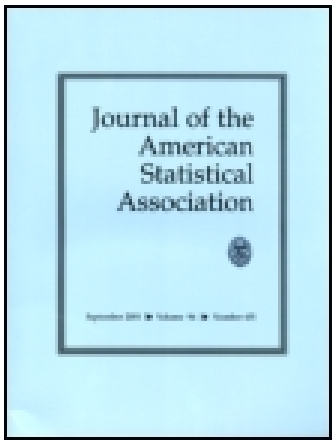

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity

Juhee Lee, Peter F. Thall, Yuan Ji & Peter Müller

Accepted author version posted online: 27 Jun 2014.Published online: 06 Jul 2015.

To cite this article: Juhee Lee, Peter F. Thall, Yuan Ji & Peter Müller (2015) Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity, Journal of the American Statistical Association, 110:510, 711-722, DOI: 10.1080/01621459.2014.926815

To link to this article: http://dx.doi.org/10.1080/01621459.2014.926815

Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity

Juhee LEE, Peter F. THALL, Yuan JI, and Peter MÜLLER

This article proposes a phase I/II clinical trial design for adaptively and dynamically optimizing each patient's dose in each of two cycles of therapy based on the joint binary efficacy and toxicity outcomes in each cycle. A dose-outcome model is assumed that includes a Bayesian hierarchical latent variable structure to induce association among the outcomes and also facilitate posterior computation. Doses are chosen in each cycle based on posteriors of a model-based objective function, similar to a reinforcement learning or Q-learning function, defined in terms of numerical utilities of the joint outcomes in each cycle. For each patient, the procedure outputs a sequence of two actions, one for each cycle, with each action being the decision to either treat the patient at a chosen dose or not to treat. The cycle 2 action depends on the individual patient's cycle 1 dose and outcomes. In addition, decisions are based on posterior inference using other patients' data, and therefore, the proposed method is adaptive both within and between patients. A simulation study of the method is presented, including comparison to two-cycle extensions of the conventional 3 + 3 algorithm, continual reassessment method, and a Bayesian model-based design, and evaluation of robustness. Supplementary materials for this article are available online.

KEY WORDS: Adaptive design; Bayesian design; Dynamic treatment regime; Latent probit model; Phase I-II clinical trial; Q-learning.

1. INTRODUCTION

Medical treatment often involves multiple cycles of therapy. Physicians routinely choose a patient's treatment in each cycle adaptively based on the patient's history of treatments and clinical outcomes. In such settings, a patient's therapy is not one treatment, but rather a sequence of treatments, each chosen using an adaptive algorithm of the general form "observe → treat → observe → treat → ⋯" etc. This paradigm is known as a dynamic treatment regime (DTR) (Lavori and Dawson 2001; Murphy, van der Laan, and Robins 2001; Murphy 2003; Moodie, Richardson, and Stephens 2007), multistage treatment strategy (Thall, Millikan, and Sung 2000; Thall, Sung, and Estey 2002) or treatment policy (Lunceford, Davidian, and Tsiatis 2002; Wahed and Tsiatis 2004). In oncology, treatment in each cycle may be a chemical or biological agent, radiation therapy, or some combination of these. DTRs also are used for chronic diseases, including behavioral disorders (Collins et al. 2005; Almirall, Ten Have, and Murphy 2010) and drug or alcohol addiction (Murphy, Collins, and Rush 2007; Murphy et al. 2007). Unfortunately, most clinical trial designs ignore the actual DTRs being used, and instead evaluate the treatments given initially as if patient outcome were due to them alone, rather than the entire DTR.

There is an extensive literature on adaptive dose-finding designs for phase I and phase I/II clinical trials (see Chevret 2006; Yin 2012). In actual conduct of such trials, the attending physician uses a DTR to make multicycle decisions for each patient. Depending on the patient's history of doses and outcomes, the dose given in each cycle may be above, below, or the same as the dose given previously, or therapy may be terminated due to excessive toxicity or poor efficacy. Since typical early-phase trial designs ignore such within-patient multicycle decision making, the "optimal" dose chosen by such a design actually pertains only to the first cycle of therapy.

While statistical methods for DTRs have seen limited application in actual clinical trials (Rush, Trivedi, and Fava 2003; Thall, et al. 2007; Wang et al. 2012), recently there has been extensive research to develop or optimize DTRs in medicine, including semiparametric methods (Wahed and Tsiatis 2006), reinforcement learning (Zhao et al. 2011), and sequential multiple assignment randomized trials (Murphy and Bingham 2009). The aim is to better reflect the intrinsically multistage, adaptive structure of what physicians actually do, in both trial design and analysis of observational data. This methodology had its origins in research to define and estimate causal parameters in complex longitudinal data, pioneered by Robins (1986, 1993, 1997, 1998), and applied to the analysis of AIDS data (Hernan, Brumback, and Robins 2000; Robins, Hernan, and Brumback 2000).

The problem of optimizing each patient's doses given in multiple cycles based on efficacy and toxicity in phase I/II trials has not been addressed formally. Phase I/II designs typically optimize the initial dose using between-patient adaptive rules. A review is given by Zohar and Chevret (2007). For phase I trials involving multiple cycles of therapy, Braun, Yuan, and Thall (2005) proposed a Bayesian design with between-patient adaptive rules based on time-to-toxicity to optimize the number of cycles ("schedule") given a fixed dose. Braun et al. (2007) extended this to allow per-administration dose to vary, and jointly optimized dose and schedule, using a criterion similar to that of the time-to-event continual reassessment method (TiTE CRM, Cheung, and Chappell 2000). Li et al. (2008) proposed an approach to optimizing dose and schedule for two nested schedules and bivariate binary outcomes, using an isotonic transformation to obtain matrix ordered toxicity probabilities with

Juhee Lee, Department of Statistics, Ohio State University, Columbus, OH 43210 (E-mail: *juheelee2@gmail.com*). Peter F. Thall, Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, Houston, TX 77230-1402 (E-mail: *rex@mdanderson.org*). Yuan Ji, Center for Clinical and Research Informatics, North Shore University Health System, Evanston, IL 60201 (E-mail: *yji@northshore.org*). Peter Müller, Department of Mathematics, University of Texas, Austin, TX 78712 (E-mail: *pmueller@math.utexas.edu*). Yuan Ji's research is supported by NIH R01/NCI CA132897. Peter Thall's research was supported by NIH/NCI R01 CA 83932 and NIH/NCI 5 P50 CA140388. Peter Müller's research was supported by NIH/NCI R01 CA157458-01A1.

DOI: 10.1080/01621459.2014.926815

order-restricted inferences. While few of these methods include within-patient adaptive rules applied after the first cycle, the phase I design proposed by Zhang and Braun (2013) to optimize dose and schedule accounts for multiple within-patient administrations.

Here, we address the problem of adaptively optimizing each patient's dose in each of two cycles of therapy in a phase I/II trial based on binary efficacy and toxicity. This is the simplest case of the general multicycle phase I/II trial design problem, which may be formulated with ordinal or time-to-event outcomes and an arbitrary number of cycles. We address the simpler two-cycle problem because it still is much more complicated than the one-cycle case. Our goals are to provide a practical trial design and establish a basis for subsequently developing methods for more complex settings. We employ a model-based Bayesian objective function, defined in terms of (efficacy, toxicity) utilities, structurally similar to reinforcement learning (Sutton and Bartow 1998) or Q-learning functions (Watkins 1989). Our method chooses a dose in each cycle to maximize the posterior expected mean of the objective function, applying a modified recursive Bellman equation (1957) that assumes, for the decision in cycle 1, that one will behave optimally in cycle 2. At the end of the trial, the method provides an optimal two-stage regime consisting of an optimal cycle 1 dose, and an optimal function of the patient's cycle 1 dose and outcomes that either chooses a cycle 2 dose or says to not treat the patient in cycle 2. This is very different from simply choosing two "optimal" doses, one for each cycle, with the "optimal" cycle 2 dose ignoring each patient's cycle 1 data. Because all decisions are based on posterior quantities computed using all patients' data, the method is adaptive both within and between patients.

Section 2 describes the proposed decision-theoretical two-cycle method, DTM2, including the Bayesian probability model, an algorithm for prior calibration, and posterior computation. Utility-based decision criteria are presented in Section 3. A simulation study is summarized in Section 4. We close with a discussion in Section 5.

2. DOSE-OUTCOME MODEL

The model used by DTM2 exploits the idea underlying the multivariate probit model, introduced by Ashford and Sowden (1970). A vector of unobserved, correlated latent multivariate normal variables is defined to induce association among a vector of observed binary variables, by defining each observed variable as the indicator that its corresponding latent variable is greater than 0. The DTM2 model is an elaboration of a multivariate probit model that includes hierarchical structures. It provides a computationally feasible basis for the task at hand. We will exploit the MCMC methods for computing posteriors for latent variable models provided by Albert and Chib (1993) and developed further by Chib and Greenberg (1998) for posterior computation via Gibbs sampling.

Let n_t denote the number of patients accrued and given at least one cycle of treatment up to trial (calendar) time t, and index patients by $i = 1, \ldots, n_t$. Our dose-outcome model does not depend on numerical dose values, and we identify the doses under consideration by the indexes $1, \ldots, m$. For treatment cycle $c = 1, 2$, denote the ith patient's dose by $d_{i,c}$, outcome indicators $Y_{i,c} \in \{0, 1\}$ for toxicity and $Z_{i,c} \in \{0, 1\}$ for efficacy, and the 2-cycle vectors $\boldsymbol{d}_i = (d_{i,1}, d_{i,2})$, $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2})$, and $\boldsymbol{Z}_i = (Z_{i,1}, Z_{i,2})$. Let $X_t = \{(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{d}_i) : i = 1, \ldots, n_t\}$ denote the observed data from all patients at t. Although the doses $\boldsymbol{d}_i$ are actions rather than parameters or random outcomes, throughout the article we will abuse probability notation slightly by including them to the right of the conditioning bar. Since actual clinical decision rules must allow a given patient's therapy to be terminated, for example, if the patient is cured, has progressive disease, or unacceptable toxicity (see Wang et al. 2012), here possible actions in cycle c may be either a dose, $d_{i,c}$, or the decision to give no treatment, which we index by 0. We denote the possible actions in either cycle by $D = \{0, 1, \ldots, m\}$.

We construct a joint distribution for $[Y_i, Z_i|d_i]$ by defining these binary outcomes in terms of four real-valued latent variables, $\boldsymbol{\xi}_i = (\xi_{i,1}, \xi_{i,2})$ for $\boldsymbol{Y}_i$ and $\boldsymbol{\eta}_i = (\eta_{i,1}, \eta_{i,2})$ for $\boldsymbol{Z}_i$, with $(\boldsymbol{\xi}_i, \boldsymbol{\eta}_i)$ following a multivariate normal distribution having means that vary with $\boldsymbol{d}_i$. Denoting the indicator of event A by $\mathrm{I}(A)$, we assume $Y_{i,c} = \mathrm{I}(\xi_{i,c} > 0)$ and $Z_{i,c} = \mathrm{I}(\eta_{i,c} > 0)$, so the distribution of $[\boldsymbol{Y}_i, \boldsymbol{Z}_i|\boldsymbol{d}_i]$ is induced by that of $[\boldsymbol{\xi}_i, \boldsymbol{\eta}_i|\boldsymbol{d}_i]$. The structure of our hierarchical model for two cycles is similar to the nonhierarchical model for multiple toxicities in one cycle of therapy used by Bekele and Thall (2004). To construct the model, we first define a conditional likelihood for the cycle-specific latent variable pairs $[\xi_{i,c}, \eta_{i,c}|d_{i,c}]$, for $c = 1, 2$ by using patient-specific random effects (u_i, v_i) that characterize dependence among the outcomes between and within cycles. Denote the univariate normal distribution with mean μ and variance σ^2 by $N(\mu, \sigma^2)$, with pdf $\phi(\cdot|, \mu, \sigma^2)$.

We begin the construction by assuming the following Level 1 and Level 2 priors:

Level 1 Priors on the Latent Variables. For patient i in cycle c given dose $d_{i,c} = d$,

$$\begin{aligned} &\xi_{i,c}|u_i, \bar{\xi}_{c,d}, \sigma_\xi^2 \sim \mathrm{N}(\bar{\xi}_{c,d} + u_i,\ \sigma_\xi^2) \\ &\quad \text{and} \quad \eta_{i,c} \mid v_i, \bar{\eta}_{c,d}, \sigma_\eta^2 \sim \mathrm{N}(\bar{\eta}_{c,d} + v_i,\ \sigma_\eta^2), \end{aligned} \tag{1}$$

with $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$ conditionally independent given (u_i, v_i) and fixed σ_ξ^2 and σ_η^2. Level 2 priors of the patient effects, (u_i, v_i), and mean cycle-specific dose effects, $(\bar{\xi}_{c,d}, \bar{\eta}_{c,d})$, are as follows:

Level 2 Priors on (u_i, v_i). For patients $i = 1, \ldots, n$,

$$u_i, v_i|\rho,\ \tau^2 \overset{\text{iid}}{\sim} \mathrm{MVN}_2(\mathbf{0}_2, \Sigma_{u,v}), \tag{2}$$

where MVN_k denotes a k-variate normal distribution, $\mathbf{0}_2 = (0, 0)$ and $\Sigma_{u,v}$ is the 2×2 matrix with all diagonal elements τ^2 and all off-diagonal elements $\rho\tau^2$. The hyperparameters, $\rho \in (-1, 1)$ and τ^2, are fixed. This Level 2 prior induces association, parameterized by (ρ, τ^2), among $(\xi_{i,1}, \eta_{i,1}, \xi_{i,2}, \eta_{i,2})$ via the latent variable model (1), and thus among the corresponding toxicity and efficacy outcomes, $(Y_{i,1}, Z_{i,1}, Y_{i,2}, Z_{i,2})$.

Level 2 Priors on $(\bar{\xi}_{c,d},\ \bar{\eta}_{c,d})$. Let $\bar{\boldsymbol{\xi}}_c = (\bar{\xi}_{c,1}, \ldots, \bar{\xi}_{c,m})$ and $\bar{\boldsymbol{\eta}}_c = (\bar{\eta}_{c,1}, \ldots, \bar{\eta}_{c,m})$. Denote by $\bar{\boldsymbol{\xi}}_{c,-d}$ the vector $\bar{\boldsymbol{\xi}}_c$ with $\bar{\xi}_{c,d}$ deleted, and let $\bar{\boldsymbol{\eta}}_{c,-d}$ denote $\bar{\boldsymbol{\eta}}_c$ with $\bar{\eta}_{c,d}$ deleted.

We assume

$$p(\bar{\xi}_{c,d}|\bar{\boldsymbol{\xi}}_{c,-d}) \propto \phi(\bar{\xi}_{c,d}|\xi_{c,0}, \sigma^2_{\xi_{c,0}})1(\bar{\xi}_{c,d-1} < \bar{\xi}_{c,d} < \bar{\xi}_{c,d+1})$$
$$p(\bar{\eta}_{c,d}|\bar{\boldsymbol{\eta}}_{c,-d}) \propto \phi(\bar{\eta}_{c,d}|\eta_{c,0}, \sigma^2_{\eta_{c,0}})1(\bar{\eta}_{c,d-1} < \bar{\eta}_{c,d} < \bar{\eta}_{c,d+1}). \quad (3)$$

The order constraints ensure that $\xi_{i,c}$ and $\eta_{i,c}$ increase stochastically in dose, hence the per-cycle probabilities of toxicity and efficacy both increase with dose. If this assumption is not appropriate, such as trials of biologic agents, these constraints may be dropped.

Collecting terms from (1), (2), and (3), the 12 fixed parameters that determine all of the Level 1 and Level 2 priors are $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\xi}_0, \boldsymbol{\eta}_0, \boldsymbol{\sigma}^2_{\xi_0}, \boldsymbol{\sigma}^2_{\eta_0}, \sigma^2_\xi, \sigma^2_\eta, \tau^2, \rho)$ where $\boldsymbol{\xi}_0 = (\xi_{1,0}, \xi_{2,0})$, $\boldsymbol{\eta}_0 = (\eta_{1,0}, \eta_{2,0})$, $\boldsymbol{\sigma}^2_{\xi_0} = (\sigma^2_{\xi_{1,0}}, \sigma^2_{\xi_{2,0}})$, and $\boldsymbol{\sigma}^2_{\eta_0} = (\sigma^2_{\eta_{1,0}}, \sigma^2_{\eta_{2,0}})$. Denote $\bar{\boldsymbol{\xi}} = (\bar{\boldsymbol{\xi}}_1, \bar{\boldsymbol{\xi}}_2), \bar{\boldsymbol{\eta}} = (\bar{\boldsymbol{\eta}}_1, \bar{\boldsymbol{\eta}}_2), \boldsymbol{\mu}_{\boldsymbol{d}_i} = (\bar{\xi}_{1,d_{i,1}}, \bar{\xi}_{2,d_{i,2}}, \bar{\eta}_{1,d_{i,1}}, \bar{\eta}_{2,d_{i,2}})$, and the covariance matrix

$$\Sigma_{\xi,\eta} = \begin{bmatrix} \sigma^2_\xi + \tau^2 & \tau^2 & \rho\tau^2 & \rho\tau^2 \\ & \sigma^2_\xi + \tau^2 & \rho\tau^2 & \rho\tau^2 \\ & & \sigma^2_\eta + \tau^2 & \tau^2 \\ & & & \sigma^2_\eta + \tau^2 \end{bmatrix}.$$

The joint distribution of $[\boldsymbol{\xi}_i, \boldsymbol{\eta}_i|\boldsymbol{d}_i, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}]$ is computed by integrating over (u_i, v_i), yielding

$$\boldsymbol{\xi}_i, \boldsymbol{\eta}_i|\boldsymbol{d}_i, \bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}} \overset{\text{iid}}{\sim} \text{MVN}_4\left(\boldsymbol{\mu}_{\boldsymbol{d}_i}, \Sigma_{\xi,\eta}\right). \quad (4)$$

The mean vector $\boldsymbol{\mu}_{\boldsymbol{d}}$ is a function of the dose levels, and does not depend on numerical dose values. The hyperparameters, τ^2 and ρ, induce associations between cycle 1 and cycle 2 and between efficacy outcomes and toxicity outcomes. For example, if $-1 < \rho < 0\,(0 < \rho < 1)$, this model implies that efficacy and toxicity are negatively (positively) associated, that is, higher (lower) toxicity is associated with lower efficacy.

Denote $\boldsymbol{\theta} = (\bar{\boldsymbol{\xi}}, \bar{\boldsymbol{\eta}})$. Integrating over (u_i, v_i) and suppressing $\tilde{\boldsymbol{\theta}}$ and patient index i, the joint likelihood for the observables of a patient is given by

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{d}, \boldsymbol{\theta}) &= \Pr(Y_1 = y_1, Y_2 = y_2, Z_1 = z_1, Z_2 = z_2|\boldsymbol{d}, \boldsymbol{\theta}) \\ &= \Pr(\gamma_{1,y_1} \le \xi_1 < \gamma_{1,y_1+1}, \gamma_{1,y_2} \le \xi_2 < \gamma_{1,y_2+1}, \\ &\qquad \gamma_{2,z_1} \le \eta_1 < \gamma_{2,z_1+1}, \gamma_{2,y_2} \le \eta_2 < \gamma_{2,z_2+1}|\boldsymbol{d}, \boldsymbol{\theta}) \\ &= \int_{\gamma_{1,y_1}}^{\gamma_{1,y_1+1}} \int_{\gamma_{1,y_2}}^{\gamma_{1,y_2+1}} \int_{\gamma_{2,z_1}}^{\gamma_{2,z_1+1}} \int_{\gamma_{2,z_2}}^{\gamma_{2,z_2+1}} \\ &\qquad \phi(\boldsymbol{\xi}, \boldsymbol{\eta}|\boldsymbol{\mu}_{\boldsymbol{d}}, \Sigma_{\xi,\eta})\mathrm{d}\eta_2\mathrm{d}\eta_1\mathrm{d}\xi_2\mathrm{d}\xi_1, \end{aligned}$$

where the cutoff vectors $(\gamma_{10}, \gamma_{11}, \gamma_{12})$ for Y_c and $(\gamma_{20}, \gamma_{21}, \gamma_{22})$ for Z_c both are $(-\infty, 0, \infty)$, for $c = 1, 2$. The conditional distribution of the cycle 2 outcomes (Y_2, Z_2) given the cycle 1 outcomes $(Y_1 = y_1, Z_1 = z_1)$ is

$$\begin{aligned} p(y_2, z_2|y_1, z_1, \boldsymbol{d}, \boldsymbol{\theta}) &= \Pr(Y_2 = y_2, Z_2 = z_2 \\ &\qquad |Y_1 = y_1, Z_1 = z_1, \boldsymbol{d}) \\ &= \Pr(\gamma_{1y_2} \le \xi_2 < \gamma_{1,y_2+1}, \gamma_{2z_2} \le \eta_2 \\ &< \gamma_{2,z_2+1}|\gamma_{1y_1} \le \xi_1 < \gamma_{1,y_1+1}, \gamma_{2,z_1} \\ &\le \eta_{i1} < \gamma_{2,z_1+1}, \boldsymbol{d}) \\ &= \frac{p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{d}, \boldsymbol{\theta})}{p(y_1, z_1|d_1, \boldsymbol{\theta})}, \end{aligned} \quad (5)$$

where the cycle 1 bivariate marginal is computed as the double integral

$$p(y_1, z_1|d_1, \boldsymbol{\theta}) = \int_{\gamma_{1y_1}}^{\gamma_{1,y_1+1}} \int_{\gamma_{2z_1}}^{\gamma_{2,z_1+1}} \phi([\xi_1, \eta_1]|\boldsymbol{\mu}^1_{d_1}, \Sigma^1_{\xi,\eta})\mathrm{d}\eta_1\mathrm{d}\xi_1 \quad (6)$$

with

$$\boldsymbol{\mu}^1_{d_1} = \begin{bmatrix} \bar{\xi}_{1,d_1} \\ \bar{\eta}_{1,d_1} \end{bmatrix} \text{ and } \Sigma^1_{\xi,\eta} = \begin{bmatrix} \sigma^2_\xi + \tau^2 & \rho\tau^2 \\ \rho\tau^2 & \sigma^2_\eta + \tau^2 \end{bmatrix}.$$

3. DECISION CRITERIA

3.1 Adaptive Dose Selection

To define our decision rules, we distinguish between doses and actions. The action in cycle 1 either chooses a dose from the set $\{1, \ldots, m\}$ of doses under consideration or makes the decision to not give the patient any treatment. Recall that we denote this decision by 0 for convenience, and we will denote the possible actions by $D = \{0, 1, \ldots, m\}$. If the optimal cycle 1 action is $d_1 = 0$ at any point in the trial then the study is terminated. Otherwise, the patient receives d_1 for cycle 1 and $d_2 \in D$ for cycle 2, where d_2 is a function of the cycle 1 dose and outcomes, (d_1, Y_1, Z_1), and the current data, X, from all patients. For example, if the cycle 1 dose d_1 produced toxicity, $Y_1 = 1$, then a possible cycle 2 action is $d_2(d_1, 1, 1, X) = d_1 - 1$ if $Z_1 = 1$, and $d_2(d_1, 1, 0, X) = 0$ if $Z_1 = 0$. Similarly, if $d_1 = 1$, the lowest dose level, and $Y_1 = 1$ was observed, then it may be that $d_2(d_1, 1, Z_1, X) = 0$ regardless of whether $Z_1 = 0$ or 1. In general, a two-cycle regime is far more general than a dose pair chosen from $D \times D$, and a regime for which d_2 ignores the patient's cycle 1 dose and outcomes, (d_1, Y_1, Z_1), is unlikely to be optimal. In the DTR literature, (d_1, Y_1, Z_1) would be called "tailoring variables." Optimizing $\boldsymbol{d} = (d_1, d_2)$ is the focus of our design.

3.2 Objective Function

We construct an objective function by using the basic ideas in Bellman (1957), starting in cycle 2 and working backward. Our method relies on per-cycle utilities $U(y, z)$ that quantify the desirability of outcome $(Y_c, Z_c) = (y, z)$ in cycle $c = 1$ or 2. Depending on the level of marginalization and aggregation over cycles and patients, many variations of the objective function defined below may be obtained. We will generically refer to all of these as "utility" or "objective function" when we want to highlight that a particular expected utility is a function of known quantities and the action only, and thus can be used to select the optimal action. For convenience, one may fix $U(0, 1) = 100$ and $U(1, 0) = 0$, which are the respective utilities for the best and worst possible outcomes, and elicit the intermediate values $U(0, 0)$ and $U(1, 1)$ from the physicians planning the trial, although any function with $U(1, 0) < U(1, 1), U(0, 0) < U(0, 1)$ may be used. In our simulations, we will use the numerical utilities $U(1, 0) = 0, U(0, 0) = 35, U(1, 1) = 65, U(0, 1) = 100$.

In the language of Q-learning (Watkins 1989; Murphy 2005; Zhao et al. 2011), for cycle c, d_c is the "action" and $U(Y_c, Z_c)$ is the "reward," with (d_1, Y_1, Z_1) the "state" prior to taking action d_2 in cycle 2. Ideally, baseline covariates such as age, disease severity, or performance status would comprise the patient's state for $c = 1$, although in practice even in the single-cycle

phase I-II setting choosing covariate-specific doses is quite complicated (see Thall, Nguyen, and Estey 2008).

Given a patient's cycle 1 data (d_1, Y_1, Z_1), the mean utility of action d_2 in cycle 2 is

$$\begin{aligned} Q_2(d_2, d_1, Y_1, Z_1, \boldsymbol{\theta}) &= E\{U(Y_2, Z_2)|d_2, d_1, Y_1, Z_1, \boldsymbol{\theta}\} \\ &= \sum_{y_2=0}^{1}\sum_{z_2=0}^{1} U(y_2, z_2)p(y_2, z_2 \\ &\quad | \, d_2, d_1, Y_1, Z_1, \boldsymbol{\theta}), \end{aligned} \tag{7}$$

and we define the cycle 2 objective function

$$\begin{aligned} q_2(d_2, d_1, Y_1, Z_1, X) = E\{&Q_2(d_2, d_1, Y_1, Z_1, \boldsymbol{\theta}) \\ &| \, d_2, d_1, Y_1, Z_1, X\}. \end{aligned} \tag{8}$$

If $d_2 = 0$, that is, no treatment in cycle 2, then $p(Y_2 = 0, Z_2 = 0 \mid d_2 = 0, d_1, Y_1, Z_1, \boldsymbol{\theta}) = 1$ and $q_2(d_2 = 0, d_1, Y_1, Z_1, X) = U(0, 0)$, the utility of having neither toxicity nor efficacy. If $d_2 \neq 0$, then $q_2(d_2, d_1, Y_1, Z_1, X)$ is a posterior expected utility of giving dose d_2 in cycle 2 given (d_1, Y_1, Z_1). This underscores the importance of requiring $U(0, 0) > U(1, 0)$, that it is more desirable to have neither toxicity nor efficacy than to have toxicity and no efficacy. Given (d_1, Y_1, Z_1) and current data X, the optimal cycle 2 action, $d_2^{\text{opt}}(d_1, Y_1, Z_1, X) = \text{argmax}_{d_2} q_2(d_2, d_1, Y_1, Z_1, X)$, subject to dose acceptability rules discussed in Section 3.3.

Next, we move backward to the cycle 1 optimization assuming that $q_2(d_2^{\text{opt}}, d_1, Y_1, Z_1, X)$ is known for all (d_1, Y_1, Z_1). The expected utility of giving dose d_1 given $\boldsymbol{\theta}$ is

$$\begin{aligned} Q_1(d_1, \boldsymbol{\theta}) &= E\{U(Y_1, Z_1)|d_1, \boldsymbol{\theta}\} \\ &= \sum_{y_1=0}^{1}\sum_{z_1=0}^{1} U(y_1, z_1)p(y_1, z_1|d_1, \boldsymbol{\theta}). \end{aligned}$$

To define the overall objective function, we discount the cycle 2 payoff using the fixed parameter $0 < \lambda < 1$, as is done traditionally in Q-learning. The expected entire future utility of giving dose d_1 in cycle 1, assuming that d_2^{opt} will be taken in cycle 2, is

$$\begin{aligned} q_1(d_1, \boldsymbol{X}) &= E\big[E\{U(Y_1, Z_1) + \lambda q_2\big(d_2^{\text{opt}}(d_1, Y_1, Z_1, X), \\ &\qquad d_1, Y_1, Z_1, X\big)|\boldsymbol{\theta}, d_1\}|d_1, X\big] \\ &= E\{Q_1(d_1, \boldsymbol{\theta})|d_1, X\} \\ &\quad + \lambda \sum_{y_1=0}^{1}\sum_{z_1=0}^{1} q_2(d_2^{\text{opt}}(d_1, y_1, z_1, X), \\ &\qquad d_1, y_1, z_1, X)p(y_1, z_1|d_1, X), \end{aligned} \tag{9}$$

where $p(y_1, z_1|d_1, X)$ is the posterior expected density for (y_1, z_1). Letting $q_1(d_1, X) = (1 + \lambda)U(0, 0)$ for $d_1 = 0$, the optimal cycle 1 action, d_1^{opt}, maximizes this quantity over D.

Maximizing q_1 and q_2 yields the optimal actions $\boldsymbol{d}^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}})$, where d_1^{opt} is either a dose or 0, d_2^{opt} is applicable only when $d_1^{\text{opt}} \neq 0$, d_2^{opt} is a function of $(d_1^{\text{opt}}, Y_1, Z_1)$, and both are functions of X. If new data from other patients are obtained between administration of d_1^{opt} and optimization of $q_2(d_2(d_1), X)$, so X changes while waiting to evaluate the patient's cycle 1 outcomes (Y_1, Z_1), then the posterior and hence the patient's d_2^{opt} might change. This may be made precise by elaborating the notation to account for relationships between timing of the patient's cycles and calendar time. We avoid this complexity since the point is clear.

3.3 Dose Acceptability

We include dose acceptability criteria, motivated by ethical considerations, since maximizing a posterior utility-based objective function, per se, is not enough to allow a dose to be administered. The problem is that, while the optimal policy under a given utility function is mathematically well-defined, it is only an indirect solution of an optimization in expectation. An important case is that where no dose is acceptably safe and efficacious in either cycle 1 or cycle 2, consequently it is not ethical to treat a patient using any dose and the trial must be stopped. Moreover, in some applications, the decision-theoretical solution might turn out to have undesirable features not anticipated when specifying the outcomes, model, and utility function. This problem is one reason why many physicians are reluctant to use formal decision-theoretical methods for clinical decision making. Spiegelhalter et al. (2004, chap. 3.14) discuss this issue. We mitigate these concerns by adding three additional dose acceptability criteria that restrict the set of solutions when maximizing (8) and (9).

The first constraint is that an untried dose level may not be skipped when escalating. This says that one does not fully trust decisions based on any assumed model and decision criteria, especially with the small amounts of data available early in the trial. Let d_1^M denote the highest dose among those that have been tried in cycle 1 and d_2^M the highest dose among those that have been tried in either cycle. The search for optimal actions is constrained so that $1 \leq d_1 \leq \min(d_1^M + 1, m)$ and $1 \leq d_2 \leq \min(d_2^M + 1, m)$. The second constraint does not allow escalating a patient's dose in cycle 2 if toxicity was observed in cycle 1, $Y_1 = 1$. The third criterion, defined in terms of expected utility, is to avoid giving undesirable dose pairs. For cycle 2, we say that *action d_2 is unacceptable* if it violates the no-skipping rule, escalates after $Y_1 = 1$, or $q_2(d_2, d_1, Y_1, Z_1, X) < U(0, 0)$, that is, the posterior expected utility of treating the patient with d_2 given (d_1, Y_1, Z_1, X) is smaller than that obtained by not treating the patient at all. We denote the set of acceptable cycle 2 doses for a patient with cycle 1 data (d_1, Y_1, Z_1) by $A_2(d_1, Y_1, Z_1, X)$. Thus, a given d_2 may be acceptable for some (d_1, Y_1, Z_1) but not acceptable for others.

Table 1 illustrates true expected cycle 2 utilities of d_2 conditional on (d_1, Y_1, Z_1) using simulation Scenario 4, discussed below in Section 4. Assume that $\boldsymbol{\theta}^{\text{true}}$ and $\tilde{\boldsymbol{\theta}}^{\text{true}}$ are known, and suppress $\tilde{\boldsymbol{\theta}}^{\text{true}}$. The $Q_2(d_2, d_1, Y_1, Z_1, \boldsymbol{\theta}^{\text{true}})$ in Table 1 is similar to (7). For example, the values of $Q_2(a_2, d_1 = 3, Y_1 = 0, Z_1 = 0, \boldsymbol{\theta}^{\text{true}})$ given in the first row of the third box from the top are (35.98, 39.49, 39.70, 36.30, 27.17) for $d_2 =$(1, 2, 3, 4, 5), respectively. Since $Q_2(5, 3, 0, 0, \boldsymbol{\theta}^{\text{true}}) < U(0, 0)$, $d_2 = 5$ is not acceptable. The other dose levels are acceptable, so $A_2(3, 0, 0, \boldsymbol{\theta}^{\text{true}}) = \{1,2,3,4\}$, with $d_2^{\text{opt}}(d_1 = 3, Y_1 = 0, Z_1 = 0, \boldsymbol{\theta}^{\text{true}}) = 3$. When $(d_1, Y_1, Z_1) = (3, 1, 0)$, no $d_2 \in \{1, \ldots, m\}$ produces expected utility greater than $U(0, 0)$, and $d_2 = 3, 4, 5$ are not allowed due to the no-escalation-after toxicity rule. Thus, $A_2(3, 0, 0, \boldsymbol{\theta}^{\text{true}})$ is the empty set and $d_2^{\text{opt}}(d_1 = 3, Y_1 = 1, Z_1 = 0, \boldsymbol{\theta}^{\text{true}}) = 0$. The last column

Table 1. True expected utilities in Scenario 4 assuming that $\boldsymbol{\theta}^{\text{true}}$ is known. $q_1(d_1, \boldsymbol{\theta}^{\text{true}})$ = the true expected total utility. Entries under d_2 in columns 4–8 are the true expected cycle 2 utilities, $q_2(d_2, \boldsymbol{\theta}^{\text{true}})$

			d_2					
d_1	$q_1(d_1, \boldsymbol{\theta}^{\text{true}})$	(Y_1, Z_1)	1	2	3	4	5	d_2^{opt}
1	66.19	(0,0)	37.32	41.40	41.85	38.54	*29.66*	3
		(0,1)	48.18	55.07	56.80	54.00	45.04	3
		(1,0)	**30.60**	**33.58**	**33.07**	**29.57**	**23.27**	*NT*
		(1,1)	**41.20**	**47.16**	**47.96**	**44.80**	**38.23**	*NT*
2	73.38	(0,0)	36.53	40.33	40.66	37.30	*28.38*	3
		(0,1)	45.06	51.39	52.92	50.08	41.15	3
		(1,0)	*30.13*	**32.88**	**32.26**	**28.70**	**22.28**	*NT*
		(1,1)	38.36	**43.71**	**44.31**	**41.12**	**34.54**	1
3	80.61	(0,0)	35.98	39.49	39.70	36.30	*27.17*	3
		(0,1)	43.31	49.20	50.62	47.73	38.69	3
		(1,0)	*30.29*	*32.82*	**32.12**	**28.43**	**21.53**	*NT*
		(1,1)	37.34	42.27	**42.75**	**39.46**	**32.48**	2
4	69.71	(0,0)	37.17	40.90	41.40	38.19	*28.65*	3
		(0,1)	44.37	50.46	52.14	49.47	40.09	3
		(1,0)	*32.13*	*34.89*	*34.39*	**30.68**	**22.91**	*NT*
		(1,1)	39.17	44.32	45.02	**41.74**	**33.91**	3
5	68.16	(0,0)	38.08	42.02	42.80	39.83	*30.06*	3
		(0,1)	45.26	51.52	53.46	51.04	41.51	3
		(1,0)	*33.00*	35.88	35.52	*31.85*	**23.69**	2
		(1,1)	40.05	45.32	46.16	42.95	**34.75**	3

NOTE: Expected utilities in bold are those for d_2 violating the no-escalation-after-$Y_1 = 1$ rule. Expected utilities in *italics* are those for unacceptable d_2 based on the utility-based criterion.

of the table lists $d_2^{\text{opt}}(d_1, Y_1, Z_1, \boldsymbol{\theta}^{\text{true}})$ for all combinations of (d_1, Y_1, Z_1).

To identify acceptable cycle 1 dose levels, we assume that $d_2^{\text{opt}}(d_1, Y_1, Z_1, X)$ is chosen from $A_2(d_1, Y_1, Z_1, X)$. For cycle 1, we say that *action* d_1 *is unacceptable* if it violates the no-skipping rule or satisfies the utility-based criterion

$$q_1(d_1, X) < U(0, 0) + \lambda U(0, 0). \tag{10}$$

This says that d_1 is unacceptable in cycle 1 if it yields a smaller posterior expected utility than not treating the patient. We denote the set of acceptable cycle 1 doses by $A_1 \subset D$. Note that, while $A_1(X)$ is adaptive between patients since it is a function of other patients' data, $A_2(d_1, Y_1, Z_1, X)$ is adaptive both between and within patients.

The second column of Table 1 illustrates true expected total utilities over two cycles under simulation Scenario 4. Assuming that $\boldsymbol{\theta}^{\text{true}}$ are known, the column gives values of

$$E\{U(Y_1, Z_1) + \lambda Q_2(d_2^{\text{opt}}(d_1, Y_1, Z_1, \boldsymbol{\theta}^{\text{true}}), d_1, Y_1, Z_1, \boldsymbol{\theta}^{\text{true}}) | \boldsymbol{\theta}^{\text{true}}, d_1\},$$

where $d_2^{\text{opt}}(d_1, Y_1, Z_1, \boldsymbol{\theta}^{\text{true}})$ can be derived in the last column of the table. The true expected total utility satisfies (10) for all the $d_1 \in D$, hence all d_1 are acceptable. From the table, the optimal pair of actions is $d_1^{\text{opt}} = 3$ and $d_2^{\text{opt}} = 3, 3, 0$, and 2 for $(Y_1, Z_1) = (0, 0), (0, 1), (1, 0)$, and $(1, 1)$, respectively, listed in the fourth row of Table 3.

3.4 Adaptive Randomization

While $\boldsymbol{d}^{\text{opt}}$ yields the best clinical outcomes, the reliability of the process over the entire trial can be improved by including adaptive randomization (AR) among $\boldsymbol{d}$ giving values of the objective function near the maximum at $\boldsymbol{d}^{\text{opt}}$. While this may seem counterintuitive, using AR decreases the probability of getting stuck at a suboptimal $\boldsymbol{d}$ and also has the effect of treating more patients at doses having larger utilities, on average. The problem that a "greedy" search algorithm may get stuck at suboptimal actions, and the simple solution of introducing some additional randomness into the search process, have been known for years in the optimization literature (see Tokic 2010). However, this has been dealt with only very recently in dose-finding (Bartroff and Lai 2010; Azriel, Mandel, and Rinott 2011; Braun, Kang, and Taylor 2012; Thall and Nguyen 2012).

To implement AR, we first define ϵ_i to be a function decreasing in patient index i, and denote $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. We define the set of ϵ_i*-optimal doses for cycle 1* to be

$$D_{i,1} = \{d_1 : |q_1(d_{1,i}^{\text{opt}}, X) - q_1(d_1, X)| < \epsilon_i, d_1 \in A_{i,1}(X)\}.$$

The set $D_{i,1}(X)$ contains d_1 in $A_{i,1}(X)$ having posterior mean utility within ϵ_i of the maximum posterior mean utility. Similarly, we define the set of $(\epsilon_i/2)$-optimal doses for cycle 2 given $(d_{i,1}, Y_{i,1}, Z_{i,1})$ to be

$$\begin{aligned} D_{i,2} = \{d_2 : &|q_2(d_{i,2}^{\text{opt}}(d_{i,1}, Y_{i,1}, Z_{i,1}, X), Y_{i,1}, Z_{i,1}, d_{i,1}, X) \\ &- q_2(d_2, d_{i,1}, Y_{i,1}, Z_{i,1}, X)| < \epsilon_i/2, \\ &d_2 \in A_{i,2}(d_{i,1}, Y_{i,1}, Z_{i,1}, X)\}. \end{aligned}$$

We use $\epsilon_i/2$ because $q_2(d_2, d_1, Y_1, Z_1, X)$ is the posterior expected utility for cycle 2 only. For cycles $c = 1, 2$, patients are randomized fairly among the doses in $D_{i,c}$, which we call AR($\boldsymbol{\epsilon}$). In practice, the numerical values of ϵ_i depend on the numerical range of $U(y, z)$, and must be determined by preliminary trial simulations.

3.5 Trial Design

Our illustrative trial studied in the simulations is constructed to mimic a typical phase I-II chemotherapy trial with five dose levels, but accounting for two cycles of therapy. The maximum sample size is $n = 60$ patients with a cohort size of 2. Based on preliminary simulations, we set $\epsilon_i = 20$ for the first 10 patients, $\epsilon_i = 15$ for the next 10 patients, and $\epsilon_i = 10$ for the remaining 40 patients. An initial cohort of 2 patients is treated at the lowest dose level in cycle 1, their cycle 1 toxicity and efficacy outcomes are observed, the posterior of $\boldsymbol{\theta}$ is computed, and actions are taken for cycle 2 of the initial cohort. Posterior computations are described in the supplementary material. If $D_{i,2} = \{0\}$, then patient i does not receive a second cycle of treatment. If $D_{i,2} \neq \{0\}$, then AR($\boldsymbol{\epsilon}$) is used to choose an action for cycle 2 from $D_{i,2}$. When (Y_2, Z_2) are observed from cycle 2, the posterior of $\boldsymbol{\theta}$ is updated. The second cohort is not enrolled until the first cohort has been evaluated for cycle 1. For all subsequent cohorts, the posterior is updated after the outcomes of all previous cohorts are observed, and the posterior expected utility, $q_{i,1}(d_1, X)$, is computed using $\lambda = 0.8$. If $D_{i,1}(X) = \emptyset$ for any interim X, then $d_{i,1}(X) = 0$, and the trial is terminated. If $D_{i,1} \neq \emptyset$, then a cycle 1 dose is chosen from $D_{i,1}$ using AR($\boldsymbol{\epsilon}$). Once the outcomes in cycle 1 are observed,

the posterior is updated. Using $(d_{i,1}, Y_{i,1}, Z_{i,1}, X)$ and ϵ_i, $D_{i,2}$ is searched. If $D_{i,2}$ contains 0 only, then $d_{i,2} = 0$ and a cycle 2 dose is not given to patient i. Otherwise, $d_{i,2}$ is selected from $D_{i,2}(d_{i,1}, Y_{i,1}, Z_{i,1}, X)$ using AR($\boldsymbol{\epsilon}$). All adaptive decisions are made based on the most recent data X, hence a new $\boldsymbol{d}_{\text{select}}$ may be chosen using partial data from recent patients for whom (Y_1, Z_1) but not (Y_2, Z_2) have been evaluated. The above steps are repeated until either the trial has been stopped early or $N = 60$ has been reached, and in this case a final optimal two-cycle regime $\boldsymbol{d}_{\text{select}}$ is chosen. The aim is that d_{select} should be the true $d_1^{\text{opt}} \in \{1, \ldots, m\}$ and $d_2^{\text{opt}}(d_1^{\text{opt}}, y_1, z_1) \in D$. The recommendation for phase III is d_{select}, rather than a single "optimal" dose as is done conventionally.

We compared DTM2 design with four other designs: two-cycle extensions of the continual reassessment method (CRM, O'Quigley, Pepe, and Fisher 1990), a Bayesian phase I-II method using toxicity and efficacy odds ratios (TEOR, Yin, Li, and Ji 2006), and two (3 + 3) methods. One (3 + 3) method implicitly targets a dose with $\text{P}(\text{Y}_1 = 1) \leq 0.17$, called (3 + 3)a, and the other implicitly targets a dose with $\text{P}(\text{Y}_1 = 1) \leq 0.33$, called (3 + 3)b. We extended each one-cycle method to account for a second cycle. For both (3 + 3) methods, we used the deterministic rule in cycle 2 that if $Y_1 = 1$ then the dose is lowered by 1 level ($d_2 = d_1 - 1$) and if $Y_1 = 0$ then the first dose is repeated ($d_2 = d_1$). The (3 + 3)a method, coupled with this deterministic rule for cycle 2, is a very commonly used method in actual phase I clinical trials.

For cycle 1 in the extended CRM (ECRM), we assumed the usual model $\Pr(Y_1 = 1|d_1) = \text{p}_{d_1}^{\exp(\alpha)}$ and $\alpha \sim \text{N}(0, 2)$ where $0 < p_1 < \cdots < p_5 < 1$ are fixed values, sometimes called the model's "skeleton." We calibrated the skeleton using the "getprior" subroutine in the package "dfcrm," setting the target toxicity probability to be 0.30, the prior guess of maximum tolerated dose 4, and the desired halfwidth of the indifference intervals 0.05 (Cheung 2011). The resulting skeleton is $(p_1, \ldots, p_5) = (0.063, 0.123, 0.204, 0.300, 0.402)$. Using this model, each patient's cycle 1 dose is that with posterior mean toxicity probability closest to 0.30. We implemented this using the R function,"crm"in dfcrm, but also imposing the no-skipping rule for cycle 1. To determine a cycle 2 dose, we used the same deterministic rule as for the extended (3 + 3) methods, with one more safety requirement. For ECRM, a cycle 2 dose is not given if $\Pr\{\Pr(Y_1 = 1 \text{ or } Y_2 = 1) > \text{p}_\text{T}|X, \boldsymbol{d}\} > \psi_T$, with $p_T = 0.5$ and $\psi_T = 0.9$, assuming independence of $\text{P}(Y_1 = 1)$ and $\text{P}(Y_2 = 1)$ for simplicity. For example, following the deterministic rule, a patient treated in cycle 1 at d_1 may be treated at $d_2 \in \{d_1 - 1, d_1\}$ depending on the cycle 1 toxicity outcome. In particular, we repeat d_1 in cycle 2 if $Y_1 = 0$. If (d_1, d_1) does not satisfy the safety requirement, then the cycle 2 treatment is not given to a patient with d_1 and $Y_1 = 0$. In addition, if the cycle 2 treatment is not allowed for any d_1 regardless of Y_1, that is, no (d_1, d_2) with $d_2 \in \{d_1 - 1, d_1\}$ satisfies the safety rule, then we lower d_1 until the cycle 2 treatment is safe for either of $Y_1 = 0$ or $Y_1 = 1$.

We extended TEOR to 2-cycles similarly to ECRM, and named this ETEOR. For ETEOR, $d_2 = 0$ if $\Pr\{\Pr(Y_1 = 1 \text{ or } Y_2 = 1) > \text{p}_\text{T}|X, \boldsymbol{d}\} > \psi_T$ or $\Pr\{\Pr(Z_1 = 0 \text{ and } Z_2 = 0) > \text{p}_\text{E}|X, \boldsymbol{d}\} > \psi_E$ with $p_T = 0.6$, $p_E = 0.8$, and $\psi_T = \psi_E = 0.9$, assuming independence of the two cycles for simplicity. In addition, we calibrated the priors of Yin, Li, and Ji (2006) using the concept of prior effective sample size (see the supplementary material for details), resulting in their $\sigma_\phi^2 = 20$, $\sigma_\psi^2 = 5$ and $\sigma_\theta^2 = 10$. We set $\bar{\pi}_T = 0.35$, $\underline{\pi}_E = 0.5$, $p^{\text{escl}} = 0.5$, $p^\star = 0.25$ and $q^\star = 0.1$, and used $\omega_d^{(3)}$ to select a dose for the next patient.

4. SIMULATION STUDY

4.1 Simulation Design

We simulated trials under each of eight dose-outcome scenarios using each of the five designs: DTM2, and the extended 3+3, ECRM, and ETEOR methods. The first seven scenarios were obtained using the model underlying DTM2, with the eighth obtained from a very different model to study robustness. To specify 2-cycle simulation scenarios, one must specify a joint distribution of (Y_1, Z_1) for each d_1 and a joint distribution of (Y_2, Z_2) as a function of (d_1, d_2, y_1, z_1). For Scenarios 1–7, the marginal probabilities of toxicity and efficacy in each cycle are given in Table 2, and we simulated data using (4), with assumed values $\sigma_\xi^{2,\text{true}} = \sigma_\eta^{2,\text{true}} = 0.5^2$, $\tau^{2,\text{true}} = 0.3^2$ and $\rho^{\text{true}} = -0.2$. We determined $\bar{\boldsymbol{\xi}}^{\text{true}}$ and $\bar{\boldsymbol{\eta}}^{\text{true}}$ by matching $\Pr(Y_\text{c} < 0) = \Phi(0 \mid \bar{\xi}_{\text{d}_\text{c}}^{\text{true}}, \sigma_\xi^{2,\text{true}} + \tau^{2,\text{true}})$ and $\Pr(Z_\text{c} < 0) = \Phi(0 \mid \bar{\eta}_{\text{d}_\text{c}}^{\text{true}}, \sigma_\eta^{2,\text{true}} + \tau^{2,\text{true}})$. We used $(\bar{\boldsymbol{\xi}}^{\text{true}}, \bar{\boldsymbol{\eta}}^{\text{true}})$ and the assumed nuisance parameters to simulate $(\boldsymbol{Y}, \boldsymbol{z})$, generated (Y_1, Z_1) from (6) using the true values of σ_ξ^2, σ_η^2, τ^2, ρ, and used (5) to generate (Y_2, Z_2) conditional on (Y_1, Z_1).

To apply DTM2, we first calibrated the hyperparameters, $\tilde{\boldsymbol{\theta}}$, using the notion of the expected sample size (ESS) as described in Morita et al. (2010). We simulated 1000 pseudosamples of $\boldsymbol{\theta}$, setting $\sigma_{\xi_{c0}}^2 = \sigma_{\eta_{c0}}^2 = 6^2$, and computed probabilities of interest, such as $\text{P}(Y_\text{c} = 0|d_\text{c})$ and $\text{P}(Z_\text{c} = 0|d_\text{c})$, based on the pseudosamples, setting $\sigma_\xi^2 = \sigma_\eta^2 = 2^2$, $\tau^2 = 1$, and $\rho = -0.5$. We determined $\tilde{\boldsymbol{\theta}}$ that gave ESS ranging from 0.5 to 2 for the quantities of interest, and used this $\tilde{\boldsymbol{\theta}}$ to determine the prior for all simulations.

To study robustness, in Scenario 8 we simulated data using the following logistic regression model. The cycle 1 marginal probabilities $(p_T(d_1), p_E(d_1))$ are the same as those of Scenario 5, with outcomes generated using true probabilities

$$
\begin{aligned}
\Pr(Y_1 = 1 \mid \text{d}_1) &= p_T(d_1),\\
\Pr(Z_1 = 1 \mid d_1, Y_1) &= \text{logit}^{-1}\{\text{logit}(p_E(d_1))\\
&\quad - 0.34(Y_1 - 0.5)\},\\
\Pr(Y_2 = 1 \mid d_1, d_2, Y_1, Z_1) &= \text{logit}^{-1}\{\text{logit}(p_\text{T}(d_1))\\
&\quad + 0.33d_2 + 0.4(Y_1 - 0.5)\\
&\quad - 0.3(Z_1 - 0.5)\},\\
\Pr(Z_2 = 1 \mid d_1, d_2, Y_1, Z_1, Y_2) &= \text{logit}^{-1}\{\text{logit}(p_E(d_1))\\
&\quad + 0.76d_2 - 0.22(Y_1 - 0.5)\\
&\quad + 2.4(Z_1 - 0.5)\\
&\quad - 1.8(Y_2 - 0.5)\}.
\end{aligned}
$$

Table 3 shows the optimal actions, d_1^{opt} and $d_2^{\text{opt}}(d_1^{\text{opt}}, Y_1, Z_1)$, under each scenario. For example, in Scenario 3, the optimal

Table 2. True marginal probabilities of toxicity and efficacy under the first seven scenarios for the simulation studies, $(p_T, p_E)^{\text{true}}$ for cycles 1 and 2

		Doses				
Scenario	Cycles	1	2	3	4	5
1	1	(0.10, 0.02)	(0.15, 0.03)	(0.30, 0.05)	(0.45, 0.08)	(0.55, 0.10)
	2	(0.13, 0.01)	(0.18, 0.02)	(0.33, 0.04)	(0.48, 0.07)	(0.58, 0.09)
2	1	(0.30, 0.50)	(0.32, 0.60)	(0.35, 0.70)	(0.38, 0.80)	(0.40, 0.90)
	2	(0.33, 0.45)	(0.35, 0.55)	(0.38, 0.65)	(0.41, 0.75)	(0.43, 0.85)
3	1	(0.05, 0.10)	(0.18, 0.13)	(0.20, 0.25)	(0.40, 0.26)	(0.50, 0.27)
	2	(0.30, 0.20)	(0.31, 0.35)	(0.32, 0.45)	(0.45, 0.65)	(0.65, 0.70)
4	1	(0.13, 0.06)	(0.15, 0.18)	(0.25, 0.35)	(0.55, 0.38)	(0.75, 0.40)
	2	(0.20,0.14)	(0.25, 0.23)	(0.35, 0.29)	(0.50, 0.32)	(0.80, 0.35)
5	1	(0.52, 0.01)	(0.61, 0.15)	(0.71, 0.20)	(0.82, 0.25)	(0.90, 0.30)
	2	(0.53,0.04)	(0.55, 0.20)	(0.62, 0.25)	(0.85, 0.27)	(0.95, 0.33)
6	1	(0.25, 0.10)	(0.28, 0.13)	(0.30, 0.25)	(0.40, 0.35)	(0.50, 0.45)
	2	(0.30, 0.20)	(0.31, 0.35)	(0.32, 0.45)	(0.43,0.65)	(0.56, 0.70)
7	1	(0.25, 0.10)	(0.28, 0.13)	(0.30, 0.25)	(0.40, 0.38)	(0.65, 0.40)
	2	(0.30, 0.20)	(0.31, 0.35)	(0.32, 0.45)	(0.43,0.65)	(0.66, 0.67)

NOTE: The true marginal probabilities of Scenario 8 are identical to those of Scenario 5.

cycle 1 action is $d_1^{\text{opt}} = 3$, and the optimal cycle 2 action is $d_2^{\text{opt}}(d_1 = 3, Y_1 = 0, Z_1) = 4$ and $d_2^{\text{opt}}(d_1 = 3, Y_1 = 1, Z_1) = 2$, regardless of Z_1.

4.2 Evaluation Criteria

We used the following summary statistics to evaluate each method's performance. Denote the outcomes of the n patients in a given trial who received at least one cycle of therapy by $\{(Y_{i,1}, Z_{i,1}), (Y_{i,2}, Z_{i,2}), i = 1, \ldots, n\}$, where $n < 60$ if the trial was stopped early. The empirical mean total utility for the n patients is $\bar{U} = \sum_{i=1}^{n}\{U(Y_{i,1}, Z_{i,1}) + U(Y_{i,2}, Z_{i,2})\}/n$, where we set $U(Y_{i,2}, Z_{i,2}) = U(0, 0)$ for patients who did not receive a second cycle of therapy. Indexing the N simulated replications of the trial by $r = 1, \ldots, N$, the empirical mean total payoff for all patents in the trial is $\bar{\bar{U}} = N^{-1}\sum_{r=1}^{N}\bar{U}^{(r)}$. One may regard $\bar{\bar{U}}$ as an index of the ethical desirability of the method for the patients in the trial, given the utility $U(y, z)$.

Table 3. Optimal treatment sequences under the eight simulation scenarios using the simulation truth, $\boldsymbol{\theta}^{\text{true}}$

		d_2^{opt}			
Scenario	d_1^{opt}	(0,0)	(0, 1)	(1,0)	(1,1)
1	0	0	0	0	0
2	5	5	5	4	4
3	3	4	4	2	2
4	3	3	3	0	0
5	0	0	0	0	0
6	5	4	4	4	4
7	4	4	4	3	3
8	5	5	3	4	4

NOTE: Based on assumption that d_1^{opt} is given, d_2^{opt} is searched for each cycle 1 outcome combination.

To evaluate performance in terms of future patient benefit, recall that DTM2 selects an optimal dose $d_{1,\text{select}}$ for cycle 1, and an optimal function $d_{2,\text{select}}$ for use in cycle 2 assuming that $d_{1,\text{select}}$ is given, with $d_{2,\text{select}}$ not defined if $d_{1,\text{select}} = 0$. Let $\boldsymbol{\theta}^{\text{true}}$ be the true parameter vector assumed for a simulation scenario. Under $\boldsymbol{\theta}^{\text{true}}$, the expected payoff in cycle 1 of giving action $d_{1,\text{select}}$ to a future patient is $Q_{1,\text{select}}(d_{1,\text{select}}) = E\{U(Y_1, Z_1)|d_{1,\text{select}}, \boldsymbol{\theta}^{\text{true}}\}$, for $d_{1,\text{select}} \neq 0$. If the rule $d_{2,\text{select}}$ is used, the expected payoff in cycle 2 is

$$Q_{2,\text{select}}(d_{2,\text{select}}) = \sum_{(y_1,z_1)\in\{0,1\}^2} E\{U(Y_2, Z_2)|d_{1,\text{select}}, d_{2,\text{select}}(y_1, z_1), y_1, z_1, \boldsymbol{\theta}^{\text{true}}\} \times p(y_1, z_1 \mid d_{1,\text{select}}, \boldsymbol{\theta}^{\text{true}}),$$

where $E\{U(Y_2, Z_2)|d_{1,\text{select}}, d_{2,\text{select}}(y_1, z_1), y_1, z_1, \boldsymbol{\theta}^{\text{true}}\}$ becomes $U(0, 0)$ if $d_{2,\text{select}}(y_1, z_1) = 0$. The total expected payoff to a future patient treated using the selected regime $\boldsymbol{d}_{\text{select}} = (d_{1,\text{select}}, d_{2,\text{select}})$ is defined to be $Q_{\text{select}}(\boldsymbol{d}_{\text{select}}) = Q_{1,\text{select}}(d_{1,\text{select}}) + \lambda Q_{2,\text{select}}(d_{2,\text{select}})$.

In addition to the criteria $\bar{\bar{U}}$ and Q_{select}, we evaluated the empirical toxicity and efficacy rates, defined as follows. Let $\delta_{i,2} = 1$ if patient i was treated in cycle 2. For each simulated trial with each method, for patients who received at least one cycle of therapy, we computed

$$\Pr(\text{Tox}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1(Y_{i,1} = 1) + \delta_{i,2}1(Y_{i,2} = 1)}{1 + \delta_{i,2}}$$

Table 4. Simulation results for the proposed method DTM2, and for 2-cycle extensions (3 + 3)a, (3 + 3)b, ECRM of standard phase I methods, and the 2-cycle extension ETEOR of the phase I–II method of Li et al. (2006)

Scenarios	Criterion	DTM2	(3+3)a	(3+3)b	ECRM	ETEOR
1	$\bar{\bar{U}}$	66.48	59.27	58.81	56.56	61.90
	Q_{select}	57.77	54.36	52.30	51.75	52.43
	Pr(Tox)	0.25	0.22	0.23	0.27	0.25
	Pr(Eff)	0.07	0.03	0.03	0.05	0.07
	% completed trials	2.3	88.6	96.5	99.6	4.4
2	$\bar{\bar{U}}$	136.35	124.36	118.32	115.86	122.13
	Q_{select}	135.76	103.85	104.48	102.43	108.47
	Pr(Tox)	0.39	0.30	0.33	0.36	0.35
	Pr(Eff)	0.72	0.58	0.55	0.56	0.60
	% completed trials	99.4	39.2	64.7	95.6	78.2
3	$\bar{\bar{U}}$	94.23	85.95	85.75	89.93	88.04
	Q_{select}	84.39	77.98	80.14	78.43	78.47
	Pr(Tox)	0.38	0.27	0.27	0.30	0.26
	Pr(Eff)	0.38	0.27	0.27	0.33	0.28
	% completed trials	79.4	96.6	99.2	100.0	78.50
4	$\bar{\bar{U}}$	75.84	81.81	80.12	85.40	84.94
	Q_{select}	69.49	74.92	75.76	78.67	78.87
	Pr(Tox)	0.51	0.25	0.26	0.29	0.28
	Pr(Eff)	0.29	0.22	0.21	0.29	0.27
	% completed trials	96.7	83.2	94.7	99.4	81.7
5	$\bar{\bar{U}}$	66.65	52.87	52.72	50.41	NA
	Q_{select}	50.64	40.66	40.70	40.61	NA
	Pr(Tox)	0.84	0.43	0.44	0.53	NA
	Pr(Eff)	0.35	0.08	0.04	0.03	NA
	% completed trials	0.4	6.8	20.0	9.0	0.0
6	$\bar{\bar{U}}$	96.43	82.82	79.27	81.50	86.14
	Q_{select}	92.78	70.30	71.28	71.29	76.24
	Pr(Tox)	0.45	0.28	0.32	0.32	0.32
	Pr(Eff)	0.41	0.24	0.23	0.25	0.29
	% completed trials	90.9	51.5	74.7	97.6	58.3
7	$\bar{\bar{U}}$	91.88	82.66	79.31	80.99	86.32
	Q_{select}	84.91	70.28	71.27	71.16	76.34
	Pr(Tox)	0.47	0.28	0.32	0.32	0.32
	Pr(Eff)	0.38	0.24	0.22	0.25	0.29
	% completed trials	90.3	51.4	73.6	97.5	58.7
8	$\bar{\bar{U}}$	95.92	80.24	76.09	79.83	80.75
	Q_{select}	93.22	68.23	69.26	69.28	70.73
	Pr(Tox)	0.54	0.34	0.36	0.37	0.34
	Pr(Eff)	0.45	0.25	0.22	0.27	0.27
	% completed trials	84.7	49.4	73.2	97.6	57.9

NOTE: $\bar{\bar{U}}$ = mean empirical utility, Q_{select} = mean payoff of $\boldsymbol{d}_{\text{select}}$. Empirical percentages Pr(Tox) and Pr(Eff) include patients who received at least cycle 1 of treatment.

and

$$\Pr(\text{Eff}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1(Z_{i,1}=1)+\delta_{i,2}1(Z_{i,2}=1)}{1+\delta_{i,2}}.$$

4.3 Simulation Results

A total of $N = 1000$ trials were simulated under each scenario for each of the five designs studied. The simulation results are summarized in Table 4. For the each of the five trial designs, Table 4 gives $\bar{\bar{U}}$, Q_{select}, the empirical per-cycle toxicity and efficacy probabilities and the percent of trials completed with $d_{1,\text{select}} \in \{1,\ldots,m\}$. A difference in $\bar{\bar{U}}$ or Q_{select} that may be considered "large" is about 5, since this translates to, on average, a difference of about 0.13 in Pr(Tox), while a difference 1 may be considered small.

In Scenario 1, Table 2 shows that doses d = 1,2,3 are safe, d = 4,5 are overly toxic, and all doses have very low efficacy.

In this case there is little benefit from any dose. The value $\bar{\bar{U}} = 66.48$ for DTM2 in Table 4 is close to the utility $U(0, 0) + 0.8U(0, 0) = 66$ of $(d_1 = 0, d_2 = 0)$. The utility-based stopping rule of DTM2 correctly terminates the trial 97.7% of the time. Similarly, ETEOR terminates 95.6% of the trials before reaching the maximum number of patients due to the low efficacy rates. In contrast, the extended versions of the 3 + 3 and ECRM are very likely to run the trial to completion, essentially because they ignore efficacy. This provides a stark illustration of the fact that there is little benefit in exploring the safety of an agent if it is inefficacious, and methods that ignore efficacy are very likely to make this mistake. This has little to do with the two-cycle structure, and it also can be seen when comparing one-cycle phase I-II (efficacy and toxicity) to phase I (toxicity only) methods. Thus, DTM2 and ETEOR are the only reasonable designs in Scenario 1, and DTM2 is superior in terms of both $\bar{\bar{U}}$ and Q_{select}.

In Scenario 2, Table 2 shows that the toxicity probabilities increase with dose from 0.30 to 0.40 in cycle 1 and from 0.33 to 0.43 in cycle 2, while the efficacy probabilities are quite high in both cycles, increasing with dose from 0.50 to 0.90 in cycle 1 and from 0.45 to 0.85 in cycle 2. Thus, if one considers toxicity probabilities around 0.40 to be acceptable tradeoffs for these very high efficacy rates, then there is a substantial payoff for escalating to higher doses. The utility function reflects this, with the optimal action $d_1^{\text{opt}} = 5$ and $d_2^{\text{opt}}(5, Y_1, Z_1) = 4$ or 5 (Table 3). DTM2 obtains larger values of $\bar{\bar{U}}$ and Q_{select} due to much larger Pr(Eff) and slightly larger Pr(Tox), compared to all of the four methods.

In Scenario 3, $d_1^{\text{opt}} = 3$, with $d_2^{\text{opt}} = 4$ if $Y_1 = 0$ in cycle 1 and $d_2^{\text{opt}} = 2$ if $Y_1 = 1$ (Table 3). This illustrates the within-patient adaptation of DTM2. The (3 + 3)a, (3 + 3)b, and ECRM methods select $d_1^{\text{opt}} = 3$ often since the toxicity probability of $d_1 = 3$ is close to 0.30, but they never select $d_2^{\text{opt}} = 2$ for patients with $(d_1, Y_1) = (3, 0)$ because their deterministic rules ignore Z_1 and do not allow escalation of dose levels for cycle 2 even with $Y_1 = 0$. Again, DTM2 achieves the largest $\bar{\bar{U}}$, Q_{select}, and Pr(Eff), with slightly larger Pr(Tox).

Scenario 4 is a challenging scenario for DTM2, and is favorable for the other four designs. In Scenario 4, $d_1^{\text{opt}} = 3$ since its toxicity probability 0.25 is closest to 0.30. In addition, $d_2^{\text{opt}}(d_1^{\text{opt}}, Y_1, Z_1)$ is exactly the same as the cycle 2 dose levels chosen by the deterministic rules of (3 + 3)a, (3 + 3)b and ECRM, except for $(Y_1, Z_1) = (0, 1)$, which only occurs about 5% of the time. From Table 1, the true expected utility of $d_2 = 2$ given $(d_1, Y_1, Z_1) = (3, 0, 1)$ is 32.82, which is very close to $U(0, 0)$. Thus, the three methods, (3 + 3)a, (3 + 3)b, and ECRM, are likely to select d_1^{opt} by considering only toxicity outcomes and select d_2^{opt} following their deterministic rules. CRM selects $d_1^{\text{opt}} = 3$ most of time, leading to the largest $\bar{\bar{U}}$ and Q_{select}. Similar performance is observed for ETEOR as well because d_1^{opt} is considered optimal by ETEOR, and it uses the same deterministic rule for cycle 2. The smaller values of $\bar{\bar{U}}$ and Q_{select} for DTM2 are because it does a stochastic search to determine the optimal actions, using much more general criteria than the other methods. Table 1 shows that, for $(d_1, Y_1) = (3, 1)$, the expected cycle 2 utilities are smaller than or very close to $U(0, 0)$ for all the cycle 2 doses, so all cycle 2 doses are barely acceptable or not acceptable. However, $d_1 = 5$ is acceptable and, given $d_1 = 5$, many cycle 2 doses are acceptable, and DTM2 often explores higher doses in cycle 1 than d_1^{opt}. This scenario illustrates the price one may pay for using more of the available information to explore the dose domain more extensively based on an efficacy-toxicity utility-based objective function.

In Scenario 5, the lowest dose is too toxic and, therefore, even $d_1 = 1$ is unacceptable. As expected, all methods terminate the trial early most of time, with DTM2 stopping trials due to the low posterior expected utilities caused by the high toxicity rate at d_1.

Scenarios 6 and 7 have identical true toxicity and efficacy rates for doses 1, 2, and 3, while for doses 4 and 5, Scenario 7 has higher toxicity rates and lower efficacy rates so that its d_1^{opt} is a dose lower than d_1^{opt} of Scenario 6. Since dose 3 has a toxicity rate closest to 0.3 in the both scenarios, the other four methods perform very similarly in the two scenarios. However, DTM2 again has much higher $\bar{\bar{U}}$ and Q_{select} values compared to all of the other methods in these scenarios.

Recall that Scenario 8 is included to evaluate robustness, with joint probabilities generated using a model very different from that underlying DTM2. It thus is remarkable that, in terms of both $\bar{\bar{U}}$ and Q_{select}, DTM2 has far superior performance compared to all four other methods. Essentially, this is because DTM2 allows a higher rate of toxicity as a tradeoff for much higher efficacy, while the phase I methods (3 + 3)a, (3 + 3)b, and ECRM all ignore efficacy, and the other phase I-II method, ETEOR, terminates the trial early much more frequently. The superior performance of DMT2 in Scenario 8 may be attributed to its use of a 2-cycle utility function to account for efficacy-toxicity tradeoffs and also as a basis for its early stopping rule. More generally, it appears that DTM2 is quite robust to the actual probability mechanism that generates the outcomes.

To assess sensitivity to association among the outcomes Y_1, Z_1, Y_2, Z_2 in the two cycles, we evaluated each method's performance with and without association in Scenarios 3, 6, and 7. We let the true $(\sigma_\xi^2, \sigma_\eta^2, \tau^2, \rho)$ be either $(0.2, 0.05, 1, -0.5)$ or $(0.5^2, 0.5^2, 0, 0)$. The first set of values induces high association between outcomes both within and across cycles, while the second set of values induces no association. This leads to different true expected utilities in each cycle and thus to different optimal decisions, as shown in Table 5. The results are summarized in Table 6. While performance changes depending on the assumed

Table 5. Optimal sequence of treatments under scenarios 3, 6, and 7, assuming different values of $(\sigma_\xi^{2,\text{true}}, \sigma_\eta^{2,\text{true}}\tau^{2,\text{true}}, \rho^{\text{true}})$ to induce either high association or no association between outcomes

		d_2^{opt}			
Scenario	d_1^{opt}	(0,0)	(0, 1)	(1,0)	(1,1)
3-High Assoc.	3	4	3	NT	2
3-No Assoc.	3	4	4	2	2
6-High Assoc.	5	5	4	NT	3
6-No Assoc.	5	4	4	4	4
7-High Assoc.	4	4	4	NT	3
7-No Assoc.	4	4	4	3	3

Table 6. Simulation results under scenarios 3, 6, and 7, assuming different values of ($\sigma_\xi^{2,\text{true}}$, $\sigma_\eta^{2,\text{true}}\tau^{2,\text{true}}$, ρ^{true}) to induce either high association or no association between outcomes

Scenarios	Criterion	DTM2	(3 + 3)a	(3 + 3)b	ECRM	ETEOR
3	$\bar{\bar{U}}$	97.06	85.68	85.24	88.56	89.85
	Q_{select}	86.18	78.53	80.53	76.58	79.27
	Pr(Tox)	0.37	0.27	0.28	0.31	0.26
High Assoc.	Pr(Eff)	0.38	0.26	0.26	0.33	0.29
	% completed trials	97.7	96.6	99.2	99.9	77.5
3	$\bar{\bar{U}}$	92.05	85.96	85.44	90.06	87.88
	Q_{select}	82.22	77.83	80.04	79.35	78.75
	Pr(Tox)	0.41	0.27	0.27	0.30	0.26
No Assoc.	Pr(Eff)	0.36	0.26	0.26	0.33	0.28
	% completed trials	98.2	96.6	99.2	99.9	77.5
6	$\bar{\bar{U}}$	101.37	85.54	81.74	83.20	89.87
	Q_{select}	95.18	72.43	73.35	71.40	77.67
	Pr(Tox)	0.42	0.26	0.29	0.31	0.31
High Assoc.	Pr(Eff)	0.43	0.25	0.22	0.26	0.31
	% completed trials	91.1	51.5	74.7	97.5	59.9
6	$\bar{\bar{U}}$	94.63	82.45	78.73	81.51	85.11
	Q_{select}	90.85	69.76	70.75	71.53	76.11
	Pr(Tox)	0.46	0.29	0.32	0.33	0.32
No Assoc.	Pr(Eff)	0.40	0.24	0.22	0.26	0.28
	% completed trials	91.6	51.5	74.7	98.2	57.0
7	$\bar{\bar{U}}$	96.67	85.40	81.63	82.94	90.00
	Q_{select}	87.63	72.44	73.35	71.28	77.84
	Pr(Tox)	0.44	0.26	0.29	0.31	0.31
High Assoc.	Pr(Eff)	0.41	0.25	0.22	0.26	0.31
	% completed trials	90.7	51.4	73.6	97.9	60.2
7	$\bar{\bar{U}}$	89.91	82.25	78.64	81.34	85.29
	Q_{select}	82.89	69.74	70.74	71.23	76.20
	Pr(Tox)	0.48	0.29	0.32	0.32	0.32
No Assoc.	Pr(Eff)	0.37	0.24	0.22	0.25	0.28
	% completed trials	90.8	51.4	73.6	97.5	57.3

true values, in all cases DTM2 is again superior to all four other methods.

5. DISCUSSION

Practical application of DTM2 requires substantial input from the physicians, including specification of outcomes, doses, prior values, and numerical utilities. Such key input from the physicians, and preliminary validation by computer simulation, have provided a practical basis for use of model-based outcome-adaptive methods in many actual phase I-II dose-finding trials (see de Lima et al. 2008). In the design process, computer simulation also may be used to conduct sensitivity analyses in the prior or the numerical utilities so that the physicians may adjust their values. For trial conduct, a database and data entry procedure must be established, with the database updated in real time as patients are treated and evaluated in each cycle. The actual data used by DTM2 are simple, however, consisting of $(d_1, Y_1, Z_1, d_2, Y_2, Z_2)$. Accounting for two cycles rather than only one is not a substantial complication compared to usual adaptive trials, since all clinical protocols contain rules for adaptive within-patient decision making.

DTM2 provides the 2-cycle regime $\boldsymbol{d}_{\text{select}}$ for phase III evaluation, rather than only a selected d_1 or 2-cycle pair (d_1, d_2). This is an important improvement, since it more accurately reflects actual clinical practice and is likely to improve the chance of success in phase III. This is because phase I methods based on toxicity alone are likely to fail to identify higher doses having higher efficacy and acceptable toxicity, and thus are more likely to select an ineffective dose for phase II evaluation. Moreover, our comparisons to the 2-cycle extension ETEOR of the phase I-II design of Yin, Li, and Ji (2006), which also uses efficacy, show the advantage of optimizing a utility-based Q-function for decision making.

Several important practical extensions should be noted. While DTM2 uses recent patients' partial data if only their cycle 1 outcomes have been evaluated, this may be refined by using event time data to enhance inferences. A useful extension would to use toxicity or efficacy follow up times from patients treated and but not fully evaluated, employing predictive probabilities similarly to Bekele et al. (2008), or taking the approach of Zhao et al.

(2011). Bivariate ordinal (Y_c, Z_c) outcomes with more than two levels may be accommodated by extending the model to include more cutoffs in the latent variables, and eliciting corresponding utilities, as in Thall and Nguyen (2012). Extension to accommodate this case is complex, however, since there would be many more elementary outcomes and thus many more model parameters. Numerous ad hoc adaptive methods for choosing a patient's doses in cycles after the first actually are used in clinical practice. For example, if Y and Z each have four levels, then for two cycles there would be 16 elementary outcomes, rather than 4, (ξ, η) would be eight-dimensional, and $\Sigma_{\xi,\eta}$ would be an 8×8 matrix. Since many actual regimes involve more than two cycles, while in theory the decision criterion can be generalized to accommodate this in a straightforward manner, the dimensions of the outcomes and decisions become much larger. This strongly suggests that, to deal with the general multicycle case in a practical way, a more parsimonious model will be needed.

6. SUPPLEMENTARY MATERIALS

Supplementary materials discuss prior calibration and posterior computation.

[Received April 2013. Revised January 2014.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 66–99. [712]

Almirall, D., Ten Have, T., and Murphy, S. A. (2010), "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation," *Biometrics*, 66, 131–139. [711]

Ashford, J. R., and Sowden, R. R. (1970), "Multi-Variate Probit Analysis," *Biometrics*, 26, 535–546. [712]

Azriel, D., Mandel, M., and Rinott, Y. (2011), "The Treatment Versus Experiment Dilemma in Dose-Finding Studies," *Journal of the Statistical Planning and Inference*, 141, 2759–2768. [715]

Bartroff, J., and Lai, T. L. (2010), "Approximate Dynamic Programming and Its Applications to the Design of Phase I Cancer Trials," *Statistical Science*, 25, 255–257. [715]

Bekele, B. N., Ji, Y., Shen, Y., and Thall, P. F. (2008), "Monitoring Late Onset Toxicities in Phase I Trials Using Predicted Risks," *Biostatistics*, 9, 442–457. [720]

Bekele, B. N., and Thall, P. F. (2004), "Dose-Finding Based on Multiple Toxicities in a Soft Tissue Sarcoma Trial," *Journal of the American Statistical Association*, 99, 26–35. [712]

Bellman, R. E. (1957), *Dynamic Programming*, Princeton, NJ: Princeton University Press. [712,713]

Braun, T. M., Kang, S., and Taylor, J. M. G. (2012), "A Phase I/II Trial Design When Response is Unobserved in Subjects With Dose-Limiting Toxicity," *Statistical Methods in Medical Research* [online]. [715]

Braun, T. M., Thall, P. F., Nguyen, H., and de Lima, M. (2007), "Simultaneously Optimizing Dose and Schedule of a New Cytotoxic Agent," *Clinical Trials*, 4, 113–124. [711]

Braun, T. M., Yuan, Z., and Thall, P. F. (2005), "Determining a Maximum Tolerated Schedule of a Cytotoxic Agent," *Biometrics*, 61, 335–343. [711]

Cheung, Y.-K. (2011), *Dose Finding by the Continual Reassessment Method*, London: Chapman & Hall/CRC Biostatistics Series. [716]

Cheung, Y.-K., and Chappell, R. (2000), "Sequential Designs for Phase I Clinical Trials With Late-Onset Toxicities," *Biometrics*, 56, 1177–1182. [711]

Chevret, S. C. (2006), *Statistical Methods for Dose Finding Experiments*, Chichester, UK: Wiley. [711]

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 3471. [712]

Collins, L. M., Murphy, S. A., Nair, V. N., and Strecher, V. J. (2005), "A Strategy for Optimizing and Evaluating Behavioral Interventions," *Annals of Behavioral Medicine*, 30, 65–73. [711]

de Lima, M., Champlin, R. E., Thall, P. F., Wang, X., Cook, J. D., Martin, T. G., McCormick, G., Qazilbash, M., Kebriaei, P., Couriel, D., Shpall, E. J., Khouri, I., Anderlini, P., Hosing, C., Chan, K. E., Patah, P. A., Caldera, Z., Jabbour, E., and Giralt, S. (2008), "Phase I/II Study of Gemtuzumab Ozogamicin Added to Fludarabine, Melphalan and Allogeneic Hematopoietic Stem Cell Transplantation for High-Risk CD33 Positive Myeloid Leukemias And Myelodysplastic Syndrome," *Leukemia*, 22, 258–264. [720]

Hernan, M., Brumback, B., and Robins, J. (2000), "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men," *Epidemiology*, 11, 561–570. [711]

Lavori, P. W., and Dawson, R. (2001), "Dynamic Treatment Regimes: Practical Design Considerations," *Statistics in Medicine*, 20, 1487–1498. [711]

Li, Y., Bekele, B. N., Ji, Y., and Cook, J. D. (2008), "Dose-Schedule Finding in Phase I/II Clinical Trials Using a Bayesian Isotonic Transformation," *Statistics In Medicine*, 27, 4895–4913. [711,718]

Lunceford, J., Davidian, M., and Tsiatis, A. A. (2002), "Estimation of the Survival Distribution of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials," *Biometrics*, 58, 48–57. [711]

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007), "Demystifying Optimal Dynamic Treatment Regimes," *Biometrics*, 63, 447–455. [711]

Morita, S., Thall, P. F., and Mueller, P. (2010), "Evaluating the Impact of Prior Assumptions in Bayesian Biostatistics," *Statistics in Biosciences*, 2, 1–17. [716]

Murphy, S. (2003), "Optimal Dynamic Treatment Regimes" (with discussion), *Journal of the Royal Statistical Society,* Series B, 65, 331–366. [711]

——— (2005), "A Generalization Error for Q-Learning," *Journal of Machine Learning Research*, 6, 1037–1097. [713]

Murphy, S., and Bingham, D. (2009), "Screening Experiments for Developing Dynamic Treatment Regimes," *Journal of American Statistical Association*, 104, 391–408. [711]

Murphy, S. A., Collins, L. M., and Rush, A. J. (2007), "Customizing Treatment to the Patient: Adaptive Treatment Strategies," *Drug and Alcohol Dependence*, 88, S1–S3. [711]

Murphy, S. A., Lynch, K. G., Oslin, D., Mckay, J. R., and TenHave, T. (2007), "Developing Adaptive Treatment Strategies in Substance Abuse Research," *Drug and Alcohol Dependence*, 88s, S24–S30. [711]

Murphy, S. A., van der Laan, M. J., and Robins, J. M. (2001), "Marginal Mean Models for Dynamic Regimes," *Journal of the American Statistical Association*, 96, 1410–1423. [711]

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. [716]

Robins, J. M. (1986), "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Survivor Effect," *Mathematical Modeling*, 7, 1393–1512. [711]

——— (1993), "Analytic Methods for Estimating HIV Treatment and Cofactor Effects," in *Methodological Issues of AIDS Mental Health Research*, eds. D. G. Ostrow and R. Kessler, New York: Plenum Publishing, pp. 213–290. [711]

——— (1997), *Causal Inference from Complex Longitudinal Data Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics (120), ed. M. Berkane, New York: Springer Verlag, pp. 69–117. [711]

——— (1998), "Marginal Structural Models," *Proceedings of the American Statistical Association Section on Bayesian Statistics*, 1–10. [711]

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560. [711]

Rush, A. J., Trivedi, M., and Fava (2003), "Depression IV: STAR*D Treatment Trial For Depression," *American Journal of Psychiatry*, 160, 237. [711]

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester, UK: Wiley. [714]

Sutton, R. S., and Barto, A. G. (1998), *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press. [712]

Thall, P. F., Millikan, R., and Sung, H.-G. (2000), "Evaluating Multiple Treatment Courses in Clinical Trials," *Statistics in Medicine*, 19, 1011–1028. [711]

Thall, P. F., and Nguyen, H. Q. (2012), "Adaptive Randomization to Improve Utility-Based Dose-Finding With Bivariate Ordinal Outcomes," *Journal of the Biopharmaceutical Statistics*, 22, 785–801. [715,721]

Thall, P. F., Nguyen, H. Q., and Estey, E. H. (2008), "Patient-Specific Dose-Finding Based on Bivariate Outcomes and Covariates," *Biometrics*, 64, 1126–1136. [714]

Thall, P. F., Sung, H.-G., and Estey, E. H. (2002), "Selecting Therapeutic Strategies Based on Efficacy and Death in Multi-Course Clinical Trials," *Journal of the American Statistical Association*, 97, 29–39. [711]

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R., and Tannir, N. M. (2007), "Bayesian and Frequentist Two-Stage Treatment Strategies Based on Sequential Failure Times Subject to Interval Censoring," *Statistics in Medicine*, 26, 4687–4702. [711]

Tokic, M. (2010), "Adaptive ϵ-Greedy Exploration in Reinforcement Learning Based on Value Differences," in *Advances in Artificial Intelligence*, Heidelberg, Germany: Springer Verlag, pp. 203–210. [715]

Wahed, A. S., and Tsiatis, A. A. (2006), "Semiparametric Estimation of Survival Distribution for Treatment Policies in Two-Stage Randomization Designs in Clinical Trials with Censored Data," *Biometrics*, 93, 163–177. [711]

——— (2004), "Optimal Estimator for the Survival Distribution and Related Quantities for Treatment Policies in Two-Stage Randomization Designs in Clinical Trials," *Biometrics*, 60, 124–133. [711]

Wang, L., Rotnitzky, A., Lin, X., Millikan, R., and Thall, P. F. (2012), "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer" with discussion, *Journal of American Statistical Association*, 107, 493–520. [711,712]

Watkins, C. J. C. H. (1989), "Learning From Delayed Rewards," unpublished PhD thesis, Cambridge University. [712,713]

Yin, G. (2012), *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*, New York: Wiley. [711]

Yin, G., Li, Y., and Ji, Y. (2006), "Bayesian Dose-Finding in Phase I/II Clinical Trials Using Toxicity and Efficacy Odds Ratios," *Biometrics*, 62, 777–789. [716,720]

Zhang, J., and Braun, T. M. (2013), "A Phase I Bayesian Adaptive Design to Simultaneously Optimize Dose and Schedule Assignments Both Between and Within Patients," *Journal of the American Statistical Association*, 108, 892–901. [712]

Zhao, Y., Zheng, D., Socinski, M. A., and Kosorok, M. R. (2011), "Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer," *Biometrics*, 67, 1422–1433. [711,713,721]

Zohar, S., and Chevret, S. (2007), "Recent Developments in Adaptive Designs for Phase I/II Dose-Finding Studies," *Journal of Biopharmaceutical Statistics*, 17, 1071–1083. [711]