Approximate Dynamic Programming and Its Applications to the Design of Phase I Cancer Trials

Jay Bartroff and Tze Leung Lai

Abstract. Optimal design of a Phase I cancer trial can be formulated as a stochastic optimization problem. By making use of recent advances in approximate dynamic programming to tackle the problem, we develop an approximation of the Bayesian optimal design. The resulting design is a convex combination of a "treatment" design, such as Babb et al.'s (1998) escalation with overdose control, and a "learning" design, such as Haines et al.'s (2003) *c*-optimal design, thus directly addressing the treatment versus experimentation dilemma inherent in Phase I trials and providing a simple and intuitive design for clinical use. Computational details are given and the proposed design is compared to existing designs in a simulation study. The design can also be readily modified to include a first stage that cautiously escalates doses similarly to traditional nonparametric step-up/down schemes, while validating the Bayesian parametric model for the efficient model-based design in the second stage.

Key words and phrases: Dynamic programming, maximum tolerated dose, Monte Carlo, rollout, stochastic optimization

1. INTRODUCTION

In typical Phase I studies in the development of relatively benign drugs, the drug is initiated at low doses and subsequently escalated to show safety at a level where some positive response occurs, and healthy volunteers are used as study subjects. This paradigm does not work for diseases like cancer, for which a nonnegligible probability of severe toxic reaction has to be accepted to give the patient some chance of a favorable response to the treatment. Moreover, in many such situations, the benefits of a new therapy may not be known for a long time after enrollment, but tox-

icities manifest themselves in a relatively short time period. Therefore, patients (rather than healthy volunteers) are used as study subjects, and given the hoped-for (rather than observed) benefit for them, one aims at an acceptable level of toxic response in determining the dose. Current designs for Phase I cancer trials, which are sequential in nature, are an ad hoc attempt to reconcile the objective of finding a *maximum tolerated dose* (MTD) with stringent ethical demands for protecting the study subjects from toxicities in excess of what they can tolerate. It treats groups of three patients sequentially, starting with the smallest of an ordered set of doses. Escalation occurs if no toxicity is observed in all three patients; otherwise an additional three patients are treated at the same dose level. If only one of the six patients has toxicity, escalation again continues; otherwise the trial stops, with the lower dose declared as MTD. As pointed out by Storer (1989), these designs, commonly referred to as 3-plus-3 designs, are difficult to analyze, since even a strict quantitative definition of MTD is lacking, "although it should be taken to mean some percentile of a tolerance distribution with

Jay Bartroff is Assistant Professor, Department of Mathematics, University of Southern California, 3620 South Vermont Ave, KAP 108, Los Angeles, CA 90089, USA (e-mail: bartroff@usc.edu). Tze Leung Lai is Professor, Department of Statistics, and by courtesy, Department of Health Research and Policy and Institute of Computational and Mathematical Engineering, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305, USA (e-mail: lait@stanford.edu).

respect to some objective definition of clinical toxicity," and the "implicitly intended" percentile seems to be the 33rd percentile (related to 2/6). Storer (1989) also considered three other "up-and-down" sequential designs for quantile estimation in the bioassay literature and performed simulation studies of their performance in estimating the 33rd percentile. Subsequent simulation studies by O'Quigley et al. (1990) showed the performance of these designs to be "dismal," for which they provided the following explanation: "Not only do (these designs) not make efficient use of accumulated data, they make use of no such data at all, beyond say the previous three, or sometimes six, responses." They proposed an alternative design, called the *continual reassessment method* (CRM), which uses parametric modeling of the dose–response relationship and a Bayesian approach to estimate the MTD or, more generally, the dose level x such that the probability $F(x)$ of a toxic event is p (1/3 in the case of MTD).

Letting $\theta = (\alpha, \beta)'$ and assuming the usual logistic model

$$(1) \qquad F_\theta(x) = 1/\{1 + e^{-(\alpha + \beta x)}\}$$

for the probability of a toxic response at dose level x, the problem of optimal choice of n dose levels to estimate the MTD seems to be covered by the theory of nonlinear designs. A well-known difficulty in nonlinear design theory is that the optimal design for parameter estimation involves the unknown parameter vector. To circumvent the difficulty, it has been proposed that the design be constructed sequentially, using observations made to date to estimate θ by maximum likelihood and choosing the next design point by using the MLE to replace the unknown parameter value in the optimal design; see Fedorov (1972). If θ is known, then a target probability p of response is attained at the level x_θ that solves $F_\theta(x_\theta) = p$, that is, $x_\theta = [\log(p/(1-p)) - \alpha]/\beta$. Wu (1985) proposed to use at stage $t + 1$ the *certainty equivalence* (or plug-in) level $x_{\hat{\theta}_t}$, where $\hat{\theta}_t$ is the MLE of θ based on (x_i, y_i), $1 \le i \le t$, and y_i is the binary response at dose level x_i. Using some approximations, he also derived a recursive representation of $x_{\hat{\theta}_t}$ and showed that it is asymptotically equivalent (as $t \to \infty$) to the adaptive stochastic approximation rule of Lai and Robbins (1979). The likelihood version of CRM proposed by O'Quigley and Shen (1996) in response to the comments of Korn et al. (1994) on Bayesian designs is in fact a variant of Wu's (1985) design. Babb et al. (1998) pointed out that the symmetric nature of squared error loss used by

CRM may not be appropriate for modeling the toxic response to a cancer treatment. They proposed the *escalation with overdose control* (EWOC) method, which uses an asymmetric linear loss function that penalizes dose level $x = \text{MTD} + \delta$ for $\delta > 0$, corresponding to an overdose, more than an under-dose $x = \text{MTD} - \delta$. Whereas CRM is equivalent to estimating the MTD at each stage by the mean of the posterior distribution of x_θ, EWOC is equivalent to estimating the MTD at each stage by the ωth quantile of the posterior distribution of x_θ, where $\omega \in (0, 1/2)$ is the so-called *feasibility bound*, usually chosen to be slightly less than p. There has also been much work on designs intended to give an accurate post-experiment estimate of the MTD or other functions of the unknown parameter vector θ. For example, locally optimal designs such as c- and D-optimal designs have been investigated extensively for binary responses, and because a nonlinear model's information matrix for binary data is a function of the unknown parameters, locally optimal designs are usually applied using initial estimates, multistage methods or Bayesian priors. Haines, Perevozskaya and Rosenberger (2003) proposed a two-stage design whose first stage is a locally optimal design based on a chosen prior, which is then updated sequentially during its second stage. A comparative study of these methods is given in Section 4.

Since the parameter θ is unknown, *active* statistical learning involves setting the doses at levels that give maximum information about the function of the unknown parameters of interest, the MTD, and how to do this is a problem in nonlinear experimental design theory (Abdelbasit and Plackett, 1983; Dette et al., 2004). On the other hand, there is also an ethical issue of treating patients in the Phase I trial at dose levels below the unknown MTD for safety, and hopefully close to the MTD for efficacy. This dilemma between treatment of current patients and efficient experimentation to gather information for future patients was articulated by Lai and Robbins (1979) in a simple linear regression model $y_k = \alpha + \beta x_k + \varepsilon_k$, where, instead of the MTD, the desired level is $(y^* - \alpha)/\beta$, for some given value y^*. Whereas an asymptotic theory of how this dilemma can be resolved optimally as $n \to \infty$ was developed by Lai and Robbins (1979), it was quite recent that a tractable scheme was developed by Han, Lai and Spivakovsky (2006) to compute an approximately optimal solution for finite sample size n (number of patients enrolled in the trial).

In Section 2 we introduce a basic stochastic optimization problem that incorporates the treatment versus experimentation dilemma in the design of Phase I

cancer trials. This problem adopts a Bayesian formulation as in CRM and EWOC, for which the computation of the posterior distributions of the parameters and of the MTD is described in Section 2. Because the regression function $F_\theta(x) = E_\theta(y|x)$ for the binary response y given by (1) is nonlinear in the parameters, the stochastic optimization problem is considerably more difficult than the linear regression model $E(y|x) = \alpha + \beta x$ considered by Lai and Robbins (1979). We review in Section 2 recent advances in the field of *approximate dynamic programming*, which we use in Section 3 to develop a new tool for tackling the stochastic optimization problem. Using this tool, we derive nearly optimal *hybrid designs* in Section 3. These hybrid designs are convex combinations (and therefore hybrids) of designs that are targeted toward treating the current patient at the best guess of the MTD (e.g., EWOC and CRM) and the Haines–Perevozskaya–Rosenberger designs that are *D*- or *c*-optimal in estimating the model parameters for future patients. The weights in these convex combinations are determined by approximate dynamic programming and can be conveniently stored to provide simple table look-up schemes for the clinical user, as noted in Section 5 which gives some concluding remarks. Section 4 provides a comparative study of the hybrid design and previous designs. It also introduces a modified hybrid design that incorporates the traditional nonparametric step-up/down scheme as a cautious first stage, followed by the model-based design in the second stage of a Phase I cancer trial.

2. STOCHASTIC OPTIMIZATION RELATED TO THE TREATMENT VERSUS EXPERIMENTATION DILEMMA

To begin with, we specify a prior distribution on θ by following Babb et al. (1998) who first specify a range $[x_{\min}, x_{\max}]$ of possible dose values believed to contain the MTD, with $x_{\min}$ believed to be a conservative starting value. Rather than directly specifying the prior distribution π for the unknown parameter θ of the working model to be used in the second stage, which may be hard for investigators to do in practice, an upper bound $q > 0$ on the probability $\rho = F_\theta(x_{\min})$ of toxicity at $x_{\min}$ can be elicited from investigators; uniform distributions over $[x_{\min}, x_{\max}]$ and $[0, q]$ are then taken as the prior distributions for the MTD and $F_\theta(x_{\min})$, respectively. Let $\mathcal{F}_k$ denote the information set generated by the first k doses and responses, that is, by $(x_1, y_1), \ldots, (x_k, y_k)$. Letting η denote the MTD, it

is convenient to transform from the unknown parameters (α, β) in the two-parameter logistic model (1) to (ρ, η) via the formulas

$$(2) \quad \alpha = \frac{x_{\min} \log(1/p - 1) - \eta \log(1/\rho - 1)}{\eta - x_{\min}},$$

$$(3) \quad \beta = \frac{\log(1/\rho - 1) - \log(1/p - 1)}{\eta - x_{\min}},$$

giving

$$
\begin{aligned}
\alpha + \beta x &= ((x - \eta) \log(1/\rho - 1) \\
&\quad - (x - x_{\min}) \log(1/p - 1)) \\
(4) \\
&\quad / (\eta - x_{\min}) \\
&= \psi(x, \rho, \eta).
\end{aligned}
$$

Assuming that the joint prior distribution of (ρ, η) has density $\pi(\rho, \eta)$ with support on $[0, q] \times [x_{\min}, x_{\max}]$, the $\mathcal{F}_k$-posterior distribution of (ρ, η) has density

$$
\begin{aligned}
(5) \quad & f(\rho, \eta | \mathcal{F}_k) \\
&= C^{-1} \prod_{i=1}^{k} \left[\frac{1}{1 + e^{-\psi(x_i, \rho, \eta)}} \right]^{y_i} \\
&\qquad \cdot \left[\frac{1}{1 + e^{\psi(x_i, \rho, \eta)}} \right]^{1 - y_i} \pi(\rho, \eta),
\end{aligned}
$$

where

$$
\begin{aligned}
C &= \int_{x_{\min}}^{x_{\max}} \int_0^q \prod_{i=1}^{k} \left[\frac{1}{1 + e^{-\psi(x_i, \rho, \eta)}} \right]^{y_i} \\
&\qquad \cdot \left[\frac{1}{1 + e^{\psi(x_i, \rho, \eta)}} \right]^{1 - y_i} \\
&\qquad \times \pi(\rho, \eta) \, d\rho \, d\eta
\end{aligned}
$$

is the normalizing constant. The marginal $\mathcal{F}_k$-posterior distribution of η is then

$$(6) \quad f(\eta | \mathcal{F}_k) = \int_0^q f(\rho, \eta | \mathcal{F}_k) \, d\rho.$$

The aforementioned CRM and EWOC doses based on $\mathcal{F}_k$ are the mean and the ω-quantile of (6).

2.1 A Global Risk Function and Its Minimization

Note that using EWOC or CRM amounts to the "myopic" policy of dosing the $(k + 1)$th patient at the dose $x_{k+1} = x$ that minimizes $E[h(x, \eta)|\mathcal{F}_k]$, in which

$$(7) \quad h(x, \eta) = \begin{cases} (x - \eta)^2 & \text{for CRM,} \\ \omega(\eta - x)^+ \\ \quad + (1 - \omega)(x - \eta)^+ & \text{for EWOC,} \end{cases}$$

where $x^+ = \max(x, 0)$ and

$$E[h(x, \eta)|\mathcal{F}_k] = \int_{x_{\min}}^{x_{\max}} h(x, \eta) f(\eta | \mathcal{F}_k) \, d\eta.$$

Since the information about the dose–toxicity relationship gained from x_{k+1} and the response y_{k+1} affects the ability to safely and effectively dose the other patients $k + 2, k + 3, \ldots, n$, one potential weakness of these myopic policies is that they may be inadequate in generating information on θ for treating the rest of the patients, as well as the post-experimental estimate of the MTD for subsequent phases. To incorporate these considerations in a Phase I trial, $x_1, x_2, \ldots, x_n$ should be chosen sequentially in such a way as to minimize the *global risk*

$$(8) \qquad E\left[\sum_{i=1}^{n} h(x_i, \eta) + g(\hat{\eta}, \eta)\right],$$

in which the expectation is taken over the joint distribution of $(\rho, \eta; x_1, y_1, \ldots, x_n, y_n)$. Note that (8) measures the effect of the dose x_k on the kth patient through $h(x_k, \eta)$, its effect on future patients in the trial through $\sum_{i=k+1}^{n} h(x_i, \eta)$, and its effect on the post-trial estimate $\hat{\eta}$ through $g(\hat{\eta}, \eta)$. It can therefore be used to address the dilemma between safe treatment of current patients in the study and efficient experimentation to gather information about η for future patients. As noted in Section 1, Lai and Robbins (1979) have introduced a similar global risk function to address the dilemma between information and control in the choice of x_k in the linear regression model $y_k = \alpha + \beta x_k + \varepsilon_k$ so that the outputs $y_k, 1 \le k \le n$, are as close as possible to some target value y^*. Specifically, they consider (8) with $g = 0$ and $h(x; \alpha, \beta) = (\alpha + \beta x - y^*)^2$.

Dynamic programming is a standard approach to a stochastic optimization problem of the form (8). Define

$$(9) \quad h_k(x) = \begin{cases} E[h(x, \eta)|\mathcal{F}_k], & 0 \le k < n - 1, \\ E[h(x, \eta) \\ \quad + g(\hat{\eta}(x_1, \ldots, x_{n-1}, x), \eta)|\mathcal{F}_{n-1}], \\ \quad k = n - 1. \end{cases}$$

To minimize (8), dynamic programming solves for the optimal design $x_1^*, \ldots, x_n^*$ by backward induction that determines x_k^* by minimizing

$$(10) \quad h_{k-1}(x) + E\left[\sum_{i=k+1}^{n} h_{i-1}(x_i^*)\Big|\mathcal{F}_{k-1}, x_k = x\right]$$

after determining the future dose levels $x_{k+1}^*, \ldots, x_n^*$. Note that (10) involves computing the conditional expectation of $\sum_{i=k+1}^{n} h_{i-1}(x_i^*, \eta)$ given the dose x at stage k and the information set $\mathcal{F}_{k-1}$, and that x_k^* is determined by minimizing such conditional expectation over all x. For $i \ge k + 1$, since x_i^* is a complicated nonlinear function of the past observations and

of $y_k, x_{k+1}^*, y_{k+1}, \ldots, x_{i-1}^*, y_{i-1}$ that are not yet observed, evaluation of the aforementioned conditional expectation is a formidable task. To overcome this difficulty, we use recent advances in approximate dynamic programming, which we first review and then extend and modify for the problem of minimizing the global risk (8).

2.2 Rollout Algorithms

To begin with, consider the problem of minimizing (8) with $g = 0$ and $h(x; \alpha, \beta) = (\alpha + \beta x - y^*)^2$ in the linear regression model $y_k = \alpha + \beta x_k + \varepsilon_k$ with i.i.d. normal errors ε_i having mean 0. Assuming a normal prior distribution of (α, β), the posterior distribution of (α, β) given $\mathcal{F}_{i-1}$ is also bivariate normal with parameters $E_{i-1}(\alpha), E_{i-1}(\beta), E_{i-1}(\alpha^2), E_{i-1}(\beta^2), E_{i-1}(\alpha \cdot \beta)$, in which E_{i-1} denotes conditional expectation given $\mathcal{F}_{i-1}$. These conditional moments have explicit recursive formulas; see Section 4 of Han, Lai and Spivakovsky (2006). The myopic policy that chooses x at stage i to minimize $E[(\alpha + \beta x - y^*)^2|\mathcal{F}_{i-1}]$ is given explicitly by

$$(11) \quad \begin{aligned} \hat{x}_i &= E_{i-1}\{(y^* - \alpha)\beta\}/E_{i-1}(\beta^2) \\ &= \{y^* E_{i-1}(\beta) - E_{i-1}(\alpha\beta)\}/E_{i-1}(\beta^2). \end{aligned}$$

Although the myopic policy is suboptimal for the global risk function (8), Han, Lai and Spivakovsky (2006) use it as a substitute for the intractable x_i^* for $k + 1 \le i \le n$ in (10), in which the conditional expectation can then be evaluated by Monte Carlo simulation. This method is called *rollout* in approximate dynamic programming. The idea is to approximate the optimal policy x_k^* by minimizing (10) with $x_{k+1}^*, \ldots, x_n^*$ replaced by some known *base policy* $\hat{x}_{k+1}, \ldots, \hat{x}_n$, which ideally is some easily computed policy that is not far from the optimum. Specifically, given a base policy $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$, let $\hat{x}_k^{(1)}$ be the x that minimizes

$$(12) \quad h_{k-1}(x) + E\left[\sum_{i=k+1}^{n} h_{i-1}(\hat{x}_i)\Big|\mathcal{F}_{k-1}, \hat{x}_k = x\right],$$

and the expectation in the second term in (12) is typically evaluated by Monte Carlo simulation. The policy $\hat{\mathbf{x}}^{(1)} = (\hat{x}_1^{(1)}, \ldots, \hat{x}_n^{(1)})$ is called the *rollout* of $\hat{\mathbf{x}}$ and has been used for stochastic control problems arising in a variety of applications; see Section 2.1 of Han, Lai and Spivakovsky (2006). The rollout $\hat{\mathbf{x}}^{(1)}$ may itself be used as a base policy, yielding $\hat{\mathbf{x}}^{(2)}$, and, in theory, this process may be repeated an arbitrary number of times, yielding $\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)}, \hat{\mathbf{x}}^{(3)}, \ldots$. Letting $R(\mathbf{x}) =$

$E[\sum_{i=1}^{n} h_{i-1}(x_i)]$, Bayard ([1991](#)) showed that, regardless of the base policy, rolling out n times yields the optimal design and that rolling out always improves the base design, that is, that

$$
\begin{aligned}
R(\hat{\mathbf{x}}) &\geq R(\hat{\mathbf{x}}^{(1)}) \geq R(\hat{\mathbf{x}}^{(2)}) \geq \cdots \\
&\geq R(\hat{\mathbf{x}}^{(n)}) = R(\mathbf{x}^*)
\end{aligned}
$$
(13)

for any policy $\hat{\mathbf{x}}$, where $\mathbf{x}^*$ denotes the optimal policy.

For the global risk function (8) associated with Phase I designs, with h given by (7), one can use the myopic design EWOC or CRM as the base design in the rollout procedure. In contrast with the explicit formula (11) for the case of a linear regression model with normal errors ε_t, the posterior distribution with density function (5) does not have finite-dimensional sufficient statistics and the myopic design involves (a) bivariate numerical integration to evaluate

$$
E[h_i(x_{i+1})|\mathcal{F}_{k-1}, x_k = x]
$$

for $i \geq k$, and (b) minimization of the conditional expectation over x. The simulation studies in [4](#), in which the rollout is implemented with EWOC as the base design, show substantial improvements of the rollout over EWOC and CRM. Although (13) says that rolling out a base design can improve it and rolling out n times yields the dynamic programming solution, in practice, it is difficult to use a rollout (which is defined by a backward induction algorithm that involves Monte Carlo simulations followed by numerical optimization at every stage) as the base policy for another rollout. To overcome this difficulty, we need a tractable representation of successive rollouts, which we develop by using other ideas from approximate dynamic programming (ADP).

2.3 Combining Least Squares with Monte Carlo in ADP

The conditional expectation in (10), as a function of x, is called the *cost-to-go function* in dynamic programming. An ADP method, which grew out of the machine learning (or, more specifically, reinforcement learning) literature, is based on two statistical concepts concerning the conditional expectation. First, for given x and the past information $\mathcal{F}_{k-1}$, the conditional expectation is an expectation and therefore can be evaluated by Monte Carlo simulations, if one knows how $h_k(x_{k+1}^*), \ldots, h_{n-1}(x_n^*)$ are generated. The second concept is that, by (9), $h_i(x_{i+1})$ is a conditional expectation given $\mathcal{F}_i$, which is a regression function (or minimum-variance prediction)

of $h_i(x_{i+1})$, with regressors (or predictors) generated from $\mathcal{F}_i$. Based on a large sample (generated by Monte Carlo), the regression function can be estimated by least squares using basis function approximations, as is typically done in nonparametric regression. Combining least squares (LS) regression with Monte Carlo (MC) simulations yields the following LS-MC method for Markov decision problems in reinforcement learning. Let $\{s_t, t \geq 0\}$ be a Markov chain whose transition probabilities from state s_t to s_{t+1} depend on the action x_t at time t, and let $f_t(s, x)$ denote the cost function at time t, incurred when the state is s and the action x is taken. Consider the statistical decision problem of choosing x at each stage k to minimize the cost-to-go function

$$
\begin{aligned}
Q_k&(s, x) \\
&= E\Bigg\{ f_k(s, x) \\
&\quad + \sum_{t=k+1}^{n} f_t(s_t, x_t) \Big| s_k = s, x_k = x \Bigg\},
\end{aligned}
$$
(14)

assuming that $x_{k+1}, \ldots, x_n$ have been determined. Let

$$
V_k(s) = \min_x Q_k(s, x), \quad x_k^* = \arg\min_x Q_k(s, x).
$$
(15)

These functions can be evaluated by the backward induction algorithm of dynamic programming: $V_n(s) = \min_x f_n(s, x)$, and for $n > k \geq 1$,

$$
\begin{aligned}
V_k(s) = \min_x \{ &f_k(s, x) \\
&+ E[V_{k+1}(s_{k+1})|s_k = s, x_k = x]\},
\end{aligned}
$$
(16)

in which the minimizer yields x_k^*. The LS-MC method uses basis functions ϕ_j, $1 \leq j \leq J$, to approximate V_{k+1} by $\widehat{V}_{k+1} = \sum_{j=1}^{J} a_{k+1,j} \phi_j$, and uses this approximation together with B Monte Carlo simulations to approximate

$$
E[V_{k+1}(s_{k+1})|s_k = s, x_k = x]
$$

for every x in a grid of representative values. This yields an approximation $\widetilde{V}_k$ to V_k and also $\widehat{x}_k$ to x_k^*. Moreover, using the sample

$$
\{(s_{k,b}, \widetilde{V}_k(s_{k,b})), 1 \leq b \leq B\}
$$
(17)

generated by the control action $\widehat{x}_k$, we can perform least squares regression of $\widetilde{V}_k(s_{k,b})$ on $(\phi_1(s_{k,b}), \ldots, \phi_J(s_{k,b}))$ to approximate $\widetilde{V}_k$ by $\widehat{V}_k = \sum_{j=1}^{J} a_{k,j} \phi_j$. Further details of this approach can be found in Chapter 6 of Bertsekas ([2007](#)).

Although the problem (10) can be viewed as a Markov decision problem with the $\mathcal{F}_{t+1}$-posterior distribution being the state s_t, the state space of the Markov chain at hand is infinite-dimensional, consisting of all bivariate posterior distributions of the unknown parameter vector (α, β). If the state space were finite-dimensional, for example, $\mathbb{R}^m$, then one could approximate the value functions (15) by commonly used basis functions in nonparametric regression, such as regression splines and their tensor products; see Hastie, Tibshirani and Friedman (2001). However, in the infinite-dimensional case, there is no such simple choice of basis functions of posterior distributions, which are the states. As pointed out in Section 6.7 of Bertsekas (2007), an alternative to approximating the value functions V_k, called *approximation in value space*, is to approximate the optimal policy by a parametric family of policies so that the total cost can be optimized over the parameter vector. This approach is called *approximation in policy space* and most of its literature has focused on finite-state Markov decision problems and gradient-type optimization methods that approximate the derivatives of the costs, as functions of the parameter vector, by simulation. We now describe a new method for approximation in policy space, which uses iterated rollouts to optimize the parameters in a suitably chosen parametric family of policies.

The choice of the family of policies should involve domain knowledge and reflect the kind of policies that one would like to use for the actual application. One would therefore start with a set of real-valued basis functions of the state s_t of the Markov chain with general, possibly infinitely-dimensional, state space, on which the family of chosen policies will be based. The control policies in this family can be represented by $\pi_t(\phi_1(s_t), \ldots, \phi_m(s_t); \boldsymbol{\beta})$, which is the action taken at time t [after s_t has been observed and the basis functions $\phi_1(s_t), \ldots, \phi_m(s_t)$ have been evaluated] and in which $\boldsymbol{\beta}$ is a parameter to be chosen iteratively by using successive rollouts, with

$$\{\pi_t(\phi_1(s_t), \ldots, \phi_m(s_t); \boldsymbol{\beta}^{(j)}), 1 \leq t \leq n\}$$

being the base policy for the rollout $\mathbf{x}^{(j+1)}$. Using the simulated sample

$$\{(s_{k,b}, x_{k,b}^{(j+1)}), 1 \leq b \leq B\},$$

in which $s_{k,b}$ denotes the bth simulated replicate of s_k, least squares regression of $x_{k,b}^{(j+1)}$ on $\pi_k(\phi_1(s_{k,b}), \ldots, \phi_m(s_{k,b}); \boldsymbol{\beta})$ is performed to estimate $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(j+1)}$; nonlinear least squares is used if π_k is nonlinear in $\boldsymbol{\beta}$.

In view of (13), each iteration is expected to provide improvements over the preceding one. A concrete example of this method in a prototypical Phase I setting is given in the next section, where linear regression splines are used in iterated rollouts. In this setting the state variable s_t represents the complete treatment history up to time t in the trial—all prior distributions, doses and responses up to that time—and the cost function $f_t(s_t, x)$ will be replaced by $h_t(x)$ given by (9).

3. HYBRID DESIGNS AS BASE POLICIES FOR ITERATED ROLLOUTS

In their use of rollouts to approximate the optimum for (8) for the normal model, Han, Lai and Spivakovsky (2006), Section 3, used the structure of their problem to come up with an ingenious "perturbation of the myopic rule" as a base policy to improve the performance of the rollout, without performing second- or higher-order rollouts. In this section we explore this technique in the context of Phase I designs, using such perturbations—called here *hybrid designs*—both as base policies and as a way to represent highly complicated but efficient policies in a simple, clinically useful way. As pointed out in Section 2.1, the objective function of the dynamic programming problem (8) involves both experimentation (for estimating the MTD) and treatment (for the patients in the study). Consider the kth patient in a trial of length n ($\geq k$). If the kth patient were the last patient to be treated in the trial ($n = k$), the best dose to give him/her would be the myopic dose m_k that minimizes $h_{k-1}(x_k)$, given by (9). On the other hand, early on in the trial, especially if $n - k$ is relatively large, one expects the optimal dose to be perturbed from m_k in the direction of a dose that provides more information about the dose–response model, for the relatively large number of doses that will have to be set for the future patients. Since the optimal design theory for learning the MTD under overdose constraints, developed by Haines, Perevozskaya and Rosenberger (2003), yields a c- or D-optimal design ℓ_k, we propose to use the following *hybrid design* representation of the optimal dose sequence:

$$(18) \qquad x_k^* = (1 - \varepsilon_k)m_k + \varepsilon_k \ell_k,$$

where ℓ_k is the chosen "learning design." Of course, any dosing policy admits the representation (18) with

$$\varepsilon_k = \frac{x_k^* - m_k}{\ell_k - m_k} \cdot \mathbf{1}_{\{\ell_k \neq m_k\}}.$$

However, we will show that it is possible to use rollouts to choose ε_k of a simple form, not depending

on x_k^*, such that the resulting hybrid design given by the right-hand side of (18) is highly efficient. Similar ideas have been used in "ε-greedy policies" in reinforcement learning (Sutton and Barto, 1998, page 122).

From our simulation studies that include the example in Section 3.2, we have found that the sequential c-optimal design (Haines et al., 2003, Section 5) with c being the vector $(0, 1)'$ works well for learning design ℓ_k in (18), which we now briefly explain. In general, optimal designs such as c- and D-optimal can be characterized as optimizing some convex loss function Ψ of the information matrix $I(\theta, \xi)$ associated with the parameter value θ and a measure ξ on the space of design points (see Fedorov, 1972). Here $(nI(\theta, \xi))^{-1}$ is interpreted as the asymptotic variance of the MLE $\widehat{\theta}_n$ of θ. The optimization problem can be generalized to the sequential Bayes setting, with prior distribution π on θ, by finding the ξ that minimizes

$$(19) \quad \int \Psi[I(\theta, \xi_{k-1}) + I(\theta, \xi)]\pi(\theta|\mathcal{F}_{k-1})\, d\theta$$

at the kth stage, where ξ_{k-1} is the empirical measure of the previous design points. In the case $k = 1$, (19) is replaced by $\int \Psi[I(\theta, \xi)]\pi(\theta)\, d\theta$. For a given vector c, the c-optimal design measure ξ minimizes the asymptotic variance of the linear estimator $c'\widehat{\theta}_n$ of $c'\theta$ or, equivalently, $\Psi[I(\theta, \xi)] = c'(I(\theta, \xi))^{-1}c$. Taking the Bayesian c-optimal design with $c = (0, 1)'$ as the learning design ℓ_k in (18) gives $c'\theta = c'(\alpha, \beta)' = \beta$, hence, this design is optimal, in some sense, for learning about β or, equivalently, about the slope

$$\left.\frac{\partial}{\partial x}E(y|x)\right|_{x=\eta} = \left.\frac{\partial}{\partial x}\left(\frac{1}{1 + e^{-(\alpha + \beta x)}}\right)\right|_{x=\eta}$$
$$= \beta p(1 - p)$$

of the dose response curve (1) at the MTD, for which p is $1/3$ or some other prespecified value. This has the following connections to the stochastic optimization problem of Lai and Robbins (1979) discussed in Section 1 and to the rollout procedure of Han, Lai and Spivakovsky (2006). For the normal model discussed in Section 1 and as an asymptotic limiting case of other models, Sacks (1958) showed that the optimal value of the step size (a user-supplied parameter in the Lai–Robbins procedure affecting its convergence rate) is proportional to $(\partial/\partial x)E(y|x)$. Moreover, Han, Lai and Spivakovsky (2006), Section 3, found that in the normal model, perturbations of the myopic policy in the direction of this c-optimal design provide a base design for a rollout that has comparable performance to that of an "oracle policy."

3.1 Relating ε_k to the Uncertainty in the Bayes Estimate $E(\eta|\mathcal{F}_{k-1})$

Since the treatment versus experimentation dilemma discussed in Section 2 stems from the uncertainty in the current estimate of the MTD η, it is natural to expect that the amount of perturbation from the myopic dose m_k depends on the degree of such uncertainty, using little perturbation when the posterior distribution of η is peaked, and much more perturbation when it is spread out. This suggests choosing ε_k as a function of the posterior variance $v_{k-1}^2 = \text{Var}(\eta|\mathcal{F}_{k-1})$, whose reciprocal is called the "precision" of $E(\eta|\mathcal{F}_{k-1})$ in Bayesian parlance. Following the approach described in Section 2.3, we use functions of $s_k = v_{k-1}/v_0$ as basic features of the posterior distribution of η to approximate the ε_k in (18).

To begin, Monte Carlo simulations are performed to obtain the rollout $\mathbf{x}^{(1)}$ of EWOC, yielding a simulated sample $\{(e_{k,b}, s_{k,b}), 1 \le b \le B\}$, where $e_{k,b}$ is the bth simulated replicate of

$$(20) \quad e_k = \frac{x_k^{(1)} - m_k}{\ell_k - m_k} \cdot \mathbf{1}_{\{\ell_k \ne m_k\}},$$

which is essentially the same as (18) with (x_k^*, ε_k) replaced by $(x_k^{(1)}, e_k)$. The basic idea in Section 2.3 can be implemented via nonparametric regression of $e_{k,b}$ on $s_{k,b}$, yielding the estimated regression function g_k. Letting $\widehat{e}_k = g_k(s_k)$, the hybrid design $x_k = (1 - \widehat{e}_k)m_k + \widehat{e}_k\ell_k$ can then be used as the base policy to form the rollout $\mathbf{x}^{(2)}$, and this procedure can be repeated to obtain the iterated rollouts $\mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \dots$.

Linear regression splines, and their tensor products for multivariate regressors, provide a convenient choice of basis functions; see Section 9.4 of Hastie, Tibshirani and Friedman (2001). For the present problem, it suffices to use a truncated linear function

$$(21) \quad g_k(s) = \min\{1, (\beta_k^{(0)} + \beta_k^{(1)}s)^+\}$$
$$\text{for } s_* \le s \le s^*,$$

where s_* and s^* are the minimum and maximum of the sample values $s_{k,b}$, $1 \le b \le B$, and to extend beyond the range $[s_*, s^*]$ by

$$(22) \quad g_k(s) = \begin{cases} sg_k(s_*)/s_*, & 0 \le s \le s_*, \\ g_k(s^*), & s \ge s^*, \end{cases}$$

which agrees with the constraint $g_k(0) = 0$ and ensures that the weight assigned to experimentation does not exceed $g_k(s^*)$. A further simplification is to group the data into K blocks so that $\varepsilon_k = \varepsilon_k(s)$ does not vary with

TABLE 1
*Risk, bias and RMSE of the final MTD estimate, DLT rate and overdose rate (OD) of
EWOC, rollout (ROLL) of EWOC, and 1st and 2nd hybrid approximations*

Design	Risk	Bias	RMSE	DLT	OD
EWOC	0.84 (0.01)	−0.20 (0.010)	0.31 (0.04)	29.8% (0.7%)	21.9% (0.6%)
ROLL	0.75 (0.01)	−0.04 (0.009)	0.22 (0.03)	33.0% (0.7%)	31.2% (0.7%)
Hybrid 1	0.75 (0.02)	−0.14 (0.012)	0.29 (0.06)	33.5% (1.5%)	37.5% (1.5%)
Hybrid 2	0.71 (0.01)	−0.04 (0.005)	0.22 (0.04)	31.24% (0.9%)	27.8% (0.9%)

k within each block, since it is expected that the amount of experimentation for the initial stages depends mostly on the uncertainty about η, while for the final stages experimentation would only benefit the post-trial estimate of η.

3.2 Example and Simulation Study

We illustrate the method in Section 3.1 by applying it to the following example, in which $n = 10$ and $[x_{\min}, x_{\max}]$ is transformed to $[0, 1]$ by location and scale changes. Independent uniform priors on $[0, q]$ and $[0, 1]$ are used for $\rho = F_\theta(x_{\min})$ and the MTD η, respectively; see (2) and (3) and the sentence following it. We use $q = 1/3$ and the EWOC loss with $\omega = 1/4$ in (7), and the squared error loss $g(\widehat{\eta}, \eta) = (\widehat{\eta} - \eta)^2$ in (8). Since n is relatively small, we can assume for simplicity that $(\beta_k^{(0)}, \beta_k^{(1)})$ in (21) does not vary with k and estimate the common $(\beta^{(0)}, \beta^{(1)})$ by applying least squares regression to the sample

$$\{(e_{k,b}, s_{k,b}) : 1 \leq k \leq n, 1 \leq b \leq B\}.$$

We also simply use (21) for all s without performing the extrapolation beyond $[s_*, s^*]$. Rolling out EWOC as the base design and using $B = 2000$ simulations, the preceding procedure gave $(\beta^{(0)}, \beta^{(1)}) = (0.096, 0.02)$. Putting

$$(23) \qquad \varepsilon_k = \min\{1, (0.096 + 0.02\nu_{k-1}/\nu_0)^+\}$$

in the hybrid design

$$(24) \qquad x_k^{(1)} = (1 - \varepsilon_k)m_k + \varepsilon_k \ell_k,$$

we used $\mathbf{x}^{(1)}$ as the base policy of a second rollout, for which the preceding procedure yielded $(\beta^{(0)}, \beta^{(1)}) = (-0.72, 0.94)$. Here we used the sequential c-optimal design with $c = [0, 1]'$ as the learning design ℓ_k (see Section 3). Table 1 contains the operating characteristics, explained below, of EWOC and its rollout, the first

hybrid design $\mathbf{x}^{(1)}$ with ε_k given by (23) and the second hybrid design $\mathbf{x}^{(2)}$ in which (0.096, 0.02) in (23) is replaced by $(-0.72, 0.94)$. Each result is based on 2000 simulation runs. The values of (ρ, η) were generated from the prior distribution given by the joint uniform distribution on $[0, q] \times [x_{\min}, x_{\max}]$. Figure 1 plots the cumulative risk $R_k = \sum_{i=1}^k E[h_{i-1}(x_i)]$ of the EWOC, rollout and hybrid designs for $k = 1, \ldots, n(= 10)$. The operating characteristics in Table 1 are the Monte Carlo estimates of overall risk R_{10}, the bias and root mean squared error (RMSE) of the terminal MTD estimate $\widehat{\eta}_{10}$, the DLT rate $P(y = 1)$ and the overdose rate OD, which is the expected proportion of patients treated at doses higher than η. Standard errors are given in parentheses.

The first hybrid design, which is an approximation to the rollout design, provides more than 10% improvement in terminal risk R_{10} over the myopic policy. The second hybrid design provides an additional 5% improvement in the terminal risk R_{10}, and also smaller values of the DLT and OD rates than the rollout design. The Monte Carlo simulations used to evaluate the operating characteristics and to fit the hybrid designs were performed by using rejection sampling to simulate from the posterior distribution. At each stage, the posterior distribution of (ρ, η) is continuous and supported on the compact set $[0, q] \times [x_{\min}, x_{\max}]$, hence, the joint uniform distribution on $[0, q] \times [x_{\min}, x_{\max}]$ is a natural candidate for the instrumental distribution in rejection sampling; see also the last paragraph of Section 5.

4. A TWO-STAGE MODIFICATION AND COMPARATIVE STUDY

Babb et al. (1998) used EWOC to design a Phase I trial to determine the MTD, with $p = 1/3$, of the antimetabolite 5-fluorouracil (5-FU) for the treatment of
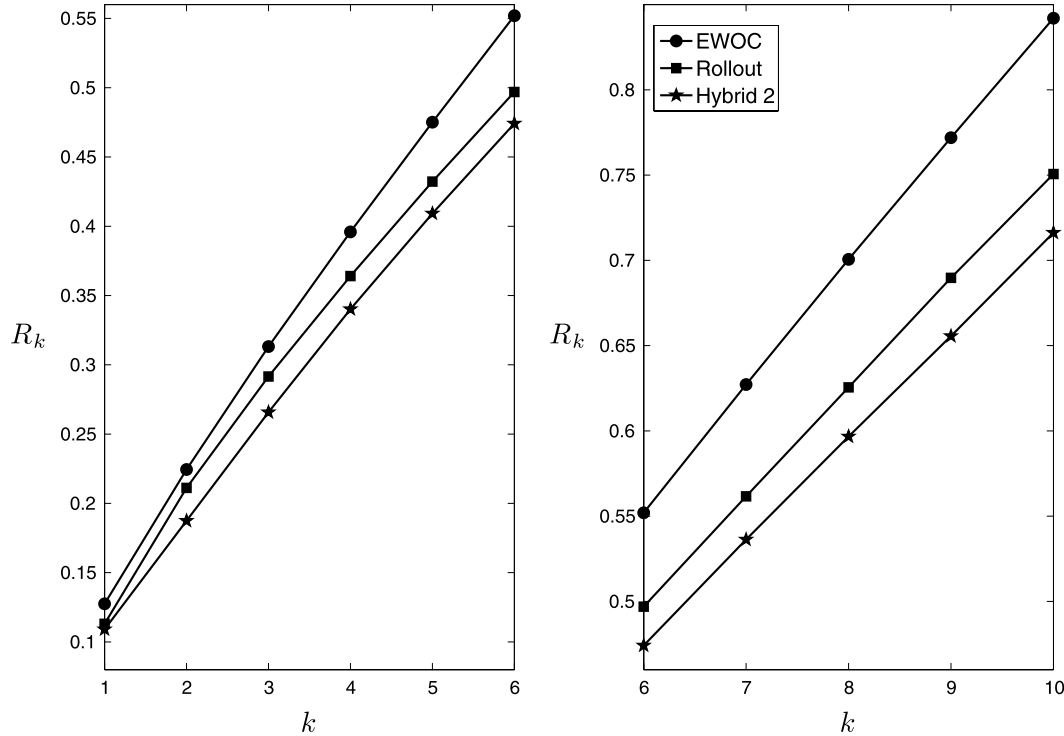
FIG. 1.   *Risk for EWOC, rollout of EWOC and hybrid designs.*

solid tumors in the colon, when taken in conjunction with fixed levels of the agents leucovorin (20 mg/m^2) and topotecan (0.5 mg/m^2). In this setting, a toxicity is considered a grade 4 hematologic or grade 3 or 4 nonhematologic toxicity within 2 weeks. As mentioned above, EWOC involves specifying pre-trial a set

$$(25) \qquad x_{\min} = \lambda_1 < \lambda_2 < \cdots \leq x_{\max}$$

of possible dose values believed to contain the MTD, where $x_{\min}$ is taken as the starting value. Based on preliminary studies of 5-FU given in conjunction with topotecan, a dose of $x_{\min} = 140$ mg/m^2 of 5-FU was believed to be safe when given with 0.5 mg/m^2 of topotecan. Also, a previous trial concluded that the MTD of 5-FU was 425 mg/m^2 when administered without topotecan, so $x_{\max}$ was taken to be 425 mg/m^2 since 5-FU has been observed to be more toxic when given with topotecan than alone. The two-parameter logistic model (1) was chosen based on previous experience with the agents, and uniform prior distributions over $[x_{\min}, x_{\max}]$ and $[0, 0.2]$ were chosen for the MTD and the probability $F_\theta(x_{\min})$, respectively. A feasibility bound of $\omega = 0.25$ was chosen for EWOC and $p = 1/3$. We compare EWOC and other previous designs with the rollout (abbreviated by ROLL)

of EWOC and the Hybrid 1 design described in Section 3.2 in this setting with $n = 24$.

To give a feel for the computational time required for the EWOC, ROLL and Hybrid 1 designs, on a desktop personal computer with a 2.66 GHz processor, the simulation of a single $n = 24$ run of the ROLL design with the EWOC base design took 49 minutes, whereas the Hybrid 1 design took 0.4 seconds and EWOC took 0.12 seconds. The Hybrid 1 design is computationally much simpler than ROLL since it does not perform rollouts of a base design, but rather calculates its dose via (24), where m_k is the EWOC dose and ℓ_k is the sequential c-optimal learning design. So although the interpolation function (23) is derived from data gathered by ROLL during its rollouts as described in 3, the Hybrid 1 design has computational time on the order of EWOC and the learning design ℓ_k, even though the computational time required for ROLL is large.

4.1 A Comparative Study

Table 2 first lists Bayesian designs, followed by non-Bayesian designs that include Wu's (1985) design, stochastic approximation (Lai and Robbins, 1979) and two 3-plus-3 dose escalation designs. The first 3-plus-3, denoted by $3+3_{10}$, uses 10 uniformly-spaced dose levels in $[x_{\min}, x_{\max}] = [140, 425]$. The second

TABLE 2
Risk, bias and RMSE of the final MTD estimate, DLT rate and MTD overdose rate (OD), with SEs in parentheses, of various designs

Design	Risk	Bias	RMSE	DLT	OD
ROLL	0.81 (0.01)	−0.069 (0.002)	0.126 (0.022)	27.68% (1.70%)	29.17% (1.87%)
Hybrid 1	0.92 (0.03)	−0.075 (0.003)	0.128 (0.028)	24.68% (0.86%)	23.48% (0.68%)
EWOC	1.13 (0.01)	−0.076 (0.003)	0.138 (0.024)	26.17% (0.98%)	19.69% (0.89%)
CRM	1.65 (0.01)	0.037 (0.003)	0.118 (0.021)	36.37% (1.10%)	62.69% (1.10%)
c-opt	1.71 (0.01)	0.060 (0.003)	0.126 (0.022)	23.44% (0.95%)	12.42% (0.74%)
D-opt	1.96 (0.02)	−0.084 (0.006)	0.143 (0.023)	13.55% (0.31%)	3.78% (0.17%)
Wu	1.77 (0.04)	0.038 (0.009)	0.122 (0.045)	23.40% (0.54%)	40.25% (0.77%)
SA	1.52 (0.02)	0.063 (0.003)	0.131 (0.022)	22.39% (0.93%)	35.56% (0.40%)
$3+3_{10}$	1.87 (0.01)	0.060 (0.003)	0.138 (0.024)	17.06% (0.84%)	0.85% (0.21%)
$3+3_{20}$	2.19 (0.02)	0.070 (0.002)	0.161 (0.025)	14.11% (0.81%)	0.75% (0.29%)

uses 20 uniformly-spaced dose levels and is denoted by $3+3_{20}$. Besides EWOC and its rollout ROLL, the Bayesian designs include CRM, the constrained D-optimal design (abbreviated by D-opt) of Haines et al. (2003) with constraint $\varepsilon = 0.05$ and the unconstrained sequential Bayesian c-optimal design (abbreviated by c-opt) with c being the vector $(0, 1)^T$. The prior density is assumed to be uniform:

$$\pi(\rho, \eta) = [q(x_{\max} - x_{\min})]^{-1}$$
$$(26) \qquad \cdot 1\{(\rho, \eta) \in [0, q] \times [x_{\min}, x_{\max}]\}$$

with $q = 0.2$, where $1(A)$ denotes the indicator of a set A. The values of (ρ, η) were generated from the prior distribution (26).

The performance of these designs is first evaluated in terms of the global risk (8), in which we use the squared error $g(\widehat{\eta}, \eta) = (\widehat{\eta} - \eta)^2$ for the MTD estimate $\widehat{\eta} = \widehat{\eta}(x_i, y_i, \dots, x_n, y_n)$. We then evaluate performance exclusively in terms of the bias and root mean squared error (RMSE) of $\widehat{\eta}$ without taking into consideration the risk to current patients, noting that the c- and D-optimal designs focus on errors of post-trial parameter estimates. Finally, since safety of the patients in the trial is the primary concern of traditional 3-plus-3 designs, performance is also evaluated in terms of the DLT rate and the probability of overdose (i.e., dose level exceeding the MTD). Each result in Table 2 is based on 2000 simulations.

The results in Table 2 show that the effects of considering the "future" patients is large, with ROLL and Hybrid 1 substantially reducing the global risk from the myopic designs: in the case of ROLL, about 30% from EWOC, 35% from CRM, and more from the 3-plus-3, c- and D-opt, SA and Wu designs. Although ROLL has somewhat smaller global risk than Hybrid 1, it is computationally much more expensive, as noted above. The results for the 3-plus-3 designs show that they are highly sensitive to the choice of $\lambda_1, \lambda_2, \dots$ in (25). The $3+3_{10}$ design, using 10 uniformly-spaced levels in $[x_{\min}, x_{\max}]$, performs perhaps surprisingly well, as it even has smaller risk than D-opt, which suffers from substantial under-dosing due to its overdose constraint of $\varepsilon = 0.05$. This seems to be because the number 10 of dose levels was a fortuitous choice given the parameter values and sample size of this study, allowing the $3+3_{10}$ design to escalate to near the MTD in most cases. However, there is often little information about the appropriate number of doses and scale before a Phase I cancer trial begins, and when dose levels are chosen over a less fortuitous range or on too fine a scale, as with the $3+3_{20}$ design, the majority of doses can end up being administered at levels far below the therapeutic range near the MTD. We emphasize that all of these designs are being evaluated using the EWOC loss function in (7) which, in particular for CRM, differs from its associated loss function; using
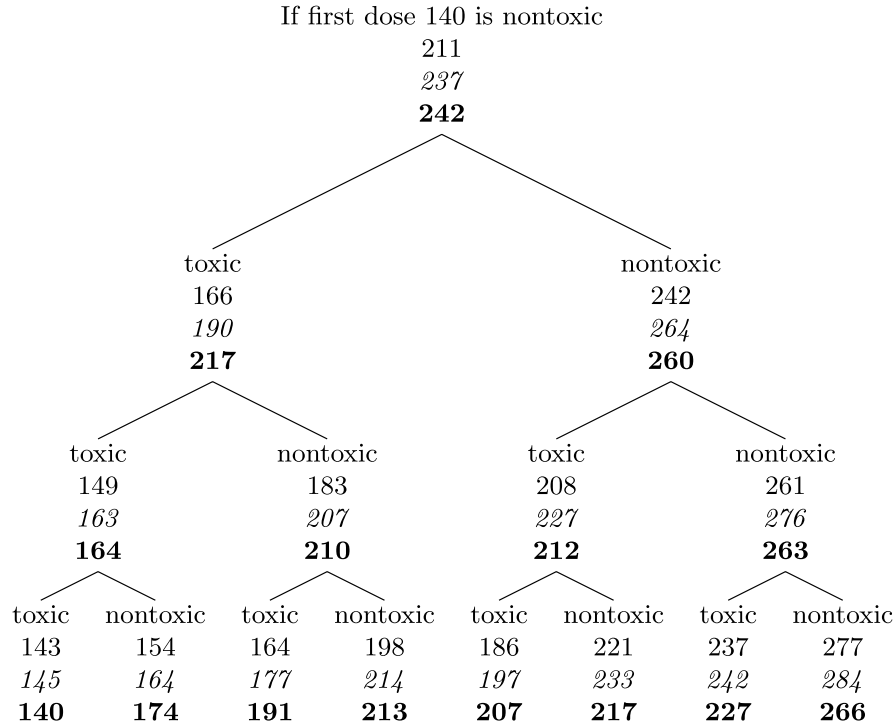
If first dose 140 is nontoxic
211
237
242

```
                    toxic                              nontoxic
                    166                                  242
                   190                                  264
                   217                                  260
           toxic        nontoxic             toxic          nontoxic
           149            183                 208              261
          163            207                 227              276
          164            210                 212              263
      toxic  nontoxic  toxic  nontoxic   toxic  nontoxic  toxic  nontoxic
      143     154      164     198       186     221      237     277
     145     164      177     214       197     233      242     284
     140     174      191     213       207     217      227     266
```

FIG. 2. *The first five dose levels given by EWOC, ROLL (in italics) and Hybrid 1 (in bold) in the 5-FU trial.*

the CRM loss function in (7) results in CRM having smaller global risk than EWOC, but again ROLL (with CRM as its base design) yields smaller global risk than both and the relationship between the other designs remains roughly unchanged.

In terms of MTD estimation accuracy, CRM and Wu have the smallest RMSE, closely followed by c-opt and ROLL; CRM and Wu also have the smallest absolute bias. It is interesting to note that the designs which explicitly account for the asymmetric underdose/overdose relationship, that is, ROLL, EWOC and D-opt (through its overdose constraint ε), are negatively biased, while the others all have positive bias.

In terms of safety, D-opt and the $3 + 3$ designs have the smallest DLT and OD rates, but in view of their large risk values, this safety comes at the cost of low doses that are nontherapeutic. CRM has high estimation accuracy and moderate risk, but also the largest DLT and OD rates because of its symmetric loss function. The remaining designs, ROLL, EWOC, Wu and SA, all have comparable DLT and OD rates, but their risk values suggest that the magnitude of the overdoses in Wu and SA are larger than EWOC, which, in turn, has larger overdoses than ROLL.

Of particular concern in phase I trials is coherence of the design (Cheung, 2005), that is, whether the next patient will be given a higher dose if the current patient experiences a toxicity, and a lower dose if the current patient does not. While a theoretical investigation of the coherence of the ROLL and Hybrid designs is beyond our scope here, as an illustrative example Figure 2 lists the first five doses given by EWOC, ROLL and Hybrid 1 in the 5-FU trial setting, assuming a nontoxic response to the first dose of $x_{\min} = 140$. Note that coherence is exhibited by all three designs in this example.

4.2 A Two-Stage Design

When one may have concerns about the validity of the Bayesian parametric model in this model-based approach, one can readily incorporate the hybrid designs as the second stage of a two-stage design. The first stage of such escalates the doses cautiously by using a modified 3-plus-3 design. For the batches of 3 in the 3-plus-3 design, we propose to combine the nonparametric step-up/down approach with a parametric model-based dose determining scheme, thereby checking the parametric model to be used for model-based escalation in the second stage. This modification of the traditional 3-plus-3 design uses a specified set of dose levels (25). Set $d_1 = \lambda_1 = x_{\min}$. In the kth group of 3 patients, 2 patients are treated at the same dose

$d_k = \lambda_j$ and 1 patient at the EWOC dose m_k, computed given the doses and responses of the previous $3(k-1)$ patients. If no DLT occurs in the group of 3 patients, d_{k+1} is increased to λ_{j+1}. If 1 DLT occurs, d_{k+1} stays the same at $d_k = \lambda_j$. Otherwise, 2 or 3 DLTS have occurred, so the trial is stopped if $d_k = x_{\min}$, and otherwise continues with d_{k+1} lowered to λ_{j-1}. (Alternatively, it may be desired to stop when 3 toxicities occur, regardless of what d_k was.) The EWOC dose m_{k+1} is updated when the process is repeated with the next group of 3 patients. This process repeats until a certain fraction of the total number n of patients has been treated, provided the trial has not been stopped at the first stage due to excess toxicities. We have found from our simulation studies that switch-over points around $n/3$ or $n/4$ seem to strike a balance between enough time for conservative dose escalation and model checking during the first stage, while leaving enough time for efficient dose escalation in the second stage.

The benefit of a first stage of conservative dose escalation occurs when, unlike in Table 2, the prior distribution of the MTD is misspecified. For example, if the true MTD falls in the left tail of the prior distribution of η, then the prior information about the MTD is biased upward, which can cause overdoses. In this situation, including an initial stage of modified dose escalation, like the modified 3-plus-3 scheme, provides additional safety by refining the prior to be more accurate when it begins to be used in the second stage. Focusing on the CRM, EWOC, ROLL and Hybrid 1 designs, Table 3 contains the results of a simulation study that considers a situation such as this, where the true MTD is the lower 15th percentile of the MTD's nominal uniform prior distribution on $[x_{\min}, x_{\max}]$. That is, the data are generated with η fixed at the 15th percentile of $[x_{\min}, x_{\max}]$ and ρ uniformly distributed over

$[0, q]$, with $q = 0.2$ as in Table 2. The nominal prior for (ρ, η) used by the Bayesian procedures in Table 3 is (26), the same as in Table 2, as are the values of the other parameters. To see the effects of the first stage of more conservative dose escalation, the operating characteristics of ROLL are recomputed using a first stage of length $n/4 = 6$; the dose levels (25) used by the modified 3-plus-3 design are 10 uniformly-spaced levels in $[x_{\min}, x_{\max}] = [140, 425]$. Adding this first stage to ROLL or Hybrid 1 substantially reduces the risk, DLT and overdose rates, as shown in Table 3, in which (a) refers to the case of $n = 24$ dose levels without the modified 3-plus-3 first stage, and (b) refers to the two-stage design using a first stage of length $n/4 = 6$ consisting of the modified 3-plus-3 design.

5. CONCLUSION

Despite their shortcomings and the development of alternative Bayesian approaches since 1990, conventional dose-escalation designs are still widely used in Phase I cancer trials because of the ethical issue of safe treatment of patients currently in the trial. However, a Phase I design also has the goal of determining the MTD for a future Phase II cancer trial, and needs an informative experimental design to meet this goal. Von Hoff and Turner (1991) have documented that the overall response rates in Phase I trials are low and that substantial numbers of patients are treated at doses that are retrospectively found to be nontherapeutic. Eisenhauer et al. (2000), page 685, have pointed out that "with a plethora of molecularly defined antitumor targets and an increasingly clear description of tumor biology, there are now more antitumor candidate therapies requiring Phase I study than ever," and that "unless more efficient approaches are undertaken, Phase I

TABLE 3
Risk, bias and RMSE of the final MTD estimate, DLT rate and MTD overdose rate (OD), with SEs in parentheses, of various designs with the MTD fixed at the lower 15th percentile of the misspecified prior

Design		Risk	Bias	RMSE	DLT	OD
ROLL	(a)	1.64 (0.02)	−0.031 (0.003)	0.142 (0.025)	30.32% (1.03%)	39.38% (1.09%)
	(b)	1.39 (0.02)	−0.025 (0.003)	0.145 (0.026)	27.41% (1.00%)	33.39% (1.05%)
Hybrid 1	(a)	1.82 (0.05)	−0.032 (0.002)	0.151 (0.027)	36.90% (1.52%)	42.31% (1.56%)
	(b)	1.69 (0.04)	−0.027 (0.003)	0.131 (0.036)	35.70% (1.51%)	41.11% (1.56%)
EWOC		2.29 (0.02)	−0.034 (0.003)	0.155 (0.028)	35.33% (1.07%)	45.98% (1.11%)
CRM		3.83 (0.02)	0.037 (0.004)	0.179 (0.032)	44.18% (1.11%)	65.12% (1.07%)

trials may be a rate-limiting step in the process of evaluation of novel anticancer agents." The hybrid designs in the previous section were motivated by developing one such "more efficient" approach.

Hybrid designs with simple interpolation functions, refined through iterated rollouts and regression, can be implemented by using simple look-up tables for the parameters in (21), and thus can be relatively simple to use for clinicians. Given computer packages to compute the standard myopic and learning designs, practitioners can use a look-up table for the values ε_k in (18) as a function of the relative posterior standard deviation v_{k-1}/v_0. For given values of the prior parameters $x_{\min}, x_{\max}, p, q$ and ω, a computer package can generate this look-up table, which can be used at every stage of the trial. We are in the process of developing open source software for this purpose.

Tighiouart, Rogatko and Babb (2005) have shown how Markov chain Monte Carlo (MCMC) can be used to compute the posterior distribution of (ρ, η) when the prior distribution is supported on $[0, q] \times [x_{\min}, \infty)$, extending the model considered above where the support of η is bounded above by $x_{\max}$. They note that priors in this class with a negative correlation structure between ρ and η result in an EWOC design with comparable accuracy for estimating the MTD but lower DLT and OD rates, relative to its performance for priors supported on $[0, q] \times [x_{\min}, x_{\max}]$. As noted in Section 4, a two-stage design can easily address the higher DLT and OD rates caused by misspecifications of such priors. On the other hand, even without a cautious first stage, the above and other generalizations of the prior of (ρ, η) can be seamlessly incorporated into our hybrid design. In fact, the model M_4 of Tighiouart et al. (2005), which has been shown to perform well in their simulation studies, has a left-truncated, hierarchical normal prior distribution on η, so the rejection sampling approach in the last paragraph of Section 3.2 can be applied here by using, say, the exponential distribution as the instrumental distribution, since its tails are upper bounds of those of the normal distribution. We can therefore still use the Monte Carlo approach laid out at the end of Section 3.2.

ACKNOWLEDGMENTS

REFERENCES

ABDELBASIT, K. M. and PLACKETT, R. L. (1983). Experimental design for binary data. *J. Amer. Statist. Assoc.* **78** 90–98. MR0696852

BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statist. Med.* **17** 1103–1120.

BAYARD, D. S. (1991). A forward method for optimal stochastic nonlinear and adaptive control. *IEEE Trans. Automat. Control* **36** 1046–1053. MR1122480

BERTSEKAS, D. P. (2007). *Dynamic Programming and Optimal Control* **2**, 3rd ed. Athena Scientific, Belmont, MA.

CHEUNG, Y. K. (2005). Coherence principles in dose-finding studies. *Biometrika* **92** 863–873. MR2234191

DETTE, H., MELAS, V. B. and PEPELYSHEV, A. (2004). Optimal designs for a class of nonlinear regression models. *Ann. Statist.* **32** 2142–2167. MR2102506

EISENHAUER, E. A., O'DWYER, P. J., CHRISTIAN, M. and HUMPHREY, J. S. (2000). Phase I clinical trial design in cancer drug development. *J. Clin. Oncol.* **18** 684.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York. MR0403103

HAINES, L. M., PEREVOZSKAYA, I. and ROSENBERGER, W. F. (2003). Bayesian optimal design for phase I clinical trials. *Biometrics* **59** 591–600. MR2004264

HAN, J., LAI, T. L. and SPIVAKOVSKY, V. (2006). Approximate policy optimization and adaptive control in regression models. *Computat. Econom.* **27** 433–452.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York. MR1851606

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. *Statist. Med.* **13** 1799–1806.

LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196–1221. MR0550144

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. MR1059105

O'QUIGLEY, J. and SHEN, L. Z. (1996). Continual reassesment method: A likelihood approach. *Biometrics* **52** 163–174.

SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405. MR0098427

STORER, B. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45** 925–937. MR1029610

SUTTON, R. S. and BARTO, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

TIGHIOUART, M., ROGATKO, A. and BABB, J. S. (2005). Flexible Bayesian methods for cancer phase I clinical trials: Dose escalation with overdose control. *Stat. Med.* **24** 2183–2196. MR2146926

VON HOFF, D. and TURNER, J. (1991). Response rates, duration of response, and dose response effects in phase I studies of antineoplastics. *Investigational New Drugs* **9** 115–122.

WU, C. F. J. (1985). Efficient sequential designs with binary data. *J. Amer. Statist. Assoc.* **80** 974–984. MR0819603