Journal of Biopharmaceutical Statistics, 16: 623–638, 2006

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400600860394

Taylor & Francis
Taylor & Francis Group

ADAPTIVE DOSE SELECTION USING EFFICACY-TOXICITY TRADE-OFFS: ILLUSTRATIONS AND PRACTICAL CONSIDERATIONS

Peter F. Thall and John D. Cook
Department of Biostatistics and Applied Mathematics, The University of Texas, Houston, Texas, USA

Elihu H. Estey
Department of Leukemia, The University of Texas, Houston, Texas, USA

The purpose of this paper is to describe and illustrate an outcome-adaptive Bayesian procedure, proposed by Thall and Cook (2004), for assigning doses of an experimental treatment to successive cohorts of patients. The method uses elicited (efficacy, toxicity) probability pairs to construct a family of trade-off contours that are used to quantify the desirability of each dose. This provides a basis for determining a best dose for each cohort. The method combines the goals of conventional Phase I and Phase II trials, and thus may be called a "Phase I-II" design. We first give a general review of the probability model and dose-finding algorithm. We next describe an application to a trial of a biologic agent for treatment of acute myelogenous leukemia, including a computer simulation study to assess the design's average behavior. To illustrate how the method may work in practice, we present a cohort-by-cohort example of a particular trial. We close with a discussion of some practical issues that may arise during implementation.

Key Words: Acute leukemia; Adaptive design; Bayesian design; Phase I clinical trial; Phase I-II clinical trial.

1. INTRODUCTION

The purpose of this paper is to review and illustrate by example an outcome-adaptive procedure, proposed by Thall and Cook (2004), that uses both efficacy (E) and toxicity (T) to choose doses of an experimental agent for successive cohorts of patients in an early-phase clinical trial. The method has three basic components. The first is a Bayesian model for the joint probabilities of efficacy (also referred to as "response") and toxicity as functions of dose. The second component consists of criteria for deciding which doses have both acceptably high efficacy and acceptably low toxicity. The third component is based on several elicited (efficacy, toxicity) probability pairs that are considered by the physician to be equally desirable targets. These targets are used to construct a family of efficacy-toxicity trade-off contours

Received October 18, 2005; Accepted May 26, 2006
Address correspondence to Peter F. Thall, Department of Biostatistics and Applied Mathematics, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX, USA; E-mail: rex@mdanderson.org

that provide a basis for quantifying the desirability of each dose. Based on the posterior computed from the current data, each cohort is treated with the most desirable dose from the set of acceptable doses. The method may be called a "Phase I-II" design because it combines the goals of conventional Phase I and Phase II trials, including evaluating toxicity, evaluating efficacy, and finding an acceptable dose.

We begin in Section 2 with a general review of the probability models, a least-squares method for determining the prior's hyperparameters from elicited information, and the dose-finding algorithm. In Section 3, we illustrate the method by an application to a trial of an experimental biologic agent used in combination with conventional chemotherapy for treatment of acute myelogenous leukemia (AML), including computer simulations to describe the design's average behavior. Section 4 provides a cohort-by-cohort illustration of how the method may work in a particular trial. Finally, in Section 5 we discuss some practical and ethical issues that should be considered when implementing the method.

2. REVIEW OF THE MODEL AND METHOD

2.1. Defining the Outcomes

In practice, the definitions of *efficacy* and *toxicity* will vary widely depending on the particular clinical setting. Likewise, the probabilities of events that are considered acceptable in each category will also depend on the particular disease being treated, the trial's entry criteria, and the rates of efficacy and toxicity that may be expected with whatever standard therapies may be available.

For example, one may contrast leukemias, such as AML, with solid tumors, such as lung or breast cancer. Because AML originates in the bone marrow whereas solid tumors do not, and because chemotherapy is only marginally more toxic for cancer cells than for the analogous normal cells, if chemotherapy used to treat AML is to be effective then it must be more toxic for normal bone marrow cells than chemotherapy used to treat solid tumors. Consequently, rates of marrow toxicity, such as infection due to low white blood cell counts, considered unremarkable in AML patients would be considered unacceptable in solid tumor patients. On the other hand, nausea and vomiting are much more likely to limit dose escalation of drugs used for therapy of solid tumors. In general, there often is a positive association between toxicity and efficacy, and this phenomenon provides the rationale for conventionally using toxicity alone to identify a "maximally tolerated dose." In AML, efficacy is defined as induction of disease remission, because it is a precondition for survival. Thus, in cases where standard therapy is very unlikely to be effective for AML, it is reasonable to accept considerably higher probabilities of toxicity with an experimental therapy than in cases where there is a greater chance of success with standard therapy and hence less need for success with experimental therapy. We will refer to this example in Section 3. These examples underscore the point that both E and T are complex events, highly dependent on the particular medical setting. Because the method described here is based on the probabilities of E and T, as well as bounds on these probabilities, it is essential that each outcome be defined collaboratively by the physician and statistician in a manner that is appropriate to the particular trial at hand.

2.2. Probability Model

Let Y_E and Y_T be the respective indicators of E and T, x the standardized dose, and θ the vector of model parameters, and denote the marginal probabilities of these events for a patient given dose x by $\pi_E(x, \theta) = Pr(Y_E = 1 \mid x, \theta)$ and $\pi_T(x, \theta) = Pr(Y_T = 1 \mid x, \theta)$. In practice, if $d_1, \ldots, d_K$ are the raw doses, then one may use $x_j = \log(d_j)$, cd_j, or some other transformation, possibly centered at the mean, to stabilize numerical computations. For simplicity, we will consider only the bivariate binary case where the outcome $\mathbf{Y} = (Y_E, Y_T)$ takes on four possible values: $(1, 1)$ if both E and T occur, $(1, 0)$ if E occurs without T, and so on. We will ignore the trinary outcome case, which may arise in several ways. For example, in the above setting this may be appropriate if E and T are disjoint events and thus $\mathbf{Y} = (1, 1)$ cannot occur, or if $\mathbf{Y} = (1, 1)$ and $(0, 1)$ are combined into the single event "toxicity." In treatment of solid tumors, the five-level ordinal outcome (Complete Response, Partial Response, Stable Disease, Progressive Disease, Death) may be reduced to the three events $E = \{\text{Complete or Partial Response}\}$, $T = \{\text{Progressive Disease or Death}\}$, and $N = \{\text{Stable Disease}\}$. For details pertaining to the trinary outcome case see Thall and Cook (2004). Denote $\pi_{a,b}(x, \theta) = Pr(Y_E = a,\ Y_T = b \mid x, \theta)$ for $a, b \in \{0, 1\}$, with marginal probabilities $\pi_E(x, \theta) = Pr(Y_E = 1 \mid x, \theta) = \pi_{1,1}(x, \theta) + \pi_{1,0}(x, \theta)$ and $\pi_T(x, \theta) = Pr(Y_T = 1 \mid x, \theta) = \pi_{1,1}(x, \theta) + \pi_{0,1}(x, \theta)$. The model for $\pi_{a,b}(x, \theta)$ is determined by the two marginals

$$\text{logit}\{\pi_E(x, \theta)\} = \beta_{E,0} + \beta_{E,1}x + \beta_{E,2}x^2$$

and

$$\text{logit}\{\pi_T(x, \theta)\} = \beta_{T,0} + \beta_{T,1}x$$

and an association parameter, ψ, via the equation (suppressing x and θ)

$$\begin{aligned}\pi_{a,b} = {} & \{\pi_E\}^a\{\pi_T\}^b\{1 - \pi_E\}^{1-a}\{1 - \pi_T\}^{1-b} \\ & + \{-1\}^{a+b}\pi_E\pi_T\{1 - \pi_E\}\{1 - \pi_T\}\{e^\psi - 1\}/\{e^\psi + 1\}.\end{aligned}$$

This model is discussed by Murtaugh and Fisher (1990) and Prentice (1988). Denoting the dose of the ith patient in the trial by $x_{(i)}$, the likelihood based on the data $D_n = \{(x_{(1)}, \mathbf{Y}_1), \ldots, \{(x_{(n)}, \mathbf{Y}_n)\}$ from the first n patients is the usual product over $i = 1, \ldots, n$, of $\pi_{\mathbf{Y}i}(x_{(i)}, \theta)$.

We will assume that all elements of the six-dimensional parameter vector $\theta = (\beta_{E,0}, \beta_{E,1}, \beta_{E,2}, \beta_{T,0}, \beta_{T,1}, \psi)$ are real-valued, and that they follow independent normal priors. Although the model parameters are independent a priori, they are not independent a posteriori. We could easily assume a multivariate normal prior including nonzero correlations, but for convenience we have not done that. The prior in this application thus is characterized by a vector ξ of 12 hyperparameters, consisting of the 6 means and 6 variances of these normal distributions. We obtain these by applying the least method given in Thall and Cook (2004), as follows. First, we elicit the mean values of $\pi_E(x_j, \theta)$ and $\pi_T(x_j, \theta)$ for each $j = 1, \ldots, K$, which produces $2K$ pieces of information. We next specify the prior variances by

matching the first two moments of the distribution of each $\pi_y(x_j, \theta)$ with those of a beta distribution. Since the variance of a beta random variable with mean μ is bounded above by $\mu(1-\mu)$, we equate the prior variance of each $\pi_y(x_j, \theta)$ to $wE\{\pi_y(x_j, \theta)\}(1 - E\{\pi_y(x_j, \theta)\})$ with $w = 1/2$, i.e., half the maximum variance of the beta with the same first two moments as $\pi_y(x_j, \theta)$. One could experiment with other values of w. This yields an additional $2K$ pieces of information. We then treat the $4K$ elicited values like data, and the theoretical means and variances under the model as nonlinear functions of the 12-dimensional hyperparameter vector ξ, and use penalized nonlinear least squares to solve for ξ. Formally, if $m_{y,j}(\xi)$ denotes the prior mean of $\pi_y(x_j, \theta)$, $m^*_{y,j}$ denotes the corresponding elicited value, $s_{y,j}(\xi)$ denotes the prior standard deviation of $\pi_y(x_j, \theta)$, and $s^*_{y,j}$ the corresponding elicited value, we solve for the vector ξ that minimizes the objective function

$$h(\xi) = \sum_{y=E,T} \sum_{j=1,\ldots,K} [(m_{y,j}(\xi) - m^*_{y,j})^2 + (s_{y,j}(\xi) - s^*_{y,j})^2] + c \sum_{j \neq k} (\sigma_j - \sigma_k)^2$$

where $\sigma_1, \ldots, \sigma_6$ are the six normal standard deviations in ξ, and the second sum is a penalty term to stabilize the computations by keeping the prior standard deviations on the same domain, usually using constant $c = .10$ to $.20$. Additional details are given in Thall and Cook (2004). Of course, one may develop priors in another manner. In any case, the priors must be sufficiently uninformative to give a design for which the data dominates the prior, rather than conversely, and that has good operating characteristics. On the other hand, the prior must be sufficiently informative to result in reasonable behavior early in the trial. In this regard, priors for which $\text{var}(\beta_{j,k})$ are very large and thus the priors of $\pi_E(x_j, \theta)$ and $\pi_T(x_j, \theta)$ have most of their mass very close to 0 and 1 are very likely to lead to a design with pathological behavior, and should be avoided.

2.3. Acceptability Criteria

Let $\underline{\pi}_E$ be a fixed lower limit on $\pi_E(x, \theta)$ and $\bar{\pi}_T$ a fixed upper limit on $\pi_T(x, \theta)$, both elicited from the physician. Given the current data, D_n, a dose is considered *acceptable* if

$$Pr\{\pi_E(x, \theta) > \underline{\pi}_E \mid D_n\} > p_E$$

and

$$Pr\{\pi_T(x, \theta) < \bar{\pi}_T \mid D_n\} > p_T,$$

where p_E and p_T are fixed design parameters, usually in the range .05 to .20, that are calibrated to obtain a design with good operating characteristics. As p_E is increased, the method is more likely to exclude a dose for lack of efficacy. As p_T is increased, the method is more likely to exclude a dose for excessive toxicity. One may think of these two criteria as gatekeepers, one for E and one for T, with a dose entering the current set of acceptable doses, $A(D_n)$, only if it is able to pass through both gates. As noted earlier, the definitions of E and T and the corresponding fixed limits $\underline{\pi}_E$ and $\bar{\pi}_T$ play critical roles in the method, since only acceptable doses may be used to treat patients in the trial.

2.4. Dose Desirability and Efficacy-Toxicity Trade-Offs

Because the pair $(\pi_E(x, \theta), \pi_T(x, \theta))$ is two-dimensional, if these probabilities are to be used as criteria for selecting a best dose from $A(D_n)$, then some form of dimension reduction is needed. We address this problem by carrying out the following geometric construction. Denoting $\pi = (\pi_E, \pi_T)$, we begin by eliciting three target probability pairs, $\{\pi_1^*, \pi_2^*, \pi_3^*\}$, that the physician considers to be equally desirable targets in the two-dimensional probability domain $[0, 1]^2$. These are represented by the triangular points in Fig. 1. Each of these elicited targets embodies a trade-off between the chance of obtaining a response and the risk of toxicity. In practice, π_1^* is elicited first and is located in the interior of $[0, 1]^2$. The second point, $\pi_2^* = (\pi_{2E}^*, 0)$, corresponds to a hypothetical case in which $\pi_T = 0$, and the third point, $\pi_3^* = (1, \pi_{3T}^*)$, corresponds to a hypothetical case in which $\pi_E = 1$. That is, if toxicity is impossible, then π_{2E}^* is the lowest value of $\pi_E(x, \theta)$, and π_{3T}^* is the highest value of $\pi_T(x, \theta)$ if response is certain, that make each of the target points π_2^* and π_3^* as desirable as π_1^*. The efficacy-toxicity trade-off contour, C, is a curve that passes through these three points. Once C has been established, a family of trade-off contours partitioning [0,1] is then generated from C, and a desirability, δ is assigned to each contour in such a way that contours closer to the ideal point $\pi = (1, 0)$ have

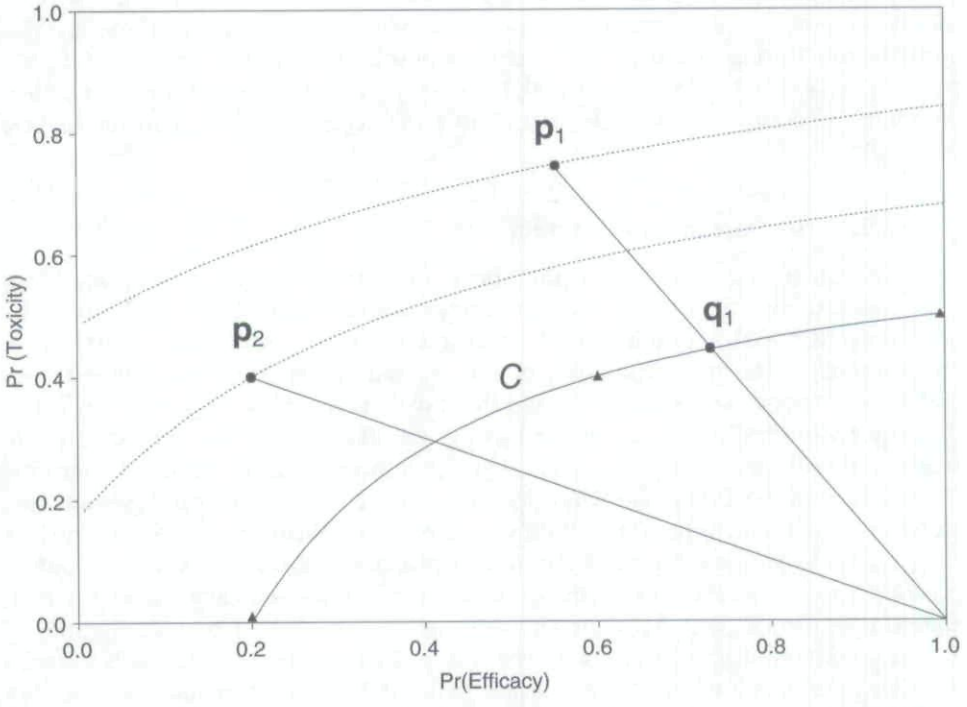

Figure 1 The targeted trade-off contour, C, is represented by the solid curve, and is generated from the three equally desirable elicited target points (.60, .40), (.20, 0), and (1, .40), which are represented by triangular points. The points $\mathbf{p}_1$, $\mathbf{q}_1$, and $\mathbf{p}_2$ represent three probability pairs, and are given along with their corresponding trade-off contours.

higher desirability, and contours farther away from (1, 0) have a lower desirability. Figure 1 illustrates how each point's desirability is computed. For $\mathbf{p}_1$, first note that $\mathbf{q}_1$ is the point where the straight line from $\mathbf{p}_1$ to (1, 0) intersects C. The desirability of $\mathbf{p}_1$ is then defined as the distance from $\mathbf{q}_1$ to (1, 0) divided by the distance from $\mathbf{p}_1$ to (1, 0) minus 1, formally $\delta(\mathbf{p}_1) = \|\mathbf{q}_1 - (1,0)\|/\|\mathbf{p}_1 - (1,0)\| - 1$, where "$\|\mathbf{a} - \mathbf{b}\|$" denotes the Euclidean distance from $\mathbf{a}$ to $\mathbf{b}$. For a given desirability δ, the contour C_δ is defined as all points in $[0, 1]^2$ having that desirability, so that, in particular, $C = C_0$. We set the desirability of the target contour equal to 0 so that all contours closer to (1, 0) than C have $\delta > 0$ and all contours farther away from (1, 0) than C have $\delta < 0$. Additional details of this construction are given in Thall and Cook (2004). The idea of efficacy-toxicity trade-off contours given here is similar to the construction of Thall et al. (2002) used for subgroup-specific treatment selection in the context of a two-stage dynamic treatment regime. An alternative terminology is to call each C_δ an *indifference set* in $[0, 1]^2$.

The software, named EffTox, that implements this methodology (Section 3.3) includes a graphical user interface for plotting the target points and the resulting target contour C during the elicitation process, so that the physician may modify the targets interactively. In practice, one should elicit the targets $\{\pi_1^*, \pi_2^*, \pi_3^*\}$ at the same time that one elicits the anticipated mean probabilities used to construct the prior. When doing this, it is essential to bear in mind that the targets represent what the physician would like to achieve, similar to specifying an alternative parameter value when constructing hypotheses to test, whereas in contrast the elicited prior means represent what the physician anticipates will actually happen. To utilize this construction during the trial, for each acceptable dose x_j we first compute the pair $E\{\boldsymbol{\pi}(x_j, \boldsymbol{\theta}) \mid D_n\} = (E\{\pi_E(x_j, \theta) \mid D_n\}, E\{\pi_T(x_j, \theta) \mid D_n\})$ of posterior means, then compute the desirability δ_j of x_j, and choose the acceptable dose with the highest desirability.

2.5. Trial Design and Conduct

To construct a trial design, the above structure is applied as follows. First, one must establish the disease and trial entry criteria, the treatment and doses, the definitions of E and T, and a tentative maximum sample size, N, and cohort size, c. Next, one elicits the upper and lower bounds $\underline{\pi}_E$ and $\bar{\pi}_T$, the prior means of $\pi_E(x_j, \theta)$ and $\pi_T(x_j, \theta)$ for each $j = 1, \ldots, K$, and the three targets $\{\pi_1^*, \pi_2^*, \pi_3^*\}$, and constructs the target contour C and resulting family of trade-off contours. The EffTox software is essential both to compute the prior's hyperparameters and the trade-off contours. Next, one uses the EffTox software to specify a set of hypothetical dose-outcome scenarios under which the trial will be simulated. A scenario consists of fixed values $\{\pi_1, \ldots, \pi_K\}$ corresponding to the outcome probabilities for the K doses, and in practice it is very useful to plot the dose-toxicity and dose-efficacy curves for each scenario, bearing in mind the numerical values of the elicited bounds $\underline{\pi}_E$ and $\bar{\pi}_T$ and the specified targets $\{\pi_1^*, \pi_2^*, \pi_3^*\}$. The trial is then simulated under each scenario in order to establish its operating characteristics (OCs), which include the selection probabilities and sample sizes at each dose and the probabilities of stopping the trial early. These values are analogous to the usual size and power figures of a conventional test of hypothesis. Due to the complexity of the design, in practice it is necessary to obtain the OCs by computer simulation. The simulation results

may be used to study the design and if necessary adjust the design parameters, including N, c, p_E, and p_T. Additionally, the physician may wish to modify the elicited prior means of $\pi_E(x_j, \theta)$ and $\pi_T(x_j, \theta)$, the bounds $\underline{\pi}_E$ and $\bar{\pi}_T$, or the targets $\{\pi_1^*, \pi_2^*, \pi_3^*\}$ based on preliminary simulation results. The EffTox software supports such simulations. As an additional check, it is also useful to see whether the design behaves reasonably at the start for specified data from the first one or two cohorts.

Formally, the rules for trial conduct are as follows:

1. Treat the first cohort at the starting dose specified by the physician.
2. For each cohort after the first, no untried dose may be skipped, either when escalating or de-escalating.
3. At any interim point in the trial,
   a. if there are no acceptable doses then terminate the trial and do not select any dose;
   b. if there is at least one acceptable dose then treat the next cohort with the dose having maximum desirability.
5. If the trial is not stopped early and there is still at least one acceptable dose at the end then select the acceptable dose having maximum desirability.

3. ILLUSTRATIVE TRIAL

3.1. A Biological Agent to Increase Sensitivity of Anti-AML Chemotherapy

Standard therapy for patients with untreated AML consists of fixed doses of idarubicin and cytosine arabinoside (IA). Although most patients achieve remission with IA, most remissions are short-lived, averaging about 9 months. The probability of response to IA in patients whose AML has relapsed is positively associated with the duration of the preceding remission. This probability is about .10 if this duration was less than 1 year or if the patient never achieved remission with the first course of IA, and median survival in these circumstances is 6 to 9 months.

A protein known as IAP (inhibitor of apoptosis protein) has been found to protect AML cells from the death (apoptosis) induced by IA. An experimental drug, which we will call X, that prevents synthesis of IAP may restore the effectiveness of IA in patients with relapsed AML. Since there has been little clinical experience with X combined with IA, a Phase I-II trial was designed in which X would be given together with a fixed dose and schedule of IA to patients with AML in relapse after a remission of less than 12 months duration, or that had not responded to the first course of IA. For the purpose of illustrating the method, we present a simplified version of the design actually used. For dose-finding, patients are evaluated 35 days after beginning treatment. E is defined as achievement of complete remission (CR) within these 35 days, in order to feasibly assess the effect of X. Toxicity is defined as death or development of life-threatening (grade 4) symptomatic toxicity by day 35. Typically, grade 4 toxicity is considered dose-limiting regardless of its association with symptoms. However, we have found that non-symptomatic grade 4 toxicity is very inconstantly associated with subsequent clinical sequelae. Thus, given the prognosis of the eligible patients, we feel it appropriate to use a less restrictive definition of toxicity.

3.2. Trial Design

The general methodology may be applied to the AML trial as follows. Four dose levels will be studied. A maximum of 36 patients will be treated in cohorts of size 3, with the first cohort given dose level 2, and not skipping any untried dose level when escalating. The starting dose must be chosen by the physician, and this is not necessarily the lowest dose being considered. In applying this method, we have found that a physician may initially specify a set of doses with the lowest as the starting dose, but when asked what he or she would do if the lowest dose were found to be unacceptably toxic, the physician often responds by saying that he or she would then add one or more lower doses and restart the trial. In such settings, we routinely ask the physician to specify such lower doses so that we may include them in the design from the start, but beginning the trial at the same initial dose originally specified, which is now no longer the lowest. Phase I designs using toxicity as the only criterion for dose-finding often start at the lowest dose for fear of excessive toxicity. However, one may argue that dose-finding trials based on efficacy should start at the *highest* dose for fear of administering an ineffective dose. Indeed, some trials using efficacy do start at the highest dose for this reason. When monitoring both efficacy and toxicity, the best starting dose depends on one's relative fear of over-treating and under-treating patients.

In the current application, the acceptability limits are $\underline{\pi}_E = .20$ and $\bar{\pi}_T = .50$, and the two acceptability criteria were applied with $p_E = p_T = .10$. Thus, in this trial x is acceptable if $Pr\{\pi_E(x, \boldsymbol{\theta}) > .20 \,|\, \text{data}\} > .10$ and $Pr\{\pi_T(x, \boldsymbol{\theta}) < .50 \,|\, \text{data}\} > .10$. Although it may seem, intuitively, that the numerical cutoffs $p_E = p_T = .10$ are too small to ensure that doses deemed acceptable are both safe and efficacious, and that large numerical values such as .90 or .95 should be used instead, in practice such large values produce a design that quickly declares no dose acceptable in virtually all cases. Another way to think of the acceptability criteria is in terms of the complementary events, which in this application would be that a dose is *unacceptable* if either $Pr\{\pi_E(x, \boldsymbol{\theta}) < .20 \,|\, \text{data}\} > .90$ or $Pr\{\pi_T(x, \boldsymbol{\theta}) > .50 \,|\, \text{data}\} > .90$. That is, x is unacceptable if it is likely that $\pi_E(x, \boldsymbol{\theta})$ is below the fixed minimal efficacy level .20 or it is likely that $\pi_T(x, \boldsymbol{\theta})$ is above the fixed maximal toxicity level .50. These criteria may be thought of as two gatekeepers, one for safety and the other for efficacy, that together determine whether a dose will be allowed to enter the pool of acceptable doses, from which the most desirable dose will then be chosen. An important point is that several values p_E and p_T should be considered when using simulation to determine design parameters, it is not necessary that $p_E = p_T$, and a general guideline is to explore values in the range .05 to .20. The same rationale that motivated our definitions of E and T led to our selection of the relatively high upper limit $\bar{\pi}_T = .50$. The lower limit $\underline{\pi}_E = .20$ reflects the belief that at least a doubling of the very low historical CR rate (0.10) is required. Although it could be argued that .20 is too low, experience suggests that greater degrees of improvement are very rarely seen and thus would be likely to shut down investigation of therapies that are improvements over standard.

The target trade-off contour is based on the three targeted trade-off probability pairs $\boldsymbol{\pi}_1^* = (.60, .40)$, $\boldsymbol{\pi}_2^* = (.20, 0)$, and $\boldsymbol{\pi}_3^* = (1.0, .50)$. In Fig. 1, these elicited targets are given as triangles, and the resulting target efficacy-toxicity trade-off contour is illustrated by the solid curve. The points labeled $\mathbf{p}_1$, $\mathbf{q}_1$, and $\mathbf{p}_2$

are included, along with their respective trade-off contours, to illustrate how the family of trade-off contours distinguishes between points in [0, 1]. For example, if $\mathbf{p}_1 = E\{\pi(x_1, \theta) \mid D_n\}$ and $\mathbf{p}_2 = E\{\pi(x_2, \theta) \mid D_n\}$, then Fig. 1 shows that x_2 is more desirable than x_1. It is important to note that a dose with $\delta < 0$ is not necessarily "undesirable," just less desirable than any dose with $E\{\pi(x, \theta) \mid D_n\}$ falling on C, where $\delta = 0$.

3.3. Computer Simulations

The OCs of the trial under each of four hypothetical dose-outcome scenarios are summarized in Table 1, and the results are also presented graphically in Fig. 2. For each scenario, the trial was simulated 1000 times. We also conducted additional preliminary simulations to examine the effects of different values of p_E and p_T, as well as other maximum sample sizes and cohort sizes. Although we do not show these additional simulations here because they would greatly increase the length of this article, we recommend that this testing should be done routinely to assess the effects of varying the design parameters and, in collaboration with the physician, this should be used as a basis for choosing a design. In Fig. 2, the true value of π for each dose is represented by its location in the unit square $[0, 1]^2$, the desirability contour for each dose is represented by a dashed line, and the size of the circular dot at each location is the selection percentage for that dose. In addition, the limits $\underline{\pi}_E = .20$ and $\bar{\pi}_T = .50$ are given as solid straight lines in each plot.

In scenario 1, dose level 1 is associated with a true response rate of $\pi_{1E} = .05$ and a true toxicity rate of $\pi_{1T} = .40$, with response and toxicity probabilities of $\pi_2 = (\pi_{2E}, \pi_{2T}) = (.10, .60)$ for dose level 2, $\pi_3 = (.20, .75)$ for level 3, and $\pi_4 = (.35, .85)$ for level 4. Thus, no dose level is acceptable in this case, and the design correctly

Table 1 Operating characteristics of the trial design; these π_E, π_T and δ are the true values under each scenario

Scenario		Dose level 1	2	3	4	None
1	π_E, π_T	.05, .40	.10, .60	.20, .75	.35, .85	
	δ	−.521	−.746	−.912	−.992	
	% Selected	5.1	16.9	6.1	3.5	68.4
	# Patients	7.7	7.8	4.7	3.6	
2	π_E, π_T	.10, .10	.40, .15	.60, .25	.65, .70	
	δ	−.185	.146	.249	−.539	
	% Selected	8.1	36.1	52.2	3.6	00.0
	# Patients	7.6	10.4	14.5	3.5	
3	π_E, π_T	.10, .10	.20, .15	.35, .16	.60, .17	
	δ	−.185	−.097	.077	.358	
	% Selected	8.0	5.8	6.6	78.0	1.6
	# Patients	7.7	4.6	4.8	18.4	
4	π_E, π_T	.30, .10	.55, .40	.40, .55	.15, .70	
	δ	.063	−.037	−.412	−.865	
	% Selected	73.2	25.8	0.3	0.1	0.6
	# Patients	21.3	12.3	1.7	0.5	

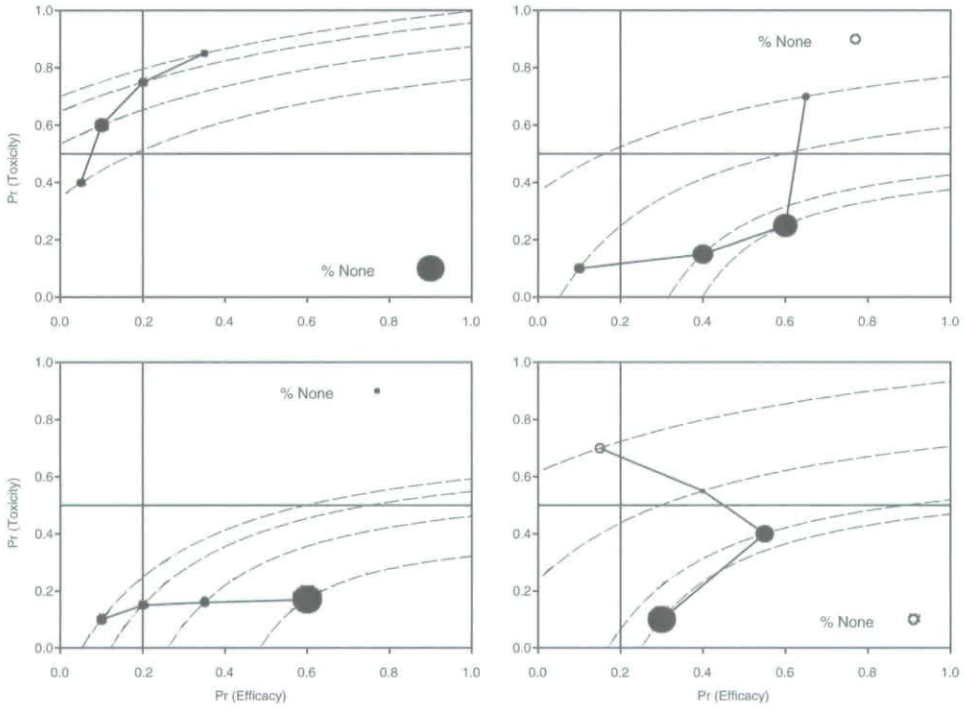

Figure 2 The four dose-outcome scenarios used in the simulation study. In each scenario, for each dose the true π is represented by its location in $[0, 1]^2$, the trade-off contour is represented by a dashed line, and the size of the circular dot is the selection percentage of that dose. The acceptability limits $\underline{\pi}_E = .20$ and $\bar{\pi}_T = .50$ are given as solid straight lines.

stops early with no dose selected 68% of the time. Note, however, that dose level 2 is nearly acceptable, with π_E only .10 below the lower limit of $\underline{\pi}_E = .20$ and π_T only .10 above the upper limit of $\bar{\pi}_T = .50$. Thus, although selecting this dose is an error, it is not a severe error. If we alter scenario 1 by adding .10 to each value of π_T, so that dose levels 2, 3, and 4 become much more toxic and hence much less desirable, then the design stops the trial early 90.8% of the time. In scenario 2, both dose levels 2 and 3 are very desirable, and these are selected 36% and 52% of the time, so the design picks a desirable dose 88% of the time. In scenario 3, the highest dose level is the most desirable by a substantial margin, and it is chosen 78% of the time. In scenario 4, dose levels 1 and 2 are most desirable, and the method picks these 73% and 26% of the time. Figure 2 shows that scenario 4 is a difficult case because the true $\pi_E(x)$ is a non-monotone function of x. In this case, the quadratic model for $\text{logit}\{\pi_E(x, \theta)\}$ comes into play, and the model and algorithm together do a good job of recognizing the non-monotonicity of $\pi_E(x)$ and avoiding the much less desirable dose levels 3 and 4.

4. A COHORT-BY-COHORT ILLUSTRATION

In addition to examining the design's average behavior by simulation, it is also worthwhile to show how the design behaves in a particular case. This is summarized

Table 2 Cohort-by-cohort illustration of the method

Cohort	Dose	Outcomes				Posterior mean $\pi_E(x_j, \theta)$, mean $\pi_T(x_j, \theta)$, and δ of each dose level			
		E^cT	E^cT^c	ET	ET^c	1	2	3	4
Prior						.124, .122	.119, .122	.146, .156	.184, .194
						$\delta = -.170$	$\delta = -.176$	$\delta = -.167$	$\delta = -.167$
1	2	0	1	2	0	.056, .224	.045, .225	.069, .285	.105, .338
						$\delta = -.334$	$\delta = -.347$	$\delta = -.375$	$\delta = -.389$
2	1	1	2	0	0	.025, .0281	.033, .269	.066, .296	.101, .318
						$\delta = -.423$	$\delta = -.402$	$\delta = -.388$	$\delta = -.361$
3	3	0	0	1	2	.042, .275	.233, .276	.573, .305	.773, .333
						$\delta = -.396$	$\delta = -.174$	$\delta = .140$	$\delta = .237$
4	4	0	0	3	0	.031, .215	.267, .400	.677, .561	.871, .661
						$\delta = -.357$	$\delta = -.292$	$\delta = -.255$	$\delta = -.366$
5	3	0	1	0	2	.033, .198	.261, .308	.674, .421	.871, .509
						$\delta = -.339$	$\delta = -.179$	$\delta = .010$	$\delta = -.062$
6	3	0	3	0	0	.028, .187	.179, .260	.528, .340	.779, .409
						$\delta = -.337$	$\delta = -.220$	$\delta = .044$	$\delta = .092$
7	4	1	0	2	0	.030, .138	.179, .277	.502, .438	.754, .568
						$\delta = -.298$	$\delta = -.237$	$\delta = -.140$	$\delta = -.230$
8	3	0	0	1	2	.037, .134	.223, .266	.580, .422	.810, .551
						$\delta = -.286$	$\delta = -.176$	$\delta = -.053$	$\delta = -.172$
9	3	0	1	0	2	.039, .122	.229, .235	.591, .373	.818, .495
						$\delta = -.275$	$\delta = -.136$	$\delta = .041$	$\delta = -.057$
10	3	2	0	0	1	.035, .128	.209, .255	.560, .404	.798, .531
						$\delta = -.285$	$\delta = -.180$	$\delta = -.037$	$\delta = -.137$
11	3	0	0	1	2	.038, .126	.235, .244	.606, .395	.829, .528
						$\delta = -.279$	$\delta = -.139$	$\delta = .014$	$\delta = -.119$
12	3	0	2	0	1	.037, .115	.220, .222	.579, .361	.810, .489
						$\delta = -.273$	$\delta = -.135$	$\delta = .052$	$\delta = -.050$

in Table 2, which gives the outcomes for each of 12 successive cohorts, along with the posterior means of $\pi_E(x_j, \theta)$ and $\pi_T(x_j, \theta)$ and the corresponding δ_j for $j = 1, 2, 3, 4$. When examining this table, it is important to bear in mind that the acceptability limits are $\underline{\pi}_E = .20$ and $\bar{\pi}_T = .50$. The outcomes in Table 2 were chosen purely for the sake of illustration. The design starts at dose level 2, de-escalates to level 1, escalates to level 3 and then to level 4, de-escalates to level 3, revisits level 4, and then settles into level 3 for the final five cohorts. One may consider this last portion of the trial, in which the final 15 patients are all treated at the same final dose level, to be its "Phase II" portion. However, in general there really is no separation between "Phase I" and "Phase II" with this design, and it may switch doses at any point based on new data. Figure 3 gives the means (solid lines) and the upper and lower 2.5 and 97.5 percentiles (dotted lines) of $\pi_E(x, \theta)$ in the left column and $\pi_T(x, \theta)$ in the right column, first under the prior, then a posteriori after 9 patients, and then after the final 36 patients. The purpose of this figure is to illustrate how the Bayesian model learns about the two dose-outcome curves as the data accumulate. The same information is summarized in Fig. 4, but in a very different way, with the four posterior distributions of $\{\pi_T(x_j, \theta), j = 1, 2, 3, 4\}$

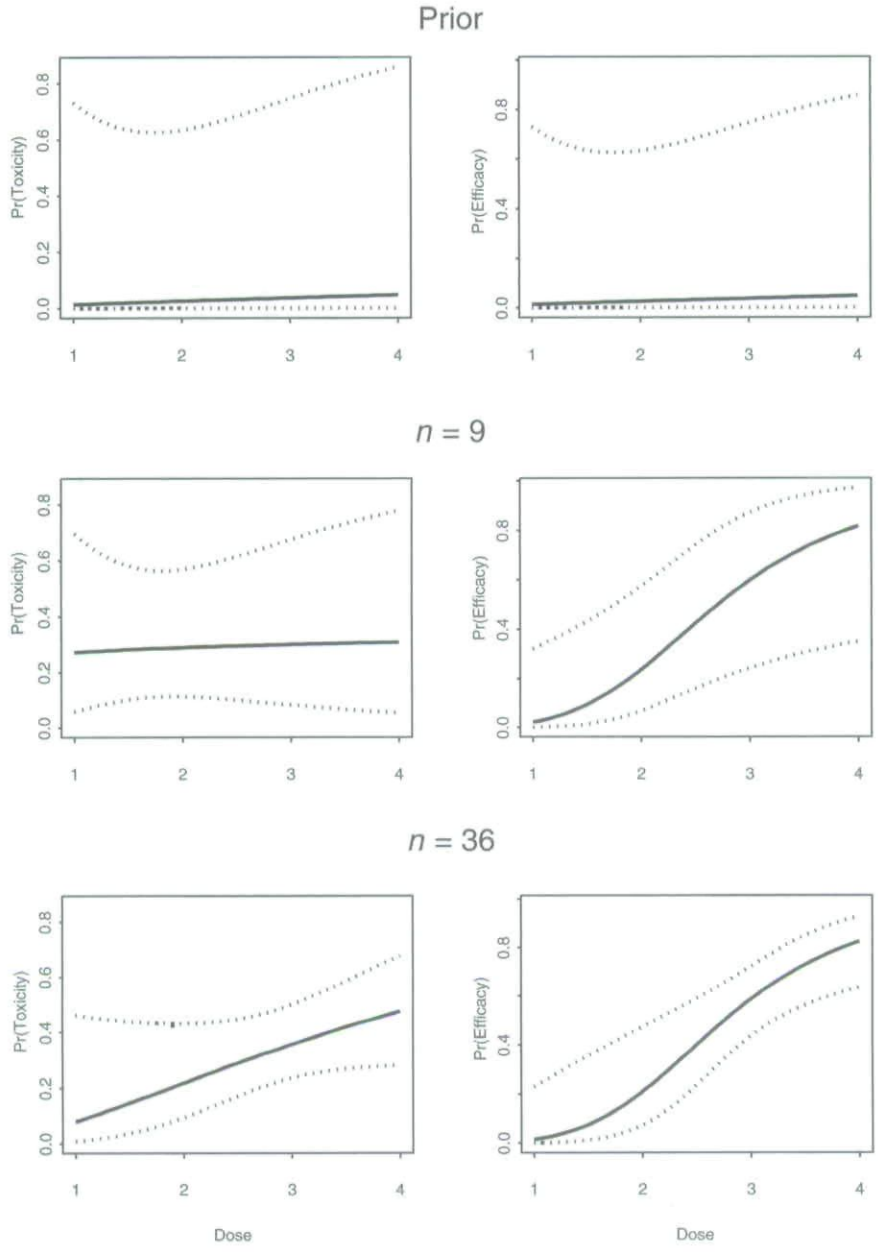

Figure 3 The means of $\pi_T(x, \theta)$ and $\pi_E(x, \theta)$ as functions of x are given as solid lines in the left and right columns, respectively, along with 95% credible intervals as functions of x given by dotted lines. The prior curves are given in the top row, the curves based on the data from 9 patients in the middle row, and the curves based on the data from 36 patients in the bottom row.

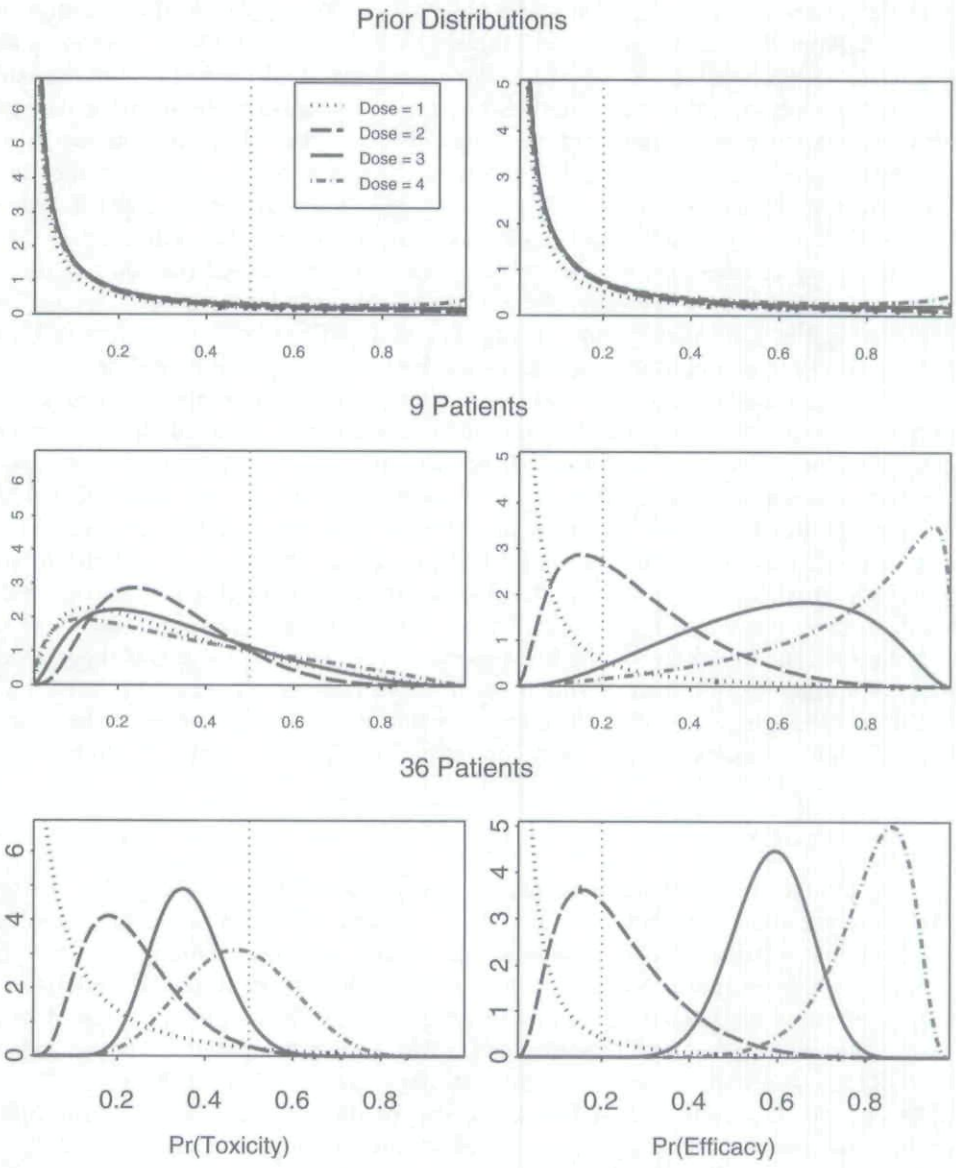

Figure 4 The distributions of $\pi_T(x_j, \theta)$ and $\pi_E(x_j, \theta)$, for doses $j = 1, 2, 3, 4$, a priori in the top row, and a posteriori after 9 and 36 patients in the middle and bottom rows.

overlaid in each plot on the left and the corresponding posteriors of $\{\pi_E(x_j, \theta), j = 1, 2, 3, 4\}$ overlaid on the right. Comparing these distributions to each other as well as to the fixed limits $\underline{\pi}_E = .20$ and $\bar{\pi}_T = .50$ shows clearly that, by the end of the trial, dose level 3 is very acceptable and obviously the most desirable.

It is important to consider how differently the trade-off-based method chooses doses compared with conventional methods that use only the toxicity data. For the first two cohorts, at $x = 2$ there were $2/3$ toxicities and $2/3$ responses, with E and

T perfectly associated as 2 patients had outcome (E, T) and the third had outcome (E^c, T^c). After de-escalating to $x = 1$ there was 1/3 toxicity and 0/3 responses at that dose. Any conventional algorithm based on toxicity alone would not escalate for the next cohort, but rather would choose $x = 1$. In contrast, the steep increase in observed response rate going from $x = 1$ to $x = 2$, and the high positive association between Y_E and Y_T, as formalized in terms of the posterior of $\pi_{a,b}(x, \theta)$ under the Bayesian model, shows that both $E\{\pi_E(x, \theta) \mid \text{data}\}$ and $E\{\pi_T(x, \theta) \mid \text{data}\}$ increase with x, that toxicity is well under control relative to the 50% upper limit, and that the highest dose, $x = 4$, is most desirable. Since we have imposed the "do not skip" rule, the next cohort is treated at $x = 3$. Thus, this decision is a consequence of considering the joint distribution of (Y_E, Y_T) as functions of x in the context of the Bayesian regression model, and assessing both the acceptability and desirability of each dose based on all of the observed data. This example also illustrates the important point that, in general, δ is not a monotone function of dose. A more general point is that with this method, as with any outcome-adaptive dose-finding method, the decision to escalate to an untried dose necessarily must be based on a prediction of what is expected or likely to occur at that higher dose. With our method, this prediction essentially is based on an extrapolation of the fitted regression model to the higher, untried dose. Note that, in this illustrative trial, the final dose turns out to be $x = 3$, which would not have been the case using a conventional method. Despite the 6/6 observed toxicities at $x = 4$ this is the second most desirable dose. Intuitively this is because 5/6 responses were also observed at this dose although, again, all of the data come into play, through the posterior under the Bayesian regression model, when computing the desirability of each dose.

5. DISCUSSION

A number of other authors have proposed methods for dose-finding using two or more outcomes, rather than only one binary toxicity indicator. Gooley et al. (1994) considered two dose-outcome curves, and were among the first to propose computer simulation as a design tool. O'Quigley et al. (2001) and Braun (2002) proposed methods that extend the continual reassessment method (CRM) for toxicity alone proposed by O'Quigley et al. (2001). Ivanova (2003) also generalized the CRM, using a play-the-winner type strategy (Zelen, 1969). Bekele and Thall (2004) proposed a method based on a scheme for differentially weighting multiple ordinal toxicities that have different degrees of clinical importance. Bekele and Shen (2005) proposed a method based on a toxicity indicator and a quantitative biologic outcome.

Our method is very different from all of these procedures, in terms of both the underlying probability model and the algorithm for selecting doses. In particular, the dose-finding methodology that we have described here is somewhat complex. It requires substantial input from the physician and a considerable amount of effort from the statistician to construct the design, and it requires specialized computer software both for constructing a design and for trial conduct. However, given the scientific and ethical advantages of the method, we feel that this extra effort is well warranted. The necessary software is freely available for download at http://biostatistics.mdanderson.org/SoftwareDownload/.

A fundamental motivation for our method is that, despite the fact that most Phase I designs are based on toxicity alone, patients enter a Phase I trial with the hope of achieving a response, that is, disease remission, not simply no toxicity. Moreover, most conventional Phase I trials do such a poor job of selecting a safe dose that the need to make dose adjustments subsequently during what is nominally a Phase II trial is routine. That is, Phase I is really also Phase II, and Phase II is really also Phase I. Our inclusion of both E and T in the dose-finding method formalizes these considerations. Although it may be claimed that a dose-finding trial with, for example, $N = 60$ is excessively large, with our proposed method this sample size accounts for both Phase I and Phase II. Moreover, because only a dose that is both safe and efficacious may be selected in the end, our method is intrinsically superior to conducting a conventional Phase I trial based on T alone followed by a Phase II trial based on E alone. The dose selection decisions made by our method often are counterintuitive to those who have been conditioned to exclude efficacy from their thinking by years of toxicity-only dose-finding. For example, common sense would have no objection to staying at a safe and effective dose, whereas the first response of toxicity-based thinking would be to escalate if toxicity is under control.

Two important aspects of our model are that we do not require that $\pi_E(x, \theta)$ be monotone in x, and that we allow $\pi_T(x, \theta)$ to be either increasing or decreasing in x. The former aspect of the model is reflected in the design's performance in scenario 4 of the simulation study, where the fixed probabilities of response are $\pi_E(x_1) = .30$, $\pi_E(x_2) = .55$, $\pi_E(x_3) = .40$, and $\pi_E(x_4) = .15$, so the quadratic term in $\eta_E(x, \theta)$ comes into play. This feature of the model and method is particularly important in dose-finding studies of biologic agents without any additional cytotoxic agents. In such settings, a wide variety of definitions of E are possible, such as achieving a specified minimum level of a particular chemical or biological reaction in the patient's blood or in a solid tumor, often as measured by a complex laboratory procedure. The form of $\pi_E(x, \theta)$ typically is not known in such settings, so the quadratic form serves as a flexible function to deal with a wide array of possibilities. Certainly, many other models are possible such as, for example, the damped logistic model

$$\pi_E(x, \theta) = \alpha[\exp(\beta_{E,0} + \beta_{E,1}x)/\{1 + \exp(\beta_{E,0} + \beta_{E,1}x)\}]$$

where $0 < \alpha < 1$, so that α is an asymptotic upper limit on $\pi_E(x, \theta)$.

We allow $\pi_T(x, \theta)$ to possibly decrease in x, if $\beta_{T,1} < 0$, because this may be the case for a particular definition of T. For example, we observed this phenomenon in an allogeneic bone marrow transplantation trial where T included systemic infection that cannot be quickly resolved by antibiotics. Larger doses of an experimental agent aimed at treating steroid refractory graft-versus-host disease may have the unexpected beneficial effect of reducing the rate of systemic infection. There are many other examples. The point is that the conventional assumption that $\pi_T(x, \theta)$ must increase with x arose from Phase I trials of cytotoxic agents, and this assumption may simply be incorrect in many Phase I or Phase I-II trials.

A final issue is determining sample size and cohort size. This is logistically important, but technically trivial. In most applications, it is very useful to examine by simulation the behavior of the design for each of an array of values of N and c.

In the illustrative trial studied in Section 3, one might study and compare the OCs of the nine designs arising from the combinations of $N = 36, 48, 60$ and $c = 2, 3, 4$. However, the choice of N should depend not only on the OCs of the design, but also on the reliability of the estimates of $\pi_E(x, \theta)$ and $\pi_T(x, \theta)$ based on the final posterior, as well as the logistics of trial conduct, and what is feasible in terms of the anticipated patient accrual rate and the available resources, including time, drug availability, laboratory facilities, and monetary costs.

ACKNOWLEDGMENT

Peter Thall's research was partially supported by NIH grant RO1 CA 83932.

REFERENCES

Bekele B. N., Thall, P. F. (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J. Am. Stat. Assoc.* 99:26–35.

Bekele, B. N., Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a Phase I/II dose-finding trial. *Biometrics* 61:343–354.

Braun, T. (2002). The bivariate continual reassessment method: extending the CRM to Phase I trials of two competing outcomes. *Controlled Clin. Trials* 23:240–256.

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

Ivanova, A. (2003). A new dose-finding design for bivariate outcomes. *Biometrics* 59:1001–1007.

Murtaugh, P. A., Fisher, L. D. (1990). Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Commun. Stat. Part A Theory Methods* 19:2003–2020.

O'Quigley, J., Hughes, M. D., Fenton, T. (2001). Dose-finding designs for HIV studies. *Biometrics* 57:1018–1029.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44:1033–1048.

Thall, P. F., Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 60:684–693.

Thall, P. F., Sung, H.-G., Estey, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *J. Am. Stat. Assoc.* 97:29–39.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* 64:131–146.