical probabilities in this prognostic group. Scenario A is obtained by changing the course 2 probabilities $(\pi_{R2}(0, 1), \pi_{D2}(0, 1))$, corresponding to salvage with T_1 following a course 1 failure with T_0, from the null values (.09, .17) to (.47, .06). The result is that, in terms of the objective function ϕ, strategy $(0, 1)$ is greatly superior to the other three strategies. Scenario A is analogous to the LFC in the one-dimensional setting, although cases with $\phi(0, 1) > \phi(0, 2) = \phi(1, 0) = \phi(2, 0)$ may be obtained in various ways. Scenario B is obtained from the null scenario by changing $(\pi_{R1}(1), \pi_{D1}(1))$ from (.16, .22) to (.52, .27) and $(\pi_{R2}(0, 1), \pi_{D2}(0, 1))$ from (.09, .17) to (.30, .46). That is, under scenario B, T_1 increases the probabilities of both response and death in both courses, a phenomenon commonly encountered in testing experimental treatments for AML. In this case, $\phi(1, 0) > \phi(0, 2) = \phi(2, 0) > \phi(0, 1)$, so that strategy $(0, 1)$ is worst and $(1, 0)$ is best. Scenario C includes a treatment–covariate interaction in which T_1 is a superior salvage treatment overall, but also increases the death rate in older patients. The probabilities characterizing this scenario were obtained by parameterizing the model using the indicator Z^*_{AGE} of older age, so that larger values of $\zeta^*_{D,\text{AGE}}(1)$ correspond to higher death rates among older patients treated with T_1 in either course. We obtained the probabilities for this scenario by starting with scenario A and increasing $\zeta^*_{D,\text{AGE}}(1)$ from 0 to 3. For example, among older patients with short CR duration treated with strategy (0,1), this has the effect of changing the two-course response and death rates from $\xi_R(0, 1) = .35$ and $\xi_D(0, 1) = .37$ under scenario A to $\xi_R(0, 1) = .17$ and $\xi_D(0, 1) = .70$ under scenario C. Many other clinical scenarios may be hypothesized, and in fact we evaluated the design under a larger set of scenarios containing the three described here.

6.2 Simulation Results

The trial was simulated 4000 times under each clinical scenario. The values in Tables 5 and 6 and reported in the text are the means over these repetitions. Each simulated dataset was fit via maximum likelihood using the full 30-parameter model specified by (5)–(9) for $s = 0, 1$, or 2 and the four strategies $(s, t) = (0, 1), (0, 2), (1, 0)$, and $(2, 0)$. The Bayesian decision criteria used in each simulated trial were computed using the approximate method described in Section 2.3. The sample size of 96 patients used throughout was chosen to obtain a correct selection probability $\geq .75$ in the (Short CR duration, Younger age) prognostic subgroup under scenario A.

The OCs in Table 5 indicate that under each of a reasonable set of possible clinical scenarios, the design has a good probability of correctly selecting the best two-course treatment strategy. The tabled correct selection probabilities are substantial improvements over the probability (.25) of guessing the best strategy in the absence of empirical evidence. Unfortunately, this practice is quite common in clinical settings where several treatment strategies are available and one strategy must be selected. The numerical results should be interpreted in terms of the numerical values of the probabilities that characterize each scenario and the fact that, of the 96 patients in the trial, on average the sample sizes in the subgroups are only 39.2 in (Short, Old), 34.4 in (Short, Young), 11.2 in (Long, Old), and 11.2 in (Long, Old).

The variation in the selection probabilities of the three inferior strategies under scenario A, from .04 to .11, is due to the facts that the course 2 sample sizes are not fixed. Rather, they depend on the numbers of failures in each course 1 treatment group and the imbalance in the course 2 randomization. In the (Short, Young) subgroup, on average $(1/3) \times 34.4 = 11.5$ patients are randomized to each of the three treatments in course 1. Because all three treatments have the same course 1 failure rate $\pi_{F1}(s, (0, 1)) = .62$ in the (Short, Young) subgroup under scenario A, this yields about $\pi_{F1}(0, (0, 1))11.5 = 7.1$ patients who fail in course 1 with T_0 and are randomized equally between T_1 and T_2 in course 2. Hence about 3.6 patients receive each of strategies $(0, 1)$ and $(0, 2)$. In contrast, on average 7.1 patients receive strategy $(1, 0)$ and 7.1 receive strategy $(2, 0)$. This also explains why on average fewer patients receive the best strategy $(0, 1)$ than receive either $(1, 0)$ or $(2, 0)$ under scenario A, which otherwise may seem counterintuitive.

Table 6 summarizes the results under scenario C, illustrating the design's ability to select the best strategy within each prognostic subgroup. Because treatment 1 has a higher death rate among older patients under this scenario, it is desirable to have a relatively low probability of selecting either strategy $(0, 1)$ or $(1, 0)$ in either of the two prognostic groups with older patients. Equivalently, it is desirable to select either $(2, 0)$ or $(0, 2)$ for older patients. This has probabilities .87 in the (Short, Old) subgroup and .76 in the (Long, Old) subgroup, and on average only 39.3 and 11.0 patients are treated in these two subgroups. These subgroup-specific selection probabilities should be compared to the value (.50) that would be obtained by guessing. The much smaller correct selection probability (.42) for the optimal strategy $(0, 1)$ in the (Long, Young) subgroup is due to its much smaller sample size of 11.3, although this probability is still much larger than the value (.25) obtained by guessing. The fact that the design performs well under scenario C may be attributed to the adaptive nature of the two-course treatment strategy and borrowing of strength by the parametric model across the various treatment strategy and prognostic subgroup combinations.

The interim decision rule (13) that drops inferior treatments has a very small effect on the selection probabilities, but yields a design that on average treats more patients with the superior strategies. For example, under scenario A, if the interim rule

Table 6. Operating Characteristics of the Design Under Scenario C, Where Strategy (0, 1) is Superior Overall but T_1 in Either Course Greatly Increases the Death Rate in Older Patients

Prognostic group				Decision probabilities		
CR Dur	Age	Treatment strategy	ϕ	Selected	Dropped early	# Patients
Short	Old	(0, 1)	−1.33	.13	[.38]	6.6
		(0, 2)	−.91	[.44]	.07	9.1
		(1, 0)	−2.32	.00	[.89]	7.6
		(2, 0)	−.91	[.43]	.16	16.1
Short	Young	(0, 1)	.55	[.73]	.06	7.4
		(0, 2)	−.39	.06	[.30]	5.3
		(1, 0)	−.39	.10	[.28]	11.0
		(2, 0)	−.39	.12	[.30]	10.6
Long	Old	(0, 1)	−.17	.24	[.30]	2.2
		(0, 2)	−.03	[.35]	.17	2.5
		(1, 0)	−2.21	.00	[.86]	2.2
		(2, 0)	−.03	[.42]	.27	4.2
Long	Young	(0, 1)	1.48	[.42]	.19	2.0
		(0, 2)	.73	.10	[.36]	1.7
		(1, 0)	.73	.24	[.28]	3.8
		(2, 0)	.73	.24	[.27]	4.7

NOTE: Correct decision probabilities are enclosed in boxes.

is not used, then the total number of patients in all prognostic groups treated with the best strategy (0, 1) drops from 21.1 to 15.4, so that about 6 more patients among the 96 receive the best treatment strategy due to interim monitoring. The effect of interim monitoring under scenario C is greater, with on average 8.7 (41.4 − 32.7) more patients among the 96 receiving one of the best strategies in their prognostic group due to the interim monitoring rule. Because dropping this rule corresponds to using an upper probability cutoff of 1 in (13), the question arises as to whether this cutoff may be calibrated to improve the design's OCs. We thus repeated the simulations summarized in Tables 5 and 6 using cutoffs .90 and .99. As the cutoff is increased over this range, the design's overall safety drops, but there is no clear pattern in its effect on the selection probabilities. It appears that the design's safety and selection probabilities may depend in a complex way on both the cutoff and the parameterization of each scenario.

The underlying probability model includes parameters characterizing not only treatments, courses, and covariate effects, but also all pairwise interactions between these three factors. A much simpler version of the model containing only main effects is given by $\eta_{k1}(s, \mathbf{Z}) = \mu_k + \alpha_k(s) + \sum_{j=1}^{q} \gamma_{kj} Z_j$ and $\eta_{k2}(s, t, \mathbf{Z}) = \eta_{k1}(t, \mathbf{Z}) + \beta_k$. Although this model's comparative simplicity may seem appealing, its use results in greatly degraded OCs. For example, the probability of correctly selecting the optimal strategy (0, 1) for (Short, Young) patients under scenario A decreases from .76 under the full model to .24 under the simpler model, whereas the respective probabilities of dropping the three inferior strategies decrease from .32, .34, and .29 (Table 5) to .12, .05, and .18. A similar question is what may result from basing the design on the empirical probabilities of response and death rather than using model-based estimates. This empirical approach reduces the correct selection probability by about .10 under each of the three scenarios. This is as expected, because the regression model borrows strength across prognostic subgroups and courses, whereas the purely empirical approach does not.

Recall that, in developing a trial design as described in Section 3, we used the historical data only to provide the marginal posterior $f(\boldsymbol{\theta}_B \mid \chi_H)$ of the baseline, nontreatment-related parameters. Repeating the simulations under models other than model 4 in Table 2 showed that the operating characteristics of the AML trial design were relatively insensitive to which model was chosen, apparently because $f(\boldsymbol{\theta}_B \mid \boldsymbol{\chi}_H)$ changed very little between these models. For example, under scenario A, using either model 1 or model 3 yielded selection probabilities all within .019 and early dropping probabilities all within .024 of the corresponding values for model 4 given in Table 5, with most of the probabilities identical to two decimal places and no systematic variation. Similarly, the number of patients treated in each course were all within .16 of the corresponding values for model 4. These differences appear to be due mainly to simulation variability.

To check the approximate Bayesian method, we recomputed the posteriors under several models in Table 2 using MCMC (Gilks, Richardson, and Spiegelhalter 1996). Each MCMC

computation was based on 100,000 runs with a burn-in sample of 10,000. The two methods gave similar posteriors, with a few large differences for parameters with a posterior mean very small relative to its standard deviation, that is, with marginal posterior centered around 0 and very disperse. Under model 4, the posterior approximate mean(std) of $\zeta_{R,\mathrm{DUR}}(1)$ was $-.28(.71)$, compared to $-.16(.72)$ using MCMC; the approximate mean(std) of $\gamma_{D,\mathrm{DUR}}$ was $-.01(.31)$, compared to $-.02(.29)$ using MCMC; and the approximate mean(std) of $\delta_{R,\mathrm{AGE}}$ was .08(.64) compared to .10(.48) using MCMC. The posterior means of the remaining 15 parameters differed by $< 8\%$, with each difference well within the posterior standard deviation.

7. DISCUSSION

We have proposed a method for the design and conduct of clinical trials in which patient outcome in each of two courses is trinomial, including the possibilities of both a desirable clinical outcome and death, and the scientific goal is to select a best two-course treatment strategy within each of several patient prognostic subgroups. The methodology is based on a generalized logistic model accounting for courses, multiple treatment strategies, and patient covariates. The method may be adapted to a wide variety of clinical settings, because its main requirements are that patient outcome can be observed relatively soon after the start of treatment and characterized by a trinomial variable. Our simulation results in the context of our motivating application indicate that, compared to more conventional selection trials, the design has attractive properties under a wide range of clinical scenarios.

We have used an approximate Bayesian method to compute the probability criteria underlying the interim monitoring rule and final selection. Given that we evaluate the design in terms of frequentist OCs and that we do not use decision theory, a natural question is whether the Bayesian formulation is needed at all. Although we consider the Bayesian formulation to be more natural, the design could be implemented by substituting a frequentist rule for the Bayesian criterion (13) and selecting treatment strategies using the MLEs $\hat{\phi}(s,t,\mathbf{Z})$. For example, the criterion for dropping an inferior strategy (u,v) could be that

$$\hat{\phi}(s,t,\mathbf{Z}) - \hat{\phi}(u,v,\mathbf{Z}) > \hat{\sigma}((s,t),(u,v),\mathbf{Z})z^*$$

for some (s,t), where $\hat{\sigma}((s,t),(u,v),\mathbf{Z})$ is an estimate of the standard deviation of the foregoing difference and z^* is chosen to control a given overall error rate among the pairwise comparisons. However, specifying the cutoff z^* in this way is essentially the same as adjusting the probability cutoff .95 used in (13), and, moreover, the ordering of the posterior means of the distinct strategies of $\phi(s,t,\mathbf{Z})$ is virtually identical to that obtained from their MLEs. Thus in practice, the two approaches should yield designs with similar properties.

Numerous extensions and modifications of the design described here are possible. Lavori and Dawson (2000) proposed a biased-coin within-subject adaptive randomization method to compare multicourse treatment strategies. A simple generalization of the AML trial design is to allow for more than two courses. This may be motivated by, for example, a trial of multiple treatments for a life-threatening infection, with the trinary outcome {alive and not infected, alive and infected, dead} in each course. In such settings, a patient may be treated until either the infection is resolved, the patient dies, or death is nearly certain regardless of additional treatment. A more complicated extension in the context of AML therapy would be to follow patients who achieve CR for an additional time period and record whether the patient is still in CR, has relapsed, or has died, with patients who relapse randomized among two or more salvage therapies. The set of possible outcomes would be more complex, because each CR is now partitioned into three subevents, and there are more treatments. If "patient success" were defined as the patient achieving CR in a given course and remaining in CR for the subsequent period, then this would be similar to the definition of patient success used in the prostate cancer trial described by Thall, Millikan, and Sung (2000). A very different type of extension would use the times to the events rather than discretizing them. This would require a multivariate event time model in place of the generalized logistic model, possibly treating the times to response and failure as nonfatal competing risks, with the distribution of subsequent survival time depending on whether response or failure has occurred. Using event times could potentially provide a more informed evaluation of treatment strategies, especially because the time to achieve response has a profound effect on subsequent survival time in AML. Such a design also could account for relapse after response and the salvage therapy administered at relapse. Practical implementation would require addressing the issues of model complexity, the logistics of continuously monitoring multiple event times, and sample size.

An important question is whether Bayesian decision theory may yield a design with better properties. Such an approach could be based on the use of ϕ as a utility function, or possibly a more complex utility that also accounts for costs, as in Stallard, Thall, and Whitehead (1999). Because such an approach is very different from that taken here, it is a topic for future research.

[Received July 2000. Revised August 2001.]

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Bryant, J., and Day, R. (1995), "Incorporating Toxicity Considerations Into the Design of Two-Stage Phase II Clinical Trials," *Biometrics*, 51, 1372–1383.

Dixon, D. O., and Simon, R. (1991), "Bayesian Subset Analysis," *Biometrics*, 47, 871–881.

Faraggi, D., and Simon, R. (1998), "Bayesian Variable Selection Method for Censored Survival Data," *Biometrics*, 54, 1475–1485.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977), *Selecting and Ordering Populations: A New Statistical Methodology*, New York: Wiley.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Jennison, C., and Turnbull, B. W. (1993), "Group Sequential Tests for Bivariate Response: Interim Analyses of Clinical Trials With Both Efficacy and Safety Endpoints," *Biometrics*, 49, 741–752.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lavori, P. W., and Dawson, R. (2000), "A Design for Testing Clinical Strategies: Biased Adaptive Within-subject Randomization," *Journal of Royal Statistical Society*, Ser. A, 163(1), 29–38.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayesian Estimates for the Linear Model" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 1–41.

Pocock, S. J., and Simon, R. (1975), "Sequential Treatment Assignment With Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–115.

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., and Cheng, S.-C. (1999), "Treatment Comparisons Based on Two-Dimensional Safety and Efficacy Alternatives in Oncology Trials," *Biometrics*, 55, 746–753.

Thall, P. F., Millikan, R. E., and Sung, H. G. (2000), "Evaluating Multiple Treatment Courses in Clinical Trials," *Statistics in Medicine*, 19, 1011–1028.

Willan, A. R., and Pater, J. L. (1985), "Hypothesis Testing and Sample Size for Bivariate Binomial Response in the Comparison of Two Groups," *Journal of Chronic Diseases*, 38, 603–608.