A Strategy for Dose-Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials

Peter F. Thall; Kathy E. Russell

Biometrics, Vol. 54, No. 1 (Mar., 1998), 251-264.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199803%2954%3A1%3C251%3AASFDAS%3E2.0.CO%3B2-J

Biometrics is currently published by International Biometric Society.

BIOMETRICS 54, 251–264
March 1998

A Strategy for Dose-Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials

Peter F. Thall* **and Kathy E. Russell**

Department of Biomathematics, Box 237,
M. D. Anderson Cancer Center, University of Texas,
1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

SUMMARY

We propose a design strategy for single-arm clinical trials in which the goals are to find a dose of an experimental treatment satisfying both safety and efficacy requirements, treat a sufficiently large number of patients to estimate the rates of these events at the selected dose with a given reliability, and stop the trial early if it is likely that no dose is both safe and efficacious. Patient outcome is characterized by a trinary ordinal variable accounting for both efficacy and toxicity. Like Thall, Simon, and Estey (1995, *Statistics in Medicine* **14,** 357–379), we use Bayesian criteria to generate decision rules while relying on frequentist criteria obtained via simulation to determine a design parameterization with good operating characteristics. The strategy is illustrated by application to a bone marrow transplantation trial for hematologic malignancies and a trial of a biologic agent for malignant melanoma.

1. Introduction

We propose a design for conducting single-arm clinical trials in which the goals are (1) to find a dose of an experimental treatment that satisfies specific safety and efficacy requirements, (2) to stop the trial early if it is likely that no dose is both safe and efficacious, and otherwise (3) to treat a sufficiently large number of patients to estimate the rates of these events at the selected dose with a given level of reliability. Patient outcome is characterized by a trinary ordinal variable accounting for both efficacy and toxicity. The proposed design involves both dose-finding and evaluation of safety and efficacy; hence, it may be regarded as a combination phase I/II. Virtually all existing phase I designs (Storer, 1989; O'Quigley, Pepe, and Fisher, 1990; Faries, 1991; Korn et al., 1994; Møller, 1995; Goodman, Zahurak, and Piantadosi, 1995; de Moor et al., unpublished manuscript; O'Quigley and Shen, 1996) rely only on toxicity while making the implicit assumption that higher doses are associated with higher response rates. In contrast, we make explicit use of both efficacy and adverse outcomes for dose-finding. Our design also is similar to the designs proposed by Gooley et al. (1994) in that we use two dose–response curves rather than one and simulation is an intrinsic part of the design process.

Our general approach is to first generate decision rules using Bayesian criteria and then evaluate the operating characteristics of the design so obtained via simulation. We then calibrate the design parameters and repeat this process until a design with good operating characteristics is obtained. This approach has been used recently by a number of authors. These include Chevret (1993), Korn et al. (1994), Goodman et al. (1995), O'Quigley and Shen (1996), and others in application of the continuous reassessment method (CRM) and its modifications and Thall, Simon, and Estey (1995, 1996) for monitoring multiple outcomes in phase II trials. Other authors who examine frequentist

[*] *Corresponding author's email address:* rex@odin.mdacc.tmc.edu

Key words: Bayesian inference; Clinical trial; Dose-finding studies; Phase I trial; Phase II trial; Safety monitoring.

properties of Bayesian decision rules include Freedman and Spiegelhalter (1989), Ho (1991), and Rosner and Berry (1995).

We present the design strategy in the context of a bone marrow transplantation (BMT) trial, where the goals are to induce moderate but not severe graft-versus-host disease (GVHD) in order to induce an accompanying graft-versus-disease effect while controlling severe toxicity. Subsequently, we use the strategy to develop a design for an apparently very different trial of the biologic agent interleuken 12 (IL-12) in malignant melanoma to illustrate the method's generality and potential breadth of application.

The remainder of the paper is organized as follows. The underlying probability model is presented in Section 2. The decision criteria and design are given in Section 3, followed by an account of computational considerations in Section 4. In Sections 5 and 6, we apply the strategy to design the BMT trial and the IL-12 trial, respectively. We discuss alternative models and robustness in Section 7 and conclude with a discussion in Section 8.

2. Dose–Response Model

In a clinical trial of patients with advanced hematologic malignancies or lymphoma conducted at the M. D. Anderson Cancer Center, the treatment strategy consisted of autologous BMT followed by administration of an experimental immunosuppressive agent plus gamma-interferon for 1 month post transplant, then abrupt withdrawal of treatment. This regimen is known to induce GVHD, which, if controlled at a moderate but not severe level, is thought to increase long-term remission duration due to a graft-versus-disease effect associated with GVHD. Nonfatal GVHD is defined to be either moderate or severe depending on whether it can or cannot be controlled by administration of steroids, with fatal GVHD defined as severe in any case. Thus, the three clinically relevant levels of GVHD are none, moderate, and severe, and these are considered to be, respectively, inefficacious, efficacious, and adverse. Patients also may suffer severe conventional toxicity, defined as grade 3 or higher, which includes regimen-related death.

Let T be the binary 0/1 indicator of severe toxicity, and let $G = 0$, 1, or 2 indicate no, moderate, or severe GVHD, respectively. We index the three outcomes given in the right-hand side of Figure 1 by the variable Y, defined by $[Y = 0] = [G = 0 \text{ and } T = 0]$, $[Y = 1] = [G = 1 \text{ and } T = 0]$, and $[Y = 2] = [G = 2 \text{ or } T = 1]$. Equivalently, we define the adverse outcome $[Y = 2]$ as severe GVHD or severe toxicity since either of these events is highly undesirable and the desired efficacy outcome $[Y = 1]$ as moderate GVHD without severe toxicity. The outcome $[Y = 0]$ occurs if the patient has neither severe toxicity nor any GVHD. In particular, Y is an index of the combined severity of G and T since $[Y \geq 1] = [G \geq 1 \text{ or } T = 1] \supset [G = 2 \text{ or } T = 1] = [Y = 2]$. We thus reduce the six elementary outcomes determined by the 3×2 cross-product of [GVHD level] × [toxicity] to these three clinically relevant events, as illustrated by Figure 1. Denote the probabilities of these three possible outcomes for a patient administered dose d by $\theta_j(d) = Pr[Y = j \mid \text{dose} = d]$, $j = 0, 1, 2,$

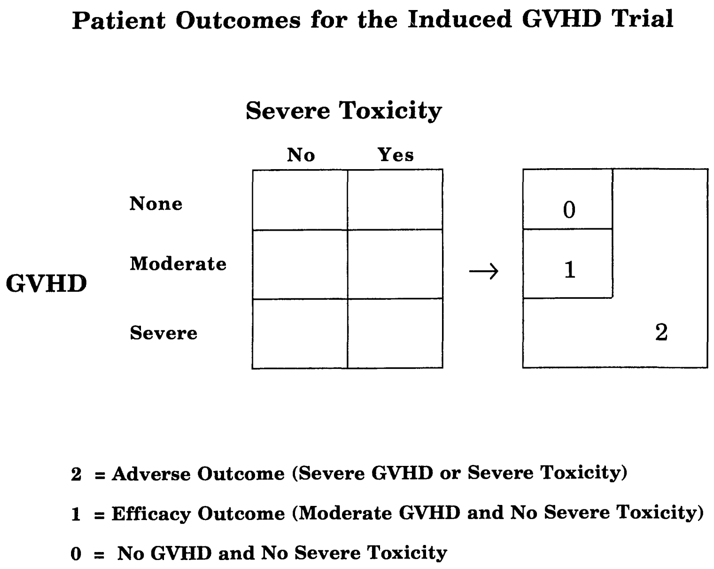
Patient Outcomes for the Induced GVHD Trial
Severe Toxicity
No
Yes
None
Moderate
Severe
GVHD
→
0
1
2
2 = Adverse Outcome (Severe GVHD or Severe Toxicity)
1 = Efficacy Outcome (Moderate GVHD and No Severe Toxicity)
0 = No GVHD and No Severe Toxicity

Figure 1. Patient outcomes for the induced GVHD trial.

with $\boldsymbol{\theta}(d) = (\theta_1(d), \theta_2(d))$. The first two goals of the trial are to find a dose d^* of the agent from among $\{2.5, 7.5, 12.5\}$ ng/ml for which, with reasonably high posterior probability, $\theta_1(d^*) \geq .50$ and $\theta_2(d^*) \leq .10$ and to stop the trial early if no dose among the three satisfies both criteria. The essential clinical difficulty is that moderate GVHD is the desired efficacy outcome while severe GVHD, in addition to severe toxicity, is an adverse outcome.

We use the following dose–response model because it is sufficiently flexible to provide a realistic representation of the actual, unknown dose–response functions. We parameterize it parsimoniously in order to compute reasonably informative posteriors based on very little data early in the trial and to allow simulation of the trial within a reasonable time frame during the design stage. Denote $\gamma_j(d) = Pr[Y \geq j \mid \text{dose} = d]$, $j = 0, 1, 2$, so that $\gamma_0(d) \equiv 1$, $\gamma_1(d) = \theta_1(d) + \theta_2(d)$, and $\gamma_2(d) = \theta_2(d)$. Since it is known that the severity of either toxicity or GVHD increases with dose of the immunosuppressive agent, we require that $\theta_2(d)$ increases (↑) and $\theta_0(d)$ decreases (↓) with d. To obtain a dose–response model with these properties, we apply McCullagh's (1980) proportional odds regression model, also known as the cumulative odds model. Writing $\text{logit}(p) = \log\{p/(1-p)\}$, our model is given by

$$\begin{aligned} \text{logit}\{\gamma_1(d)\} = \eta_1(d) &= \mu + \alpha + \beta d, \\ \text{logit}\{\gamma_2(d)\} = \eta_2(d) &= \mu + \beta d, \end{aligned} \tag{1}$$

with $\alpha > 0$ to ensure that $\boldsymbol{\theta}(d)$ is a probability distribution on $\{0, 1, 2\}$ and $\beta > 0$ to ensure the above monotonicity requirements. An important property of the model is that the probability of the desired efficacy outcome, $\theta_1(d) = \gamma_1(d) - \gamma_2(d)$, may be nonmonotone in dose, which reflects clinical experience with similar treatments. This proportional odds model may be specified equivalently by the more usual equation $Pr[Y \leq j] = e^{-\eta_{j+1}}/(1 + e^{-\eta_{j+1}})$, $j = 0, 1$, which is the most commonly applied version of the general model given by $Pr[Y \leq j] = F(-\eta_j)$ for a c.d.f. F and $\eta_j \downarrow$ in j.

We use a Bayesian formulation because it provides a natural framework to incorporate information as it accumulates and make decisions in real time during the trial. As the data (d, Y) become available from each successive cohort of patients, we repeatedly update the posterior distribution of the parameters (μ, α, β) and apply our decision criteria. Our approach is not fully Bayesian in the sense of Berry (1993, 1995) and Berry and Stangl (1996) since it does not rely on formal decision theory. Rather, we simulate the trial under each of several dose–response scenarios as a means to calibrate design parameters to ensure that the design has desirable operating characteristics.

We next specify prior distributions on the parameters (μ, α, β) to complete the probability model. To do this, we first define, in terms of the two dose–response curves $\theta_1(d)$ and $\theta_2(d)$ over the domain of d containing the doses used in the trial, an array of clinical scenarios encompassing what may reasonably be anticipated as the true state of nature. These are graphed in Figure 2, with the curve of $\theta_1(d)$ given by the solid line, $\theta_2(d)$ given by the dashed line, the fixed clinical standards $\theta_1^* = .50$ for efficacy and $\theta_2^* = .10$ for toxicity given by horizontal dotted lines, and acceptable doses indicated by arrows. In each of scenarios 1–4, exactly one dose is acceptable, $d = 2.5$, 2.5, 7.5, and 12.5, respectively. Scenarios 1 and 4 may be considered least favorable in that, at the acceptable dose, the adverse and efficacy outcome probabilities are at their respective limits .10 and .50 and the dose–response curves are rather flat. Scenario 2 is more optimistic than scenario 1 in that the respective toxicity and efficacy probabilities are .05 and .55 at $d = 2.5$ and the toxicity curve is steep. However, each of scenarios 1–4 is a difficult one in that the therapeutic window of acceptable doses, where $\theta_1(d) \geq .50$ and $\theta_2(d) \leq .10$, is rather small. In this regard, it important to bear in mind that only the three dose levels 2.5, 7.5, and 12.5 are considered. Thus, for example, although in theory the therapeutic window under scenario 2 is $1.84 \leq d \leq 5.52$, the only relevant value within this window is $d = 2.5$, where $p_1(2.5) = .55$ and $p_2(2.5) = .05$. In scenario 5, both $d = 2.5$ and $d = 7.5$ are acceptable. In scenario 6, none of the three doses are acceptable, but there exists an acceptable dose between 7.5 and 12.5. In scenarios 7, 8, and 9, no dose in the range 2.5 to 12.5 is acceptable due to uniformly insufficient efficacy (scenario 7), uniformly excessive adverse outcome probability (scenario 8), or a complete disaster where all three doses are too adverse and insufficiently efficacious (scenario 9). Because patients are treated only at the dose levels $d = 2.5$, 7.5, or 12.5, each of these scenarios is characterized by the six numerical values $\{(p_1(d), p_2(d)), d = 2.5, 7.5, 12.5\}$. While Figure 2 provides a graphical illustration of the sorts of dose–response curves that may take on these values at these three doses, their behavior for doses other than these three values is irrelvant with regard to the design and its operating characteristics. This would not be the case, however, for an extended version of the design allowing the addition of one or more new dose levels during the trial.

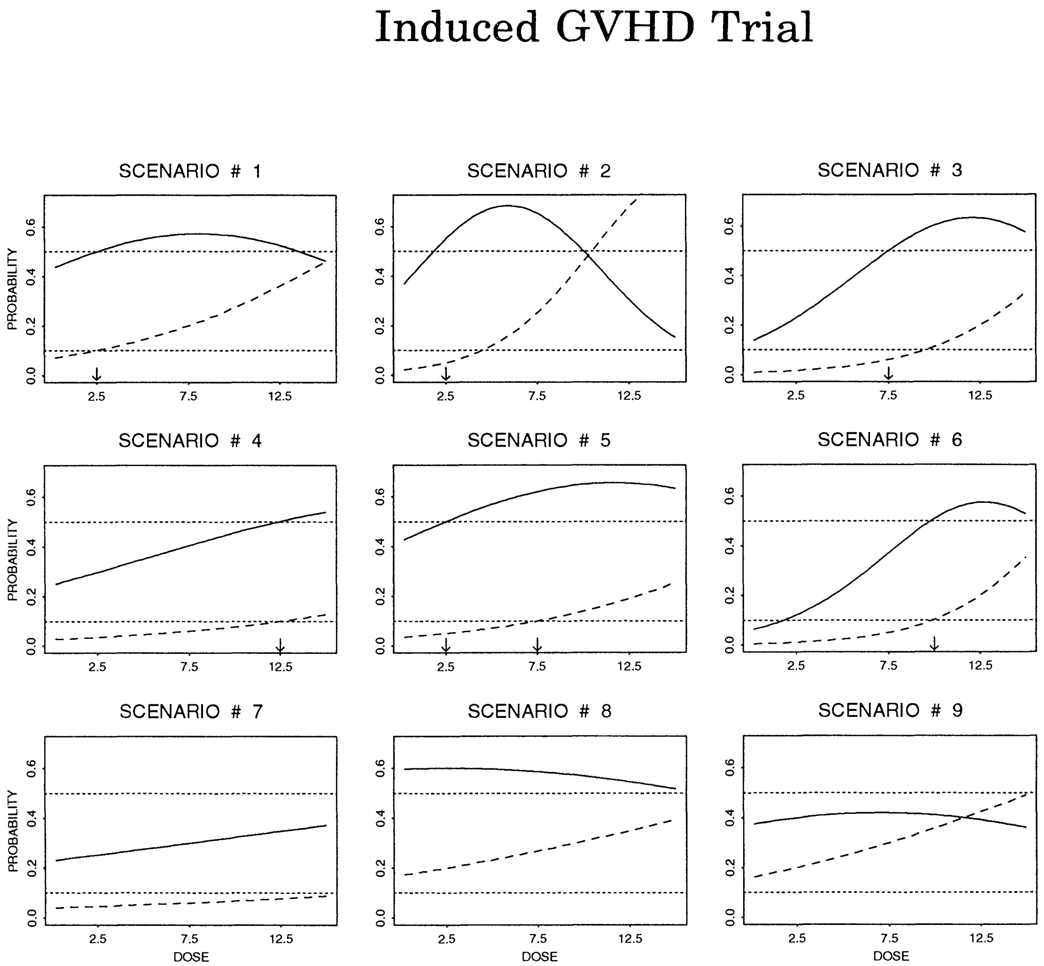

Figure 2. Dose–response scenarios for the induced GVHD trial. The solid and dashed lines denote $\theta_1(d) = Pr\,[\text{efficacy outcome} \mid d]$ and $\theta_2(d) = Pr\,[\text{adverse outcome} \mid d]$, respectively. The fixed criteria θ_1^* and θ_2^* are denoted by horizontal dotted lines. Acceptable doses under each scenario are marked by an arrow.

Denoting the mean of $\theta_j(d)$ by $p_j(d)$, we solved for the corresponding means $(\mu^o, \alpha^o, \beta^o)$ of (μ, α, β) under each scenario by equating $\text{logit}\{p_2(d)\} = \mu^o + \beta^o d$ for two values $d = x, y$ and solving for $\beta^o = [\text{logit}\{p_2(y)\} - \text{logit}\{p_2(x)\}]/(y - x)$, then $\mu^o = \text{logit}\{p_2(y)\} - \beta^o y$ and $\alpha^o = \log[p_0(x)^{-1} - 1] - \mu^o - \beta^o x$. Repeating this computation for each of the nine scenarios yields the parameters given in Table 1. Based on these numerical values and assuming that the nine scenarios to which they correspond together represent a realistic range of what may be encountered in actual conduct of the trial, we defined the domain of each parameter to encompass these and also some more extreme situations that might possibly occur. This yielded the domains $-6 \leq \mu \leq -1$, $1 \leq \alpha \leq 4$, and $.04 \leq \beta \leq .40$. We assumed independent uniform priors for the parameters on their domains, both to reflect considerable prior uncertainty and to facilitate rapid numerical computation of posteriors. Given that the numerical values of the dose–response curves at $d = 2.5$, 7.5, and 12.5 helped us determine the priors in this way, it is important to note that we initially considered a smaller number of scenarios but subsequently expanded them to the nine given in Figure 2 based on the physicians' initial reactions and suggestions.

3. The Design

3.1 *Bayesian Decision Criteria*

The primary goal of the trial is to determine if there is a dose d^* among the three doses of the experimental agent that satisfies both the efficacy and toxicity criteria. If there is such a dose, then an additional goal is to treat a sufficiently large number of patients at d^* to estimate $\boldsymbol{\theta}(d^*)$ with reasonable reliability. Safety monitoring is of paramount importance in any early phase trial,

Table 1

Fixed parameters used for simulation of the induced GVHD trial

Scenario	μ^o	α^o	β^o
1	−2.6027	2.6027	.1622
2	−3.8674	3.3499	.3692
3	−4.7994	2.9927	.2730
4	−3.5830	2.6113	.1109
5	−3.3180	3.1451	.1494
6	−5.2817	2.6217	.3116
7	−3.1673	2.1762	.0554
8	−1.5781	2.7726	.0767
9	−1.6558	1.7918	.1078

however. We thus require that the design terminate the trial early with reasonably high probability if either the lowest dose has an unacceptably high adverse outcome rate or the highest dose is not sufficiently efficacious. These goals together encompass the requirements of a usual phase I trial where dose-finding is based solely on the adverse outcome as well as those of a phase II trial where both efficacy and adverse outcomes are monitored, as in Bryant and Day (1995), Conaway and Petroni (1995, 1996), and Thall et al. (1995, 1996). The design proposed here thus may be regarded as a phase I/II hybrid since it monitors multiple outcomes and has as goals dose-finding, safety monitoring, and estimation.

To achieve the above goals, we first define the two decision criteria formally in terms of posterior probabilities given the accumulated data at any interim point in the trial. Let $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$ be fixed standards specified by the clinician. For the induced GVHD trial, $\theta_1^* = .50$ and $\theta_2^* = .10$. Given upper probability cutoffs π_1 and π_2, we consider a dose d to have unacceptably low efficacy if

$$\psi_1(d, data) = Pr[\theta_1(d) < \theta_1^* \mid data] > \pi_1 \tag{2}$$

and unacceptably high adverse outcome rate if

$$\psi_2(d, data) = Pr[\theta_2(d) > \theta_2^* \mid data] > \pi_2. \tag{3}$$

We say that a dose is acceptable if neither (2) nor (3) is the case and unacceptable if either criterion is satisfied. To determine (π_1, π_2), we simulate the design for several (π_1, π_2) pairs under each scenario and ask the clinician to choose a design parameterization based on its operating characteristics.

3.2 *Conduct of the Trial*

Given the above criteria, the trial is carried out as follows.

1. Treat patients in cohorts of size c, up to a maximum of N patients.
2. Treat the first cohort at the lowest dose level.
3. Never escalate by more than one dose level unless some patients have been treated at all intermediate dose levels.
4. If the current dose is unacceptably toxic and is
   (a) not the lowest dose level, then de-escalate one dose level.
   (b) the lowest dose level, then terminate the trial.
5. If the current dose has acceptable toxicity and unacceptably low efficacy and
   (a) the next higher dose level has acceptable toxicity, then escalate one dose level.
   (b) the next higher dose level is unacceptably toxic, then terminate the trial.
   (c) the current dose is the highest dose level, then terminate the trial.
6. If the current dose is acceptable, then treat the next cohort at the acceptable dose level having largest efficacy criterion probability $1 - \psi_1(d) = Pr[\theta_1(d) \geq \theta_1^* \mid data]$.

The first four requirements are similar to those used in the modified CRM for dose-finding in a conventional phase I trial, as described by Faries (1991), Korn et al. (1994), Møller (1995), and Goodman et al. (1995), where patient outcome is summarized in terms of a single binary indicator for the adverse event toxicity. Our design is more general than the modified CRM in that we summarize patient outcome by a trinary ordinal variable that accounts for both efficacy and adverse outcomes, and we require that the trial be terminated early if it is likely that all of the dose levels being considered are unacceptable. Consequently, our probability model, decision criteria, and the decision scheme that must be followed during conduct of the trial are more complex.

3.3 *Sample Size*

We choose the maximum sample size N to estimate the efficacy outcome probability with a given reliability if the trial is not stopped early. Under a scenario where the efficacy at the selected dose d^* equals the targeted minimum level, specifically $p_1(d^*) = \theta_1^*$, for given interval width 2δ and for each of several values of N, we determine the posterior coverage probability $Pr\,[\theta_1^* - \delta \leq \theta_1(d^*) \leq \theta_1^* + \delta \mid data_N]$. The clinician may then choose N based on its associated posterior coverage probability, along with the usual considerations of patient accrual rate, maximum feasible trial duration, monetary costs, and drug availability. This is the Bayesian analog of the common procedure of determining sample size to obtain a confidence interval of given reliability. Since the actual data based on N patients vary substantially due to the dose-finding strategy, we evaluate these posterior probabilities empirically by simulation.

4. Computing

The decision criteria $\psi(x, data) = \{\psi_1(x, data),\ \psi_2(x, data)\}$ must be evaluated for each dose x repeatedly during the trial, with updating when the data from each successive cohort become available. When simulating the design to obtain operating characteristics, this computation must be done [mean number of cohorts per trial] × [number of simulated trials] × [number of scenarios] times for each design parameterization $(\boldsymbol{\theta}^*, \boldsymbol{\pi}, N, c)$. Denote the ith patient's dose by $d_{(i)}$ and, in general, let $I[A] = 1$ if the event A occurs and 0 otherwise. The likelihood based on data $(\mathbf{Y}_n, \mathbf{d}_n)$ $= \{(Y_i,\ d_{(i)}),\ i = 1, \ldots, n\}$ is

$$\mathcal{L}(\mu, \alpha, \beta; \mathbf{Y}, \mathbf{d}) = \prod_{i=1}^{n} \prod_{j=0}^{2} \left\{\theta_j(d_{(i)}, \mu, \alpha, \beta)\right\}^{I[Y_i = j]}. \tag{4}$$

Denote the prior by $f(\mu, \alpha, \beta)$. Each decision function is the ratio of 2 three-dimensional integrals. The efficacy criterion at dose x is

$$\psi_1(x, \mathbf{Y}, \mathbf{d}) = \frac{\int I[\theta_1(x, \mu, \alpha, \beta) > \theta_1^*]\, \mathcal{L}(\mu, \alpha, \beta; \mathbf{Y}, \mathbf{d})\ f(\mu, \alpha, \beta)\, d\mu\, d\alpha\, d\beta}{\int \mathcal{L}(\mu, \alpha, \beta; \mathbf{Y}, \mathbf{d})\ f(\mu, \alpha, \beta)\, d\mu\, d\alpha\, d\beta}, \tag{5}$$

with $\psi_2(x, \mathbf{Y}, \mathbf{d})$ defined similarly. The integral in the denominator of (5) is over the three-dimensional rectangular domain of the prior, and the integral in the numerator is evaluated over the subdomain obtained by deriving the system of nonlinear inequalities in μ, α, β defined by the inequality $\theta_1(x, \mu, \alpha, \beta) > \theta_1^*$. We used the Gauss–Kronrod (GK) algorithm with 21 points (Piessens et al., 1983) to evaluate the denominator. Although application of GK to evaluate the numerator gives precise results, this proved too time-consuming for practical use in the simulations. We thus evaluated the numerator using three applications of a 10-point Gaussian formula (Abramowitz and Stegun, 1965, Section 25.4.29), which is much faster and gives results of sufficient accuracy that this coarser approximation had no effect on the actual decisions.

All simulations were based on 1000 replications. The method of L'Ecuyer and Cote (1991) was used to generate uniform random numbers. To save time, we used the dynamic programming technique of storing the first computed value of $\psi(\mathbf{x}, \mathbf{Y}_n, \mathbf{d}_n) = \{\psi(x, \mathbf{Y}_n, \mathbf{d}_n), x = 2.5,\ 7.5,\ 12.5\}$ for each cumulative data vector $\{\mathbf{Y}_n, \mathbf{d}_n\}$ and retrieving $\psi(\mathbf{x}, \mathbf{Y}_n, \mathbf{d}_n)$ from memory whenever $\{\mathbf{Y}_n, \mathbf{d}_n\}$ recurred in the course of a simulation. Each simulation of the nine scenarios took 4 to 6 hours on a DEC AlphaServer 2100 5/250 running OSF/1, depending on machine load. Computing time is not problematic during conduct of the trial, however, since each evaluation of $\psi(\mathbf{x}, \mathbf{Y}_n, \mathbf{d}_n)$ takes at most a few seconds.

5. The Induced GVHD Trial

Given $\theta_1^* = .50$ and $\theta_2^* = .10$, the remaining design parameters are the maximum sample size N, cohort size c, and cutoffs (π_1, π_2). We first studied the design's operating characteristics for $N = 30$ and 40, $c = 1, 2, \ldots, 6$ patients per cohort, and 14 different (π_1, π_2) combinations. Fixing $c = 3$ and $(\pi_1, \pi_2) = (.90, .90)$ based on these results, we then evaluated the posterior coverage probabilities

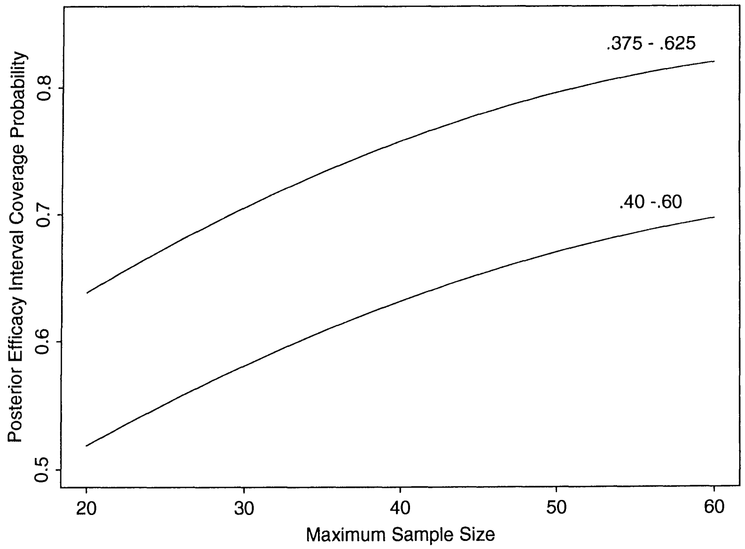

Figure 3. Posterior probabilities $Pr\,[.40 < \theta_1(d) < .60 \mid data_N]$ and $Pr\,[.375 < \theta_1(d) < .625 \mid data_N]$ as functions of maximum sample size N.

$Pr\,[.50 - \delta \leq \theta_1(2.5) \leq .50 + \delta \mid data_N]$ for $\delta = .10, .125$ and $21 \leq N \leq 60$ under scenario 2, where the true efficacy probability at $d = 2.5$ is the targeted .50. A plot of the posterior coverage probabilities is given in Figure 3. The final design chosen by the clinicians had $N = 39$. Table 2 summarizes its operating characteristics, with correct decision probabilities enclosed in boxes. The correct decision probabilities of six designs based on other (π_1, π_2) values near (.90,.90) are summarized in Table 3.

Table 2 reflects the complexity of the clinical setting and is motivated both by consideration of what sort of dose–response curves may actually be the case and by the possible decisions that may be made. In general, one would like a design that has high probabilities of picking an acceptable dose when it exists (scenarios 1–5) and of stopping the trial when no dose is acceptable (scenarios 7–9). The first seven lines of Table 2 together comprise the possible decisions, aside from the outcome that the trial reaches its maximum sample size without reaching a decision. From a frequentist point of view, one might consider the vector of probabilities of these seven decisions to be a generalization

Table 2
Operating characteristics of the induced GVHD trial design

	Scenario								
	1	2	3	4	5	6	7	8	9
Decision									
d_1 selected	$\boxed{.43}$	$\boxed{.77}$	.00	.00	$\boxed{.11}$	.00	.00	.19	.13
d_2 selected	.23	.07	$\boxed{.56}$	.05	$\boxed{.62}$	.18	.00	.02	.02
d_3 selected	.00	.00	.19	$\boxed{.60}$	.16	.22	.16	.00	.00
d_1 toxic	.16	.05	.00	.00	.02	.00	.00	$\boxed{.78}$	$\boxed{.51}$
d_1 not eff, d_2 toxic	.14	.08	.02	.02	.04	.01	.01	.01	$\boxed{.29}$
d_2 not eff, d_3 toxic	.01	.00	.15	.10	.01	$\boxed{.47}$	.04	.00	$\boxed{.04}$
d_3 not eff	.00	.00	.02	.22	.01	.07	$\boxed{.78}$	.00	$\boxed{.01}$
Sample size									
No. treated at d_1	17.0	24.8	3.3	3.7	8.0	3.1	3.5	15.1	12.7
No. treated at d_2	13.9	11.3	17.9	6.3	20.5	11.1	4.0	3.4	6.2
No. treated at d_3	1.9	0.8	14.6	22.0	9.0	15.7	13.7	0.3	1.2
Total no. patients	32.7	36.2	35.8	31.9	37.6	29.8	21.2	18.8	20.1
Toxicity	.18	.13	.12	.09	.12	.13	.06	.29	.31

of the usual power and $1 - Pr$ [Type I error] associated with a conventional test of hypothesis. In contrast with hypothesis testing, however, the goals here are not to obtain confirmatory results. Rather, the objectives are to determine whether there is an acceptable dose under this treatment strategy and, if so, to obtain reasonably reliable estimates of the patient outcome probabilities at that dose to be used as a basis for making treatment decisions and planning future trials.

Recall that $d_1 = 2.5$ is acceptable under both scenarios 1 and 2 but that the toxicity curve θ_2 is steeper under scenario 2. It appears that, under our design, this steepness provides a higher probability of determining that $d = 2.5$ is acceptable with fewer patients treated at the toxic levels $d = 7.5$ and 12.5 under scenario 2, as compared to scenario 1.

Table 3 illustrates the fact that selecting a correct dose under scenarios 1–5 and terminating the trial early under scenarios 7–9 are conflicting desiderata. This type of conflict also exists between scenarios 7 and 8 and is reflected by their early stopping probabilities. Increasing π_2 for fixed π_1 has the effect of increasing the early stopping probability under scenario 7 and decreasing it under scenario 8. Conversely, increasing π_1 for fixed π_2 greatly reduces the early stopping probability under scenario 7, with little effect under scenario 8. Additional simulations with smaller values of π_1 and π_2 (not shown) produced designs with higher early stopping probabilities under scenarios 7–9 but lower correct dose selection probabilities under scenarios 1–5. In general, smaller values of c produce higher early stopping probabilities, although the magnitude of the effect is trivial under all but two scenarios. Under scenario 4, where the highest dose 12.5 is acceptable but both dose–response curves are relatively flat, the correct selection probabilities for the cohort sizes $c = 1, \ldots, 6$ were (.46, .56, .63, .65, .69, .71), respectively. Under scenario 7, where no dose is efficacious, the corresponding early stopping probabilities were (.83, .81, .76, .73, .69, .71). These values indicate that the cohort size $c = 3$ provides a compromise between these two competing goals. The overall toxicity rates showed a small monotone decline with c, with the largest being a drop from 34% at $c = 1$ to 28% at $c = 6$ under scenario 9. In choosing cohort size, however, these results must be weighed along with the practical consideration of maximum trial duration, which increases substantially with c. While it might seem that there should be a symmetry between the designs defined by $(\pi_1, \pi_2) = (.90, .95)$ and $(.95, .90)$, this is not the case due to the fact that $(\theta_1(d), \theta_1^*)$ and $(\theta_2(d), \theta_2^*)$ are not symmetric.

6. A Biologic Agent Trial

In this section, we briefly describe a second application to illustrate the generality of the approach. This is a phase I/II trial of the biologic agent interleukin 12 (IL-12) for treatment of malignant melanoma. The goal is to find an acceptable dose from among four equally spaced IL-12 dose levels, scored $\{1, 2, 3, 4\}$ for simplicity. Here the set of possible patient outcomes is generated by the more usual 2×2 product of a binary response variable, defined as $\geq$50% tumor shrinkage, and the same binary severe toxicity outcome as before. Analogously to our approach in the BMT trial, we reduce the four elementary outcomes to the three outcomes illustrated in Figure 4. The efficacy outcome is defined to be $[Y = 1]$ = [response and no severe toxicity], the adverse outcome is $[Y = 2]$ = [severe toxicity], and the third outcome is $[Y = 0]$ = [no response and no severe toxicity]. Thus, Y is again ordinal, with $1 \equiv \gamma_0 \geq \gamma_1 = Pr$ [response or severe toxicity] $\geq Pr$ [severe toxicity] $= \gamma_2$. The trial goals, probability model, and decision rules are as before, with the substantive difference that here the fixed upper limit for the severe toxicity probability is $\theta_2^* = .33$ and the fixed lower limit for the efficacy outcome probability is $\theta_1^* = .20$. As in the BMT application, these cutoffs were specified by the clinician and are appropriate for the specific patient group and treatment regimen.

Table 3
Correct decision probabilities under six parameterizations of the induced GVHD trial design, $N = 39$ and $c = 3$

		Scenario								
π_1	π_2	1	2	3	4	5	6	7	8	9
.90	.85	.40	.72	.57	.54	.71	.52	.72	.87	.88
.90	.90	.43	.77	.56	.60	.73	.47	.78	.78	.84
.90	.95	.39	.74	.48	.69	.65	.34	.78	.66	.78
.95	.85	.46	.76	.66	.63	.74	.42	.61	.88	.86
.95	.90	.50	.80	.61	.70	.75	.38	.62	.78	.80
.95	.95	.42	.75	.50	.77	.67	.30	.66	.66	.70

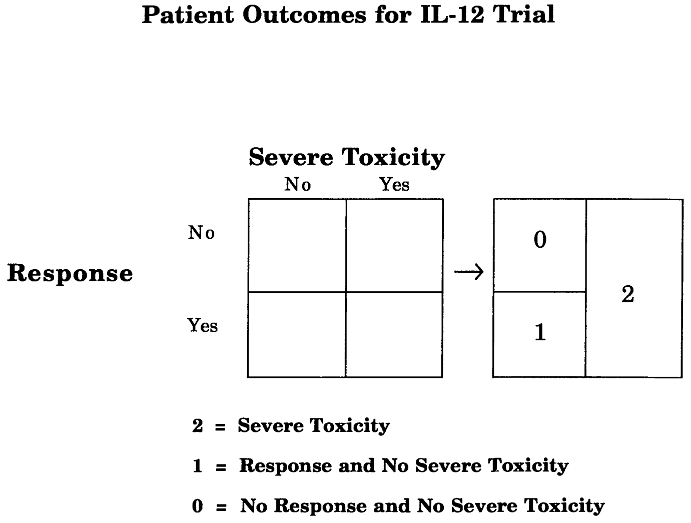
Patient Outcomes for IL-12 Trial
Severe Toxicity
No
Yes
Response
No
Yes
0
1
2
2 = Severe Toxicity
1 = Response and No Severe Toxicity
0 = No Response and No Severe Toxicity

Figure 4. Patient outcomes for the IL-12 trial.

Figure 5 illustrates the nine scenarios under which we evaluated the design's operating characteristics for the IL-12 trial, with $\theta_2(d)$ and $\theta_1(d)$ given, respectively, by the dashed and solid lines in each graph, as in Figure 2. The dose–response curves in Figure 4 appear very different from those of Figure 2, essentially because here $\theta_2^* > \theta_1^*$, while this inequality is reversed in the previous application. In each of the first four scenarios, exactly one dose is acceptable, $d = 1$, 2, 3, and 4, respectively, while either $d = 1$ or 2 is acceptable in scenario 5. Under scenario 6, there is an acceptable dose only between $d = 2$ and 3. Scenarios 7, 8, and 9 are as in the BMT trial in that all four doses are inefficacious, too toxic, or both, respectively. Table 4 summarizes a simulation study of the design for $N = 30$ and 45 and five (π_1, π_2) combinations, analogous to the study of the BMT trial. While the correct dose selection probabilities under scenarios 1–5 are relatively insensitive to N, the early stopping probabilities under scenarios 7–9 increase markedly as N is increased from 30 to 45. Based on these results along with cost and accrual considerations, the clinician selected the design with $N = 45$ and $(\pi_1, \pi_2) = (.90, .90)$, for which on average *a posteriori* $Pr[.075 \leq \theta_1(2) \leq .325 \mid \mathbf{Y}_{45}, \mathbf{d}_{45}] = .82$ under scenario 2.

7. Robustness

Our method is based on a very parsimonious parameterization of the proportional odds model. For K dose levels, the $2K$ probabilities $\{(\theta_1(d_j), \theta_2(d_j)), j = 1, \ldots, K\}$ are characterized by three parameters. Two natural questions are how well the design performs when the proportional odds assumption does not hold and how high a price is paid by use of such a parsimonious model. The overriding point with regard to robustness is that our aim is not to estimate the functions $\theta_1(d)$ and $\theta_2(d)$ over the domain of d but rather to select a dose from the set $\{d_1, \ldots, d_K\}$ that is both safe and efficacious. This is similar to the goals of the CRM in phase I where only toxicity is considered and the aim is to select a dose having mean toxicity probability closest to a given fixed standard (O'Quigley et al., 1990). Because the data in the type of trial considered here are very expensive in terms of human life, the algorithm for selecting the dose of each successive cohort must perform well early in the trial when very little data are available to update the posterior. This strongly motivates the use of a model with as few parameters as possible that, through its dose–response functions, enables each updated parameter distribution incorporating new data from the cohort treated at the most recent dose to provide new information about all the dose levels. Thus, the situation is quite different from that in which a model is fit to data *ex post facto*. We regard the regression model used here as a device to provide reliable real-time decision-making based on very small amounts of very expensive data. In this regard, our situation is similar to the phase I setting where a single binary outcome is observed on each patient. We employ our three-parameter model here in a manner analogous to that in which O'Quigley et al. (1990) employ their one-parameter model to implement the CRM in a phase I trial.

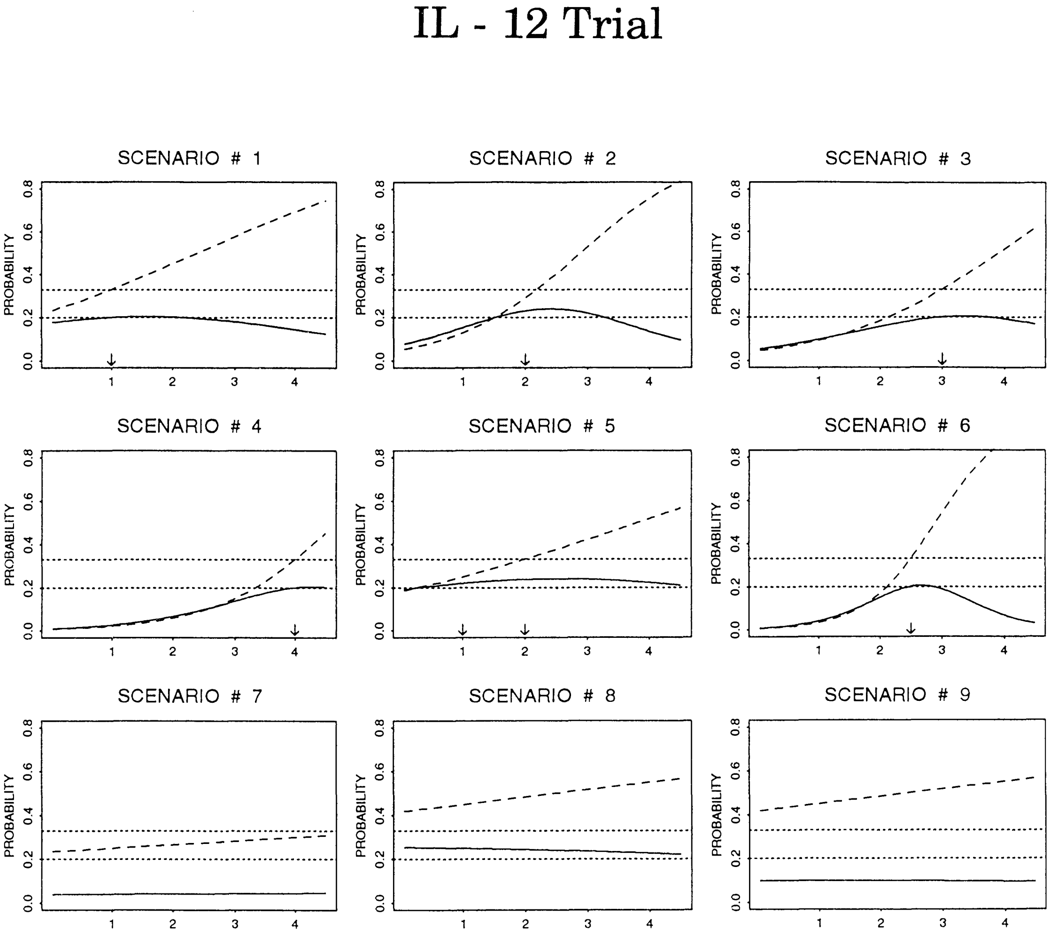

Figure 5. Dose–response scenarios for the IL-12 Trial. The solid and dashed lines denote $\theta_1(d) = Pr$ [efficacy outcome $\mid d$] and $\theta_2(d) = Pr$ [adverse outcome $\mid d$], respectively. The fixed criteria θ_1^* and θ_2^* are denoted by horizontal dotted lines. Acceptable doses under each scenario are marked by an arrow.

The design's operating characteristics depend on both the design and the fixed probabilities $\{(p_1(d_j), p_2(d_j)), j = 1, \ldots, K\}$ that characterize each scenario. Aside from the formal requirement that $0 \leq p_1(d_j) + p_2(d_j) \leq 1$, these values need not correspond to any parametric dose–response function, and any reasonable probabilities may be used. We generated these values using the proportional odds model as a convenience to allow graphical illustration of the scenarios. More importantly, the scenarios considered for each trial cover a very broad array of possibilities that the clinicians felt encompassed what might reasonably be the true state of nature. Our simulation studies indicate that the design behaves surprisingly well, on average, under rather difficult circumstances. It is also worthwhile to consider its early performance based on only the first cohort of three patients in a single trial. Table 5 presents the set of all possible cases and corresponding decisions made by the design for the induced GVHD trial. These decisions appear very sensible, and we have found the simple information in Table 5 to be very useful when explaining the design's properties to physicians.

This is not to say that other model formulations are not reasonable or that a four-parameter model is not feasible. There are several possible alternative models, starting with those obtained by simply replacing the link function logit$(\cdot)$ in expression (1) with $F^{-1}(\cdot)$ for any c.d.f. F. Two common alternatives are the standard normal c.d.f. and $F(\eta) = 1 - \exp\{-\exp(\eta)\}$. Rather than changing the link function, one may obtain an alternative model by replacing the proportional odds assumption. One possibility is the continuation ratio model, defined here by logit$\{c_j(d)\} = \eta_j$, where $c_j = Pr[Y > j \mid Y \geq j,\ d]$, $j = 0, 1$. Writing $\eta_j(d) = \alpha_j + \beta d$ for $j = 0, 1$, in this case, the efficacy and toxicity probabilities are given by $\theta_1(d) = c_0(d)\{1 - c_1(d)\}$ and $\theta_2(d) = c_0(d)c_1(d)$,

Table 4

Correct decision probabilities under 10 *parameterizations of the IL*-12 *trial design, c* = 3

		Scenario								
π_1	π_2	1	2	3	4	5	6	7	8	9
					N = **30**					
.85	.85	.44	.62	.46	.42	.66	.33	.65	.71	.85
.85	.90	.46	.65	.46	.43	.74	.26	.71	.61	.79
.90	.85	.46	.66	.49	.50	.71	.25	.59	.71	.80
.90	.90	.51	.71	.50	.51	.74	.21	.65	.62	.76
.95	.90	.57	.76	.55	.60	.79	.12	.52	.61	.69
					N = **45**					
.85	.85	.44	.67	.44	.39	.64	.51	.70	.78	.93
.85	.90	.48	.71	.46	.45	.70	.46	.77	.70	.92
.90	.85	.48	.72	.46	.50	.69	.43	.66	.80	.82
.90	.90	.53	.76	.51	.53	.75	.41	.75	.74	.89
.95	.90	.57	.78	.56	.59	.81	.25	.68	.73	.81

and the model is parameterized by $(\alpha_1, \alpha_2, \beta)$. Another alternative is the stereotype model where $\theta_j(d) = \exp(\alpha_j + \beta_j d)/\Sigma_m \exp(\alpha_m + \beta_m d)$, with β_j possibly replaced by βs_j for fixed or estimated scores $s_1, \ldots, s_K$ (Anderson, 1984). A review of regression models for ordinal outcomes is given by Greenland (1994). Perhaps the most important generalization to study next would be to extend (1) to allow different coefficients of d in each linear term, specifically define $\text{logit}\{\gamma_1(d)\} = \mu + \alpha + \beta_1 d$ and $\text{logit}\{\gamma_2(d)\} = \mu + \beta_2 d$. Since this would impose the constraint $\alpha + (\beta_1 - \beta_2)d > 0$ for all d, the Anderson score model might provide a slightly more tractable four-parameter model with heterogeneous coefficients on d.

An empirical study of the design's robustness to the proportional odds assumption could be carried out in two ways. The first would require comparative evaluation of a given set of scenarios under each of several alternative models. The two main features to study would be the type of model and the number of parameters. Although such a study is beyond the scope of the present paper, one may anticipate that designs based on models having $\theta_1(d)$ and $\theta_2(d)$ with parameterizations that can conform closely to the values $\{(p_1(d_j), p_2(d_j)),\ j = 1, \ldots, K\}$ characterizing a given scenario will perform well under that scenario but will perform less well when this is not the case. The second way to evaluate robustness would be to study the design as defined under the three-parameter proportional odds model but generate scenarios using different models. This type of study seems less useful, however, because the nine scenarios studied here already encompass a very broad array of possible dose response scenarios. In this regard, it is interesting that the dose–response curves (not shown) corresponding to those in Figure 2 but generated under the three-parameter continuation ratio model appear nearly identical to those under proportional odds. It thus seems

Table 5

Decisions based on the first cohort of the induced GVHD trial

No toxicity, no GVHD ($Y = 0$)	No toxicity, moderate GVHD ($Y = 1$)	Toxicity or severe GVHD ($Y = 2$)	Decision for second cohort
0	0	3	Stop the trial
0	1	2	Stop the trial
1	0	2	Stay at d_1
0	2	1	Stay at d_1
0	3	0	Stay at d_1
1	1	1	Stay at d_1
1	2	0	Escalate to d_2
2	1	0	Escalate to d_2
2	0	1	Escalate to d_2
3	0	0	Escalate to d_2

that the number of parameters may have a greater effect than model choice, although a detailed simulation study is required to provide definitive answers.

8. Discussion

The decision criteria (2) and (3) are similar to those employed by Thall et al. (1995, 1996) for early stopping in phase II trials with multiple outcomes. Like Thall et al., we monitor both efficacy and adverse outcomes, although the outcome set and probability model here are more specialized than their Dirichlet multinomial model. Their strategy deals with a fixed experimental treatment, however, while the design and model given here accommodate dose-changing during the trial. A subtle but important difference is that the standards θ_1^* and θ_2^* used here in the definitions of ψ_1 and ψ_2 are fixed, whereas Thall et al. use random probabilities $\boldsymbol{\theta}^S$ of the corresponding outcomes under a prior, elicited from the clinician or based on historical data, corresponding to an established standard treatment. If such a prior were available in the present context, then θ_1^* and θ_2^* could be replaced by the corresponding random values obtained from $\boldsymbol{\theta}^S$. This would require four- rather than three-dimensional numerical integrations to compute ψ_1 and ψ_2, however, and hence would increase computing time. The randomness in $\boldsymbol{\theta}^S$ also would affect the design's operating characteristics, with a decrease in the correct decision probabilities as the variability in $\boldsymbol{\theta}^S$ increases.

The method used to combine the ordinal variable G with the binary T for the induced GVHD trial may be generalized in various ways. For example, two ordinal adverse outcome variables Z_1, Z_2 and a binary 0/1 efficacy variable R may be combined into a single ordinal variable Y as follows, provided that $[R = 1]$ is irrelevant if either adverse outcome occurs. If $Z_1 = 0(1)k_1$ and $Z_2 = 0(1)k_2$, say, then first define $Z^* = j$ if $i_{a,j-1} < Z_a \leq i_{a,j}$ for $a = 1, 2$, where the grouping indices $0 = i_{a,0} < i_{a,1} < \cdots < i_{a,m_a}$, $a = 1, 2$, are chosen so that Z^* accounts for all clinically relevant adverse outcomes. Defining $[Y = 0] = [Z^* = 0, R = 0]$, $[Y = 1] = [Z^* = 0, R = 1]$, and $[Y = j] = [Z^* = j]$ for $j \geq 1$, the proportional odds model $\text{logit}\{Pr[Y \geq j \mid d]\} = \mu_j + \beta d$, with $\mu_j \downarrow$ in j, accounts for all three outcomes much more parsimoniously than a trivariate model for (Z_1, Z_2, R).

Gooley et al. (1994) proposed a design for a BMT trial where dose-finding was based on two adverse outcomes, one ↑ and the other ↓ with T-cell dose. They considered three dose–response scenarios in which the windows of acceptable doses were wide, narrow, and nonexistent, respectively, and proposed three designs based on non-Bayesian criteria. Similarly to our approach, they used simulation to evaluate the operating characteristics of each design under each scenario.

Our design assumes a single patient prognostic group for whom it is appropriate to do both dose-finding and efficacy evaluation at the selected dose. This is different from phase I/II scenarios where the phase I group has very poor prognosis while the subsequent phase II trial is conducted in patients with higher prognostic level. An extension of our design might begin dose-finding with the poorest prognosis patients and subsequently include better prognosis patients, as in a typical phase I/II setting, while accounting for prognostic level with additional covariates in the linear components of (1). The efficacy decision criterion would necessarily require a higher standard in the better prognosis group. In particular, this extension would take advantage of the Bayesian formulation's ability to make use of all the information from the successive phases. Our preliminary investigations have shown that this more general design leads to much more time-consuming simulations, however, since the numerical integrals are of higher dimension. We thus are currently investigating other methods for rapid computation of posterior probability criteria similar to $\psi(d)$.

Computer Software

Computer programs to implement the methods described here are available as the compressed file "efftox97.tar.Z" via anonymous ftp from odin.mdacc.tmc.edu in the subdirectory /pub/source. Fetching and unpacking this tar file automatically creates the subdirectories "prog" and "sim" on your computer. A menu-driven program for conducting the trial is in "prog," and "sim" contains simulation routines for computing operating charactistics and properties of posterior probability intervals. Alternatively, these programs are available from the first author via email at rex@odin.mdacc.tmc.edu, either as source code or a compiled version for use on PCs and MacIntoshes.

Résumé

Nous proposons une stratégie pour définir des plans d'essais cliniques étudiant un seul traitement dans lesquels les buts sont, d'une part, de déterminer une dose du traitement étudié satisfaisant à la fois des critères de tolérance et d'efficacité, d'autre part de traiter un nombre suffisamment

important de patients pour pouvoir estimer avec une précision donnée le taux d'événements sous la dose retenue et, enfin, d'arrêter l'essai rapidement s'il est vraisemblable qu'aucune dose efficace et bien tolérée n'existe. La réponse du patient est caractérisée par une variale ordinale à trois modalités prenant en compte la tolérance et l'efficacité. Comme Thall, Simon, et Estey (1995, *Statistics in Medicine* **44,** 357–379), nous utilisons des critères bayésiens pour générer des règles de décision, tout en nous appuyant sur des critères fréquentistes obtenus par simulation pour choisir une configuration du plan expérimental ayant de bonnes propriétés. Nous illustrons ensuite cette stratégie par un exemple d'application dans un essai de greffe de moëlle pour des hémopathies malignes et un autre exemple dans le cas de l'essai d'un agent biologique pour le mélanome malin.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions.* New York: Dover.

Anderson, J. A. (1984). Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society, Series B* **46,** 1–30.

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

Berry, D. A. (1995). Decision analysis and Bayesian methods in clinical trials. In *Recent Advances in Clinical Trial Design and Analysis,* P. F. Thall (ed), 125–154. Boston: Kluwer.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health related research. In *Bayesian Biostatistics,* D. A. Berry and D. K. Stangl (eds), 3–66. New York: Dekker.

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51,** 1372–1383.

Chevret, S. (1993). The continual reassessment method in cancer phase I clinical trials: A simulation study. *Statistics in Medicine* **12,** 1093–1108.

Conaway, M. R. and Petroni, G. R. (1995). Bivariate sequential designs for phase II clinical trials. *Biometrics* **51,** 656–664.

Conaway, M. R. and Petroni, G. R. (1996). Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* **52,** 1375–1386.

Faries, D. (1991). The modified continual reassessment method for phase I cancer clinical trials. *American Statistical Association Proceedings of the Biopharmaceutical Section* 269–273.

Freedman, L. S. and Spiegelhalter, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* **10,** 357–367.

Goodman, S. N., Zahurak, M. L., and Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* **14,** 1149–1161.

Gooley, T. A., Martin, P. J., Fisher, L. D., and Pettinger, M. (1994). Simulation as a design tool for phase I/II clinical trials: An example from bone marrow transplantation. *Controlled Clinical Trials* **15,** 450–462.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine* **13,** 1665–1677.

Ho, C. H. (1991). Some frequentist properties of a Bayesian method in clinical trials. *Biometrical Journal* **33,** 735–740.

Korn, E. L., Midthune, D., Chen, T. T., Rubinstein, L. V., Christian, M. C., and Simon, R. M. (1994). A comparison of two phase I trial designs. *Statistics in Medicine* **13,** 1799–1806.

L'Ecuyer, P. and Cote, S. (1991). Implementing a random number package with splitting facilities. *ACM Transactions on Mathematical Software* **17,** 98–111.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B* **42,** 109–142.

Møller, S. (1995). An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in Medicine* **14,** 911–922.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics* **52,** 673–684.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **46,** 33–48.

Piessens, S., deDoncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (1983). *Quadpack. A Subroutine for Automatic Integration.* Springer Series in Computational Mathematics, Volume 1. New York: Springer-Verlag.

Rosner, G. and Berry, D. A. (1995). A Bayesian group sequential design for a multiple arm randomized trial. *Statistics in Medicine* **14,** 381–394.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45,** 925–937.

Thall, P. F., Simon, R., and Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* **14,** 357–379.

Thall, P. F., Simon, R., and Estey, E. H. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology* **14,** 296–303.

Received July 1996; *revised March* 1997; *accepted June* 1997.