Using Data Augmentation to Facilitate Conduct of Phase I-II Clinical Trials with Delayed Outcomes

Ick Hoon Jin, Suyu Liu, Peter F. Thall, and Ying Yuan

October 17, 2013

Abstract

A practical impediment in adaptive clinical trials is that outcomes must be observed soon enough to apply decision rules to choose treatments for new patients. For example, if outcomes take up to six weeks to evaluate and the accrual rate is one patient per week, on average three new patients will be accrued while waiting to evaluate the outcomes of the previous three patients. The question is how to treat the new patients. This logistical problem persists throughout the trial. Various *ad hoc* practical solutions are used, none entirely satisfactory. We focus on this problem in phase I-II clinical trials that use binary toxicity and efficacy, defined in terms of event times, to choose doses adaptively for successive cohorts. We propose a general approach to this problem that treats late-onset outcomes as missing data, uses data augmentation to impute missing outcomes from posterior predictive distributions computed from partial follow-up times and complete outcome data, and applies the design's decision rules using the completed data. We illustrate the method with two cancer trials conducted using a phase I-II design based on efficacy-toxicity trade-offs, including a computer stimulation study.

Keywords: Bayesian Adaptive Clinical Design; Dose-Finding; Phase I-II Clinical Trial Design; Missing Data; Data Augmentation Algorithm; Piecewise Exponential Model.

1 Introduction

Phase I-II clinical trial designs combine conventional phase I and phase II trials by determining a dose of a new agent based on both toxicity and efficacy (Gooley et al., 1994; Thall and Russell, 1998; O'Quigley et al., 2001; Braun, 2002; Thall and Cook, 2004; Bekele and Shen, 2005; Zhang et al., 2006; Yin et al., 2006; Yuan and Yin, 2011a). Most commonly, a small phase I trial based on toxicity (Storer, 1989; O'Quigley et al., 1990; Babb et al., 1998; Conaway et al., 2004) first is conducted to choose a putatively safe dose, the "maximum tolerated dose" (MTD), and a phase II trial then is conducted (Gehan, 1969; Fleming, 1982; Simon, 1989; Thall and Simon, 1994; Thall et al., 1995; Bryant and Day, 1995), using the MTD, with efficacy the primary outcome. This conventional approach may lead to several problems. (1) Informal dose adjustments often are made in phase II if excessive toxicity is observed, which invalidates assumed properties of any efficacy-based design. (2) Operating characteristics of the entire phase I – phase II process are seldom computed. (3) Separate designs ignore the trade-off between efficacy and toxicity that often underlies therapeutic decision making. Since phase I-II designs address these problems explicitly, a natural question is why such designs are used infrequently for actual trials. While these are complex issues, the following logistical problem may play a prominent role.

In outcome-adaptive clinical trials, a major practical impediment arises if patient outcomes are not observed quickly enough to apply decision rules that choose treatments or doses for newly accrued patients. Because it is undesirable, and often impossible, to delay a new patient's treatment while waiting for previous patients' outcomes to be scored so that an adaptive statistical rule can be applied, outcome-adaptive rules may be at odds with clinical practice. In phase I-II, this problem arises if either toxicity or efficacy is not scored quickly, relative to the accrual rate. One solution is to turn away new patients and treat them off protocol. This may be less desirable than giving the experimental regimen, or impossible if no alternative treatment exists. Another solution is to give all new patients the dose or treatment that is optimal based on the most recent data. This may have very undesirable consequences if the most recent optimal dose later turns out to be overly toxic, and it is the main reason that dose-finding usually is done sequentially with small cohorts of 1, 2, or 3 patients. In Phase I-II, the opposite effect can occur if the most

recent optimal dose is safe but inefficacious. A third solution when some patients' outcomes have not yet been evaluated fully is to treat all new patients immediately, but use a dose one level below the design's current optimal dose. In Phase I, the problem of dealing with delayed ("late onset") toxicity was first addressed by Cheung and Chappell (2000), who introduced the time-to-event continual reassessment method (TiTE-CRM), and later by Braun (2006), Bekele et al. (2008), and Yuan and Yin (2011b).

As an illustration, consider a phase I-II clinical trial of chemotherapy for acute leukemia, both toxicity and efficacy events may occur at any time during a common 6-week evaluation period, doses chosen adaptively for cohorts of size 3, and accrual rate 1 patient per week. On average, the first cohort will be accrued in 3 weeks with all of their outcomes scored by week 9, when adaptive rules are applied using their data to choose the second cohort's dose. Since one also can expect to accrue 6 new patients between weeks 3 and 9, the question is how to deal therapeutically with these new patients. A second example is an autologous stem cell transplantation trial for multiple myeloma (MM) where toxicity may occur at any time during the first 30 days, but efficacy is evaluated only at 90 days post-transplant, and is defined as no detectable MM protein in the urine or blood serum and $< 5\%$ plasma cells in the bone marrow. If the accrual rate is 3 patients per month, one may expect to accrue 9 patients before efficacy is scored for any patients, so applying an adaptive rule to choose a dose for patients 4, 5, and 6, (cohort 2), using the data from patients 1, 2, and 3, (cohort 1), is not possible without delaying the second cohort's therapy. The severity of this problem increases with accrual rate and persists throughout the trial.

In this paper, we consider phase I-II designs where both efficacy (E) and toxicity (T) are characterized as binary variables, Y_E and Y_T, evaluated either during or at the end of specified time intervals, $[0,\ U_E]$ for Y_E and $[0,\ U_T]$ for Y_T. We refer to Y_E and Y_T as "delayed outcomes" because they are not observed immediately. Denote accrual rate by α. We quantify the severity of the problem using the logistical difficulty indexes $\zeta_j = U_j\alpha$ for each outcome $j = E, T$, and overall index $\zeta = \max\{\zeta_E, \zeta_T\}$. For example, if $\alpha = 1$ patient/month and $U_E = U_T = 3$ months, then $\zeta = 3$. Doubling the accrual rate to $\alpha = 2$ gives the problematic value $\zeta = 6$.

We propose a methodology that handles the problem of delayed outcomes in phase I-II by accounting for each patient's follow up time prior to evaluation of Y_E and Y_T, and treating all

Y_j's that have not yet been observed as missing values. We use data augmentation (Tanner and Wong, 1987) to impute each missing Y_j using partial follow-up times and complete outcome data. Combining observed and imputed (Y_E, Y_T) data for all patients who have been treated, we then apply the phase I-II design's decision rules. Our approach of treating delayed outcomes as missing data is similar to that of Yuan and Yin (2011b), who deal with the phase I setting with Y_T but not Y_E. Key differences are that the bivariate distribution and missingship patterns of (Y_E, Y_T) are much more complicated than those of Y_T alone. In addition, Yuan and Yin (2011b) use the EM algorithm under a frequentist framework to estimate toxicity probabilities, whereas we rely on predictive probabilities and imputatation to obtain a completed data set under a Bayesian formulation.

In Section 2, we propose a general approach to delayed outcomes in phase I-II trials. In Section 3, we present imputation and Bayesian data augmentation methods. Section 4 illustrates the method with two trials designed using an extended version of the efficacy-toxicity (Eff-Tox) trade-off based design of Thall and Cook (2004). In Section 5, we present simulations of the proposed method and comparison with alternative methods. We conclude with a discussion in Section 6.

2 Observed Outcomes and Missing Values

2.1 Data Structures

The data structure assumed by most phase I-II dose-finding method consists of assigned doses and two binary outcomes. We denote this by $\boldsymbol{\mathcal{D}}(Y) = \{(d_{[1]}, Y_{1,E}, Y_{1,T}), \ldots, (d_{[n]}, Y_{n,E}, Y_{n,T})\}$, where $d_{[i]}$ is the dose given to the ith patient and interim sample size $n = 1, \cdots, N_{max}$. To account for the fact that, at any interim decision, one or both of Y_E and Y_T may not yet be observed for a given patient, we denote the data including only the observed Y_j's by $\boldsymbol{\mathcal{D}}_{\text{obs}}(Y)$ and the unobserved ("missing") Y_j's by $\boldsymbol{\mathcal{D}}_{\text{mis}}(Y)$. Our strategy will be to use $\boldsymbol{\mathcal{D}}_{\text{obs}}(Y)$ and partial follow up time data to impute any missing Y_j's in $\boldsymbol{\mathcal{D}}_{\text{mis}}(Y)$ and thus construct a completed version of $\boldsymbol{\mathcal{D}}_{\text{obs}}(Y)$ that has the form of $\boldsymbol{\mathcal{D}}(Y)$, and then apply the phase I-II method.

Let X_j denote time to outcome $j = E, T$. Corresponding to the acute leukemia trial and the

stem cell transplantation trial described previously, we distinguish between two data structures, depending on how Y_E is observed. In Case 1, Y_E is observed in real time, and $Y_E = I(X_E \leq U_E)$. In Case 2, Y_E is evaluated only at U_E, and there is no X_E. In both cases, $Y_T = I(X_T \leq U_T)$. To simplify exposition, we focus primarily on Case 1 since it is more complex, and later explain how to deal with Case 2. For trials where efficacy is evaluated periodically X_E is interval censored. We include this by smoothing the interval censored data, placing X_E at the midpoint or a randomly chosen value of the interval where efficacy was known to occur. We accommodate patient death during therapy by defining X_T as the time to either non-fatal toxicity or death. For efficacy, if the patient dies prior to U_E we define $Y_E = 0$, which is implied by defining $X_E = \infty$. Denote $U = \max\{U_T, U_E\}$ and $V =$ follow up time, where by design $V \leq U$.

2.2 Missingship Mechanism

Denote $\pi_j(d, \theta) = \Pr(Y_j = 1 \mid d, \theta)$ for $j = E, T$, where θ is the model parameter vector. In the sequel, we often will suppress d and θ for brevity. The observed follow up time of X_j during the evaluation interval $[0,\ U_j]$ is $X_j^o = V \wedge X_j$. We assume that V is independent of X_j. The key to our method is that X_j^o provides useful information about Y_j, because

$$\Pr(Y_j = 1 \mid X_j^o = V < X_j) = 1 - \frac{\Pr(U_j < X_j)}{\Pr(X_j^o = V < X_j)}$$

must decrease as X_j^o increases from 0 to U_j. This fact also underlies the TiTE-CRM (Cheung and Chappell, 2000). To account for missing values, we use V to extend the previous definitions of Y_E and Y_T, as follows. We define

$$Y_j = \begin{cases} \mathit{missing} & \text{if } X_j > V \text{ and } V < U_j \quad (X_j^o = V), \\ 1 & \text{if } X_j \leq V \leq U_j \qquad\quad (X_j^o = X_j), \\ 0 & \text{if } X_j > V = U_j \qquad\quad (X_j^o = U_j). \end{cases}$$

That is, $Y_j = \mathit{missing}$ if the patient has not yet experienced the event and has not been fully followed to U_j, while Y_j is observed if the patient either has experienced the event ($Y_j = 1$) or has completed the defined follow-up time without the event ($V = U_j$ and $Y_j = 0$). Denote the missingship indicators $M_j = I(Y_j = \mathit{missing})$ for $j = E, T$, and $M = (M_E, M_T)$. The ith patient's data are $\mathcal{D}_i = (d_{[i]}, Y_{i,E}, Y_{i,T}, X_{i,E}^o, X_{i,T}^o)$, and $\mathcal{D}_i$ determines M_i. Our methodology

uses the actual interim data from n patients, by $\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \cdots, \mathcal{D}_n)$, to impute missing Y_j's and construct a completed binary data set $\boldsymbol{\mathcal{D}}(Y)$ for implementing the phase I-II method.

2.3 Event Time Distributions

To construct flexible survival functions, we assume piecewise exponential marginals for $[X_j|d, Y_j = 1]$, $j = E, T$, by partitioning $[0, U_j]$ into K_j intervals, $[0, h_{j,1})$, $[h_{j,1}, h_{j,2})$, $\cdots$, $[h_{j,K_j-1}, h_{j,K_j}]$, and assuming hazard $\lambda_{j,k}$ on $[h_{j,k-1}, h_{j,k})$. The marginal survival function for X_j is

$$S_j(x|Y_j = 1, \lambda_j) = \exp\left\{ -\sum_{k=1}^{K_j} w_{j,k}(x)\, \lambda_{j,k} \right\}, \qquad x > 0,$$

denoting $\lambda_j = (\lambda_{j,1}, \cdots, \lambda_{j,K_j})$, and weights $w_{j,k}(x) = h_{j,k} - h_{j,k-1}$ if $x > h_{j,k}$, $w_{j,k}(x) = x - h_{j,k-1}$ if $x \in [h_{j,k-1}, h_{j,k})$, and $w_{j,k}(x) = 0$ otherwise. Initially, we considered a more elaborate form of S_j, in which each $\lambda_{j,k}$ was replaced by $\lambda_{j,k}\gamma_j^d$ where $\gamma_j > 1$, so that the event time distributions varied with dose in a proportional hazard model with piecewise exponential baseline hazard. However, we found that fixing $\gamma_{_E} = \gamma_{_T} = 1$ did not change the method's operating characteristics, due to the fact that there is little or no information to estimate the γ_j's. To determine whether a more parsimonious survival model might give a design with similar performance, we replaced the piecewise exponential with a Weibull. However, simulations across eight dose-outcome scenarios (Supplementary Table S9) showed that the Weibull gave larger probabilities of incorrectly stopping early, and either similar or lower probabilities of selecting desirable doses.

Our imputation method requires the joint conditional survival probabilities $S(x_{_E}, x_{_T}|a, b) = \Pr(X_E > x_{_E}, X_T > x_{_T}|Y_E = a, Y_T = b)$ for $(a, b) = (0,1)$, $(1,0)$, $(1,1)$. Assuming conditional independence, $S_j(x_{_j}|Y_E, Y_T) = S_j(x_{_j}|Y_j)$ for $j = E, T$, implies that $S(x_{_E}, x_{_T}|1, 0)$ is determined by the marginal of X_E and $S(x_{_E}, x_{_T}|0, 1)$ is determined by the marginal of X_T. Determining $S(x_{_E}, x_{_T}|1, 1)$ requires accounting for association between X_E and X_T. We do this by defining a joint distribution using the Clayton copula (Clayton, 1978), given for $\phi \geq 0$ by

$$\begin{aligned} S(x_{_E}, x_{_T}|1, 1) &= \left\{ S_E(x_{_E}|Y_E = 1)^{-1/\phi} + S_T(x_{_T}|Y_T = 1)^{-1/\phi} - 1 \right\}^{-\phi} \\ &= \left\{ \exp\left\{ \sum_{k=1}^{K_E} \lambda_{E,k} w_{_{E,k}}/\phi \right\} + \exp\left\{ \sum_{k=1}^{K_T} \lambda_{T,k} w_{_{T,k}}/\phi \right\} - 1 \right\}^{-\phi}. \end{aligned}$$

The likelihood of $\mathcal{D}$ depends on Y_E, Y_T, and the censoring patterns of X_E and X_T. Denote $\lambda = (\lambda_E, \lambda_T)$, $\xi = (\phi+1)/\phi$, $\epsilon_j = I(X_j^o = X_j)$, the pdf of X_j by $f_j(\cdot)$, and

$$W(X_E^o, X_E^o) \ = S_E(X_E^o|Y_E=1)^{-1/\phi} + S_T(X_T^o|Y_T=1)^{-1/\phi} - 1.$$

When $Y_E = Y_T = 1$, there are four possible likelihoods,

$$L(\mathcal{D}|\lambda,\phi) = \begin{cases} \xi W(X_E^o, X_E^o)^{-\phi-2} \ \prod_{j=E}^{T} f_j(X_j^o|Y_j=1)\Big\{ \prod_{j=E}^{T} S_j(X_j^o|Y_j=1)\Big\}^{\xi} & \text{if } (\epsilon_{_E}, \epsilon_{_T}) = (1,1) \\ W(X_E^o, X_E^o)^{-\phi-1} f_E(X_E^o|Y_E=1) S_E(X_E^o|Y_E=1)^{\xi} & \text{if } (\epsilon_{_E}, \epsilon_{_T}) = (1,0) \\ W(X_E^o, X_E^o)^{-\phi-1} f_T(X_T^o|Y_T=1) S_T(X_T^o|Y_T=1)^{\xi} & \text{if } (\epsilon_{_E}, \epsilon_{_T}) = (0,1) \\ W(X_E^o, X_E^o)^{-\phi} & \text{if } (\epsilon_{_E}, \epsilon_{_T}) = (0,0) \end{cases}$$

For the cases $(Y_E, Y_T) = (1,0)$ or $(0,1)$, the likelihoods are given by

$$L(\mathcal{D}|\lambda,\phi) = \big[\{f_E(X_E^o|Y_E=1)\}^{\epsilon_E}\{S_E(X_E^o|Y_E=1)\}^{1-\epsilon_E}\big]^{Y_E(1-Y_T)}$$

$$\times \ \big[\{f_T(X_T^o|Y_T=1)\}^{\epsilon_T}\{S_T(X_T^o|Y_T=1)\}^{1-\epsilon_T}\big]^{Y_T(1-Y_E)}.$$

Denoting the likelihood for n patients by $L(\boldsymbol{\mathcal{D}}|\lambda,\phi) = \prod_{i=1}^{n} L(\mathcal{D}_i|\lambda,\phi)$, the posterior is $f(\lambda,\phi|\boldsymbol{\mathcal{D}})$ $\propto f(\lambda)f(\phi)L(\boldsymbol{\mathcal{D}}|\lambda,\phi)$, for priors $f(\lambda)$ and $f(\phi)$ of λ and ϕ.

3 Imputation Method

Let $\pi_{a,b}(d,\theta) = \Pr(Y_E = a, Y_T = b|d,\theta)$ for $a,b \in \{0,1\}$ denote the joint distribution of $[Y_E, Y_T \mid d]$, so $\pi_E(d,\theta) = \pi_{1,1}(d,\theta) + \pi_{1,0}(d,\theta)$ and $\pi_T(d,\theta) = \pi_{1,1}(d,\theta) + \pi_{0,1}(d,\theta)$. If no outcomes are missing the likelihood is the usual product

$$L(\boldsymbol{\mathcal{D}}(Y)|\theta) = \prod_{i=1}^{n}\prod_{a=0}^{1}\prod_{b=0}^{1}\big\{\pi_{a,b}(d_{[i]},\theta)\big\}^{I\{(Y_{i,E},Y_{i,T})=(a,b)\}}, \tag{1}$$

with posterior $f(\theta|\boldsymbol{\mathcal{D}}(Y)) \propto L(\boldsymbol{\mathcal{D}}(Y)|\theta)f(\theta)$, for prior $f(\theta)$. Since $L(\boldsymbol{\mathcal{D}}(Y)|\theta)$ cannot be computed if some $Y_{i,j}$'s are missing, we obtain a completed version of the likelihood, of the form (1), by using Bayesian data augmentation (Tanner and Wong, 1987). We iterate between an imputation step sampling missing Y_j's from their full conditionals, and a step computing the posterior using the completed data. Importantly, the imputation depends on the dose-outcome model.

Missing Y_j's are nonignorable because, at follow up time V, a patient who will not experience outcome j by U_j is more likely to have $Y_j = missing$ (M_j=1) than a patient for whom outcome j will occur (Yuan and Yin, 2011b). By Bayes' Law, $\Pr(M_j = 1|Y_j = 0) > \Pr(M_j = 1|Y_j = 1)$ implies that

$$\frac{\pi_j}{1-\pi_j} > \frac{\Pr(Y_j = 1|M_j = 1)}{\Pr(Y_j = 0|M_j = 1)}.$$

This says that the odds that $Y_j = 1$ decreases if Y_j is missing, so the missingship indicator M_j contains information about the future value of Y_j.

A complication in carrying out the imputation is that there are three possible missingship patterns: (1) $Y_E = missing$ and Y_T is observed, (2) $Y_T = missing$ and Y_E is observed, and (3) both $Y_E = missing$ and $Y_T = missing$. Since we impute missing values by sampling from their full conditional posteriors, these distributions must be specified for each missingship pattern. These posteriors are defined in terms of the following conditionals for X_E and X_T, given each of the four possible future outcome pairs, which for brevity we denote by

$$S_{ab} \;=\; \Pr(X_E > V, X_T > V|Y_E = a, Y_T = b), \qquad a, b \in \{0, 1\}. \tag{2}$$

In any case, $S_{00} = 1$ due to the fact that $\Pr(X_j > V|Y_j = 0) = 1$ for $0 \le V \le U_j$, which also implies that $S_{10} = \Pr(X_E > V|Y_E = 1, Y_T = 0)$ and $S_{01} = \Pr(X_T > V|Y_E = 0, Y_T = 1)$. Thus, only S_{11} involves a joint distribution for (X_E, X_T) given $Y_E = Y_T = 1$.

To obtain consistent estimates when imputing the missing Y_j's, we condition on the actual data $\mathcal{D}$. The following posterior conditional distributions are derived in the Appendix. When $Y_E = missing$ and Y_T is observed, we impute Y_E from its conditional posterior

$$\text{pr}(Y_E = 1|\mathcal{D}) = \left\{\frac{\pi_{1,1}S_{10}}{\pi_{1,1}S_{10} + \pi_{0,1}}\right\}^{Y_T} \left\{\frac{\pi_{1,0}S_{10}}{\pi_{1,0}S_{10} + \pi_{0,0}}\right\}^{1-Y_T}.$$

When $Y_T = missing$ and Y_E is observed, we impute Y_T from its conditional posterior

$$\text{pr}(Y_T = 1|\mathcal{D}) = \left\{\frac{\pi_{1,1}S_{01}}{\pi_{1,1}S_{01} + \pi_{1,0}}\right\}^{Y_E} \left\{\frac{\pi_{0,1}S_{01}}{\pi_{0,1}S_{01} + \pi_{0,0}}\right\}^{1-Y_E}.$$

When both Y_E and Y_T are missing, we impute (Y_E, Y_T) from the joint conditional posterior

$$\text{pr}(Y_E = y, Y_T = z|\mathcal{D}) = \frac{\pi_{y,z}S_{yz}}{\sum_{a=0}^{1}\sum_{b=0}^{1}\pi_{a,b}S_{ab}}, \quad \text{for} \;\; y, z = 0, 1$$

At the posterior step using the completed data, we sample parameters from their full conditional posteriors in two steps: (1) Sample θ from $f(\theta|\boldsymbol{\mathcal{D}}(Y))$ and (2) sample ϕ and $\lambda_{j,k}$ for each $k = 1, \cdots, K_j$ and $j = (E, T)$ from $f(\lambda, \phi|\boldsymbol{\mathcal{D}})$. This is iterated until the Markov chain converges, with posteriors computed using adaptive rejection Metropolis sampling (Gilks et al., 1995).

We now turn to Case 2, in which Y_E is evaluated only at U_E. Analytically, Case 2 is much simpler than Case 1 because in Case 2 efficacy is Y_E with no random event time involved, and Y_E is missing completely at random (MCAR) at all $V < U_E$. For the same reason, trial conduct in Case 2 is much harder logistically than in Case 1. This is because there is no X_E in Case 2 and hence no partial information about Y_E when $V < U_E$. Inference for π_E relies entirely on observed Y_E values from previously treated patients, while Y_T may be imputed by exploiting the event time data (X_T^o, δ_T) using the marginal of X_T.

Our proposed method can be applied to any phase I-II combination of probability model and decision rules based on $[Y_E, Y_T \mid d]$ with delayed outcomes. To make things concrete, we apply the method to the Efficacy-Toxicity (EffTox) trade-off method of Thall and Cook (2004). Case 1 is illustrated by the trial of chemotherapy for acute leukemia where $U_T = U_E = 42$ days, and Case 2 by the stem cell transplantation trial with $U_T = 30$ days and $U_E = 90$ days.

4 EffTox Designs

4.1 The Design Assuming Complete Data

We first briefly review the EffTox phase I-II design (Thall and Cook, 2004) assuming complete data $\mathcal{D}(Y)$. The model for $[Y_E, Y_T \mid d]$ assumes marginal probabilities $\pi_E = \text{logit}^{-1}\{\mu_E + \beta_{E,1}d + \beta_{E,2}d^2\}$ and $\pi_T = \text{logit}^{-1}\{\mu_T + \beta_{T,1}d + \beta_{T,2}d^2\}$ and uses a Gumbel copula (Murtaugh and Fisher, 1990) to obtain a joint distribution,

$$\pi_{a,b} = (\pi_E)^a(1-\pi_E)^{1-a}(\pi_T)^b(1-\pi_T)^{1-b} + (-1)^{a+b}\pi_E(1-\pi_E)\pi_T(1-\pi_T)\left(\frac{e^\psi - 1}{e^\psi + 1}\right), \qquad (3)$$

where ψ parameterizes association between Y_E and Y_T. To prevent mis-estimation with small sample sizes, we require that $\beta_{E,1} + 2\beta_{E,2}d > 0$ and $\beta_{T,1} + 2\beta_{T,2}d > 0$, corresponding to agents where both π_T and π_E increase with dose.

The *desirability* of a dose is defined by first defining the desirability $\delta(\pi_E, \pi_T)$ of each probability pair $\pi = (\pi_E, \pi_T)$ in $[0,\ 1]^2$, with $\delta(\pi_E, \pi_T)$ increasing in π_E and decreasing in π_T. Denoting the posterior means $\pi^*_{j,d} = \mathrm{E}\{\pi_j(d,\theta)|\mathcal{D}(Y)\}$ for $j = E, T$, the desirability of d is $\delta(\pi^*_{E,d}, \pi^*_{T,d})$, for use as a decision criterion. The function δ may be obtained from a target efficacy-toxicity tradeoff contour, $\mathcal{C}$, in $[0,\ 1]^2$ in several ways (cf. Thall, Cook and Estey, 2006). To avoid doses that are too toxic or inefficacious, two admissibility criteria are imposed. Given elicited fixed lower limit $\underline{\pi}_E$ on π_E and upper limit $\bar{\pi}_T$ on π_T, a dose d is *acceptable* if

$$\Pr\{\pi_E(d,\theta) > \underline{\pi}_E|\mathcal{D}(Y)\} > p_{_E} \quad \text{and} \quad \Pr\{\pi_T(d,\theta) < \bar{\pi}_T|\mathcal{D}(Y)\} > p_{_T} \tag{4}$$

for prespecified cutoffs $p_{_E}$ and $p_{_T}$. The trial starts at a dose chosen by the physician and each new cohort is treated with the acceptable dose having largest $\delta(d)$. An untried dose may not be skipped when escalating, and when escalating to an untried dose only the toxicity admissibility rule is imposed. If no dose is acceptable the trial is stopped with no dose selected.

We refer to the new version of the EffTox design that incorporates our proposed imputation methodology as the late onset (LO)-EffTox design. It differs from the EffTox design in one fundamental way, namely that event time data are exploited to compute posterior decision criteria using the Bayesian data augmentation methodology described in Section 3.

4.2 Prior Specification

In any model-based Bayesian adaptive clinical trial design, the prior must be sufficiently vague that the accumulating data dominates the posterior distribution, and thus the adaptive decisions. Thall and Cook (2004) provided a framework for establishing priors from elicited mean values of $\pi_E(d,\theta)$ and $\pi_T(d,\theta)$ that solves for prior hyperparameters using nonlinear least squares. We derive a prior based on the weakly informative prior for logistic regression proposed by Gelman et al. (2008). Following Gelman et al. (2008), for priors we assume $\mu_T, \mu_E, \beta_{T,1}, \beta_{E,1}, \beta_{T,2}, \beta_{E,2} \sim$ iid Cauchy$(0, 2.5)$, and then shift the six Cauchy prior location parameters from 0 to $\tilde{\mu} = (\tilde{\mu}_{\mu_T}, \tilde{\mu}_{\beta_{T,1}}, \tilde{\mu}_{\beta_{T,2}}, \tilde{\mu}_{\mu_E}, \tilde{\mu}_{\beta_{E,1}}, \tilde{\mu}_{\beta_{E,2}})$, to reflect prior opinion. To obtain $\tilde{\mu}$, similarly to Thall and Cook (2004), we first elicit the means $\tilde{m}_{E,d_r}$ and $\tilde{m}_{T,d_r}$ of $\pi_E(d_r,\theta)$ and $\pi_T(d_r,\theta)$ from the physician for each dose $d_r, r = 1, ..., R$. For each $j = E, T$, we use least squares to solve for $(\tilde{\mu}_{\mu_j}, \tilde{\mu}_{\beta_{j,1}}, \tilde{\mu}_{\beta_{j,2}})$

by assuming $\mathrm{E}\{\mathrm{logit}(\tilde{m}_{j,d_r})\} = \tilde{\mu}_{\mu_j} + \tilde{\mu}_{\beta_{j,1}} d_r + \tilde{\mu}_{\beta_{j,2}} d_r^2$.

To obtain a vague prior on λ, we assume $[X_j|Y_j = 1] \sim \mathrm{Unif}(0, U_j)$, which implies that the hazard at the midpoint of the subinterval $[h_{j,k-1},\ h_{j,k})$ of the partition is $\tilde{\lambda}_{j,k} = K_j/\{U_j(K_j - k + 0.5)\}$. We assume that $\lambda_{j,k} \sim Gam(\tilde{\lambda}_{j,k}/C, 1/C)$, where $Gam(a, b)$ denotes the gamma distribution with mean a/b and variance a/b^2. Thus, $\lambda_{j,k}$ has prior mean $\tilde{\lambda}_{j,k}$ and variance $C\tilde{\lambda}_{j,k}$, so C is a tuning parameter that determines $\mathrm{var}(\lambda_{j,k})$ and that can be calibrated by simulation. In preliminary simulations, we found that $C = 2$ yields a reasonably vague prior and a design with good operating characteristics. Finally, we assume $\psi \sim \mathrm{Normal}(0, 1)$ and $\phi \sim Gam(0.2, 0.2)$.

5 Computer Simulations

5.1 Clinical Trial Designs

We simulated our proposed methodology to study its behavior when applied to the EffTox design for phase I-II trials in each of Cases 1 and 2. To assess robustness, we conducted a second set of simulations in which we varied several model, design, and simulation scenario parameters. In all simulations, we considered hypothetical phase I-II trials with five raw doses (2.5, 5.0, 7.5, 10.0, 12.5). In the model, we replaced each raw dose d_r^o by the standardized dose $d_r = (.5/s)\{\log(d_r^o) - \overline{\log(d^o)}\}$, where s is the standard deviation of the centered log doses, so $d_1, \cdots, d_5$ are centered around 0 and have standard deviation .5. The trade-off contour, $\mathcal{C}$, was determined by fitting a quadratic curve to the trade-off target probability pairs $(\pi_E, \pi_T) = (0.15, 0), (0.45, 0.20), (1, 0.60)$, which gives target contour function $\pi_T = -0.0952 + 0.6239\pi_E + 0.0713\pi_E^2$. Figure 1 illustrates the target contour by a solid line, with contours on which all (π_E, π_T) have the same desirability $\delta(\pi_E, \pi_T)$ shown as dashed lines. Dose acceptability was determined by $\underline{\pi}_E = 0.25$ for efficacy and $\bar{\pi}_T = 0.35$ for toxicity, with decision cutoffs $p_E = p_T = 0.10$. In Case 2, a dose could be declared inefficacious due to the first inequality in (4) being violated only after at least one cohort was fully evaluated at that dose. In the first set of simulations, we considered trials with 16 cohorts of size three, so $N_{\max} = 48$. In practice, $N_{\max}$ should be chosen via simulation by doing a sensitivity analysis that evaluates a design's properties over a range of practically feasible $N_{\max}$ values.

The following designs are constructed to mimic dose-finding trials of either chemotherapy for acute leukemia in Case 1, or an agent that is part of a preparative regimen in the stem cell transplantation trial for multiple myeloma in Case 2. Recall that, in Case 1, X_E may occur at any time during [0, U_E], and, in Case 2, Y_E is observed at U_E. For Case 1, we assumed that U_E = U_T= 6 weeks with accrual rate $\alpha = 1.5$ patients per week. For Case 2, we assumed $U_E = 90$ days, (12.85 weeks), toxicity evaluation interval $U_T = 30$ days (4.3 weeks), and accrual rate $\alpha =$ 2.1 patients per week. The logistical difficulty indices are $\zeta = 1.5 \times 6 = 9$ for Case 1 and $\zeta =$ $2.1 \times 12.85 = 27$ for Case 2. We kept all other parameters for Case 2 the same as those in Case 1.

We assumed prior means 0.15, 0.20, 0.25, 0.30, 0.35 for $\pi_E(d_1, \theta)$, ..., $\pi_E(d_5, \theta)$ and 0.15, 0.20, 0.27, 0.35, 0.45 for $\pi_T(d_1, \theta)$, ..., $\pi_T(d_5, \theta)$. Applying the method described in Section 4.2, this gave location parameters $(\tilde{\mu}_{\mu_E}, \tilde{\mu}_{\beta_{E,1}}, \tilde{\mu}_{\beta_{E,2}}) = (-1.21,\ 0.96,\ 0.35)$ and $(\tilde{\mu}_{\mu_T}, \tilde{\mu}_{\beta_{T,1}}, \tilde{\mu}_{\beta_{T,2}})$ $= (-1.16,\ 1.39,\ 0.85)$ for the shifted Cauchy priors. For the Gamma piecewise exponential event rate priors, we assumed $K = 6$. The formula in Section 4.2 gives $(\tilde{\lambda}_{E,1}, \cdots, \tilde{\lambda}_{E,6}) =$ $(\tilde{\lambda}_{T,1}, \cdots, \tilde{\lambda}_{T,6}) =$ (0.182, 0.222, 0.286, 0.400, 0.667, 2.000) in Case 1, and $(\tilde{\lambda}_{T,1}, \cdots, \tilde{\lambda}_{T,6}) =$ (0.364, 0.444, 0.571, 0.800, 1.333, 4.000) in Case 2. We used tuning parameter $C = 2$. Thus, for example, $\lambda_{E,1} \sim Gam(0.093, 0.5)$ in Case 1.

5.2 Simulation Study Design

Each simulation scenario was specified in terms of assumed true efficacy and toxicity probabilities at each dose, $\pi_j(d_r)^{true}$, for $j = E, T$ and $r = 1, ..., 5$. We modeled association between X_E and X_T by assuming a Clayton copula (Clayton, 1978) with $\phi = 1.0$. We considered eight scenarios, illustrated in Figure 2. In Scenario 8, no dose is admissible, since d = 1,2 are inefficacious and d = 3,4,5 are too toxic. For the first simulation study, we generated X_E, X_T from Weibull distributions. For each j and d_r, the Weibull scale and shape parameters were chosen so that (1) $\Pr(X_j \leq U_j | d_r)^{true} = \pi_j(d_r)^{true}$ and (2) $\pi_{j,late}(d_r)^{true} = \Pr(U_j/2 \leq X_j \leq U_j | d_r)^{true} = 0.50$, that is, 50% of the events occurred in the second half of the evaluation interval. Because $\pi_E(d_r)^{true}$ and $\pi_T(d_r)^{true}$ vary with d_r in each scenario, the scale and shape parameters of the corresponding true Weibull distributions both vary with d_r. Each scenario was simulated 1000 times.

5.3 Simulation Results for Case 1

For Case 1, we compared the LO-EffTox design to three methods that are used in practice to deal with the late onset problem. The first method is the "One Level Down" rule. With this method, if some of the patients treated at the current optimal dose, $d^{opt} = d_r$ have not yet been evaluated fully, i.e. $Y_{i,E,d_r} = missing$ or $Y_{i,T,d_r} = missing$, then any new patient is treated at d_{r-1}. The second method is the "Look Ahead" rule (Thall et al., 1999) which says that, for each possible value $\tilde{\mathcal{D}}_{\text{mis}}(Y)$ that $\mathcal{D}_{\text{mis}}(Y)$ may take on, use the completed data $\mathcal{D}_{\text{obs}}(Y) \cup \tilde{\mathcal{D}}_{\text{mis}}(Y)$ to compute d^{opt}. If this dose is the same for all possible $\tilde{\mathcal{D}}_{\text{mis}}(Y)$, then use that dose to treat the next patient immediately. Otherwise, the only two options for new patients are to make them wait to be treated, which usually is impossible in practice, or to turn them away and treat them off protocol. The third method uses all complete cases, where both Y_E and Y_T are observed, to compute d^{opt} and treat the next patient immediately.

We will use two summary criteria to evaluate each method's performance and compare the three methods. Denote the true desirability of dose d_r by δ_r^{true} and the true set of acceptable doses by $\mathcal{A}^{true}$. The first criterion is the *desirability-weighted selected percentage*,

$$\bar{\delta} = \frac{\sum_{r=1}^{5} \delta_r^{true} \Pr(\text{select } d_r) I(d_r \in \mathcal{A}^{true})}{\sum_{r=1}^{5} \delta_r^{true} \; I(d_r \in \mathcal{A}^{true})},$$

which quantifies dose selection reliability and thus potential benefit for future patients. The second criterion is the ratio N_E/N_T where N_E and N_T denote the number of patients who experienced efficacy and toxicity, respectively. This criterion quantifies benefit to the patients in the trial, hence may be considered an index of ethical desirability.

Table 1 gives the operating characteristics of the LO-EffTox design and three competing methods. The percentage of trials with no dose selected is denoted by "None," with the numbers of patients turned away from the trial by the Look Ahead method given in parentheses. In general, LO-EffTox outperforms the One Level Down, Look Ahead rule, and Complete Case method. The One Level Down rule has much smaller correct selection rate and mean N_E. The Look Ahead rule design performs roughly equivalently to LO-EffTox in terms of dose selection percentages and N_E/N_T, but the trial durations under the Look Ahead rule are dramatically larger compared to LO-EffTox. This is because the Look Ahead rule turns many patients away,

while LO-EffTox treats all patients. Compared to the Complete Case method, LO-EffTox has either similar or higher correct selection percentages and more patients treated at doses with higher desirability. In Scenario 8, where all doses are inefficacious or too toxic, LO-EffTox has by far the largest probability of correctly stopping early and selecting no dose.

Figure 3 illustrates the results in Table 1 in terms of $\bar{\delta}$ plotted on the vertical axis and N_E/N_T on the horizontal axis, under each of Scenarios 1 – 7. Scenario 8 is not included in Figure 3 since in this case no dose is acceptable, so $\bar{\delta}$ is not relevant. Values in the upper right portion of the figure are more desirable, while values in the lower left are less desirable. Figure 3 shows that the One Level Down rule produces designs with very poor properties, in terms of both $\bar{\delta}$ and N_E/N_T. These two criteria are roughly equivalent for LO-EffTox and the Look Ahead version of EffTox for each of Scenarios 1 – 6, In Scenario 7, which has true desirability not monotone in dose, LO-EffTox has much greater $\bar{\delta}$ and much smaller N_E/N_T compared to the "Look Ahead" version of EffTox, so in this case there is no clear winner. However, since the Look Ahead rule turns away many patients and produces a very long trial, the apparent equivalence in terms of the two criteria in Figure 3 only tells part of the story. Compared to the Complete Case method, LO-EffTox has either similar or much larger $\bar{\delta}$ values and similar N_E/N_T.

5.4 Simulation Results for Case 2

For Case 2, we compared the LO-EffTox design to the One Level Down, Look Ahead, and Complete Case rules, and also the TiTE-CRM (Cheung and Chappell, 2000). We included the TiTE-CRM because, when the efficacy indicator Y_E is observed at U_E, the only time-to-event variables are X_T and V, which are the basis for the TiTE-CRM, so the TiTE-CRM is a reasonable alternative in Case 2. To implement the TiTE-CRM, we assumed the dose-toxicity model $\pi_T(d_r) = p_r^{\exp(\alpha)}$ with fixed skeleton $(p_1, \cdots, p_5) = (0.15, 0.20, 0.27, 0.35, 0.45)$, parameter α having $N(0, \sigma^2 = 2)$ prior, and target toxicity probability 0.35.

The simulation results for the Case 2 under six of the eight scenarios are summarized in Table 2. Results for the other two scenarios are summarized in Table S1. Table 2 shows that, across all scenarios, LO-EffTox greatly outperforms the One Level Down, Look Ahead, and Complete Case methods, in terms of both selection percentages and N_E/N_T. The "One Level Down" rule

is most likely to select d_3 or lower doses, and almost never selects d_4 or d_5, regardless of their δ values. It thus greatly sacrifices efficacy in many cases, as shown by the extremely small N_E values. As in Case 1, the "Look Ahead" rule has selection percentages similar to those of the LO-EffTox design, but the price is a much longer trial with many patients turned away. Because the TiTE-CRM design completely ignores efficacy, it is unlikely to select doses having acceptable toxicity and high efficacy, which is the case in nearly all scenarios. As in Case 1, for Scenario 8 LO-EffTox has by far the largest probability of correctly stopping early and selecting no dose.

5.5 Sensitivity Analyses

To investigate robustness, we conducted additional simulations in which we varied each of $\pi_{j,late}(d)^{true}$, logistical difficulty index ζ, $N_{\max}$, the event time distributions, K = number of sets in the piecewise exponential event time distribution partition, and the association parameter, ϕ. Each sensitivity analysis was conducted for all scenarios, and the results are summarized in Supplementary Tables S2 – S8. Table S2 shows that the design is not sensitive to changes in values of $\pi_{j,late}(d)^{true}$ over the range 0.10 to 0.90, illustrated in Figure 4a. Tables S3 and S4 summarize sensitivity to the logistical difficulty index for the values $\zeta = 3.0$ to 24.0 in Case 1 and $\zeta = 9.0$ to 54.0 in Case 2. As ζ increases, the ratio N_E/N_T and $\bar{\delta}$ both decrease, illustrated in Figure 4b. Table S5 shows, as expected, that the design's performance improves with larger $N_{\max}$, illustrated in Figure 4c. Simulations with X_E and X_T generated from several combinations of the Weibull and Log-logistic(Table S6) showed that LO-EffTox is robust to the true event time distributions. Table S7 shows that the number of sets in the partition of the piecewise exponential has little effect on the method's performance for $K = 6$ to 12. Varying the association parameter from $\phi = 0.1$ to 2.5 had almost no effect on performance (see Table S8).

6 Discussion

We have proposed a general methodology to address the problem of late-onset outcomes in phase I-II clinical trials. The method treats unobserved binary outcomes are as nonignorable missing data, uses data augmentation to impute the missing outcomes, and applies the design's decision

rules using the completed data. Simulations show that, in most cases, the proposed design performs better than alternative approaches to the late onset problem in phase I-II trials. Our results suggest that the general approach of imputing binary vectors $\mathbf{Y}$ by utilizing time-to-event variables used to define $\mathbf{Y}$ may improve the logistics of any outcome-adaptive procedure based on the distribution of $\mathbf{Y}$.

Appendix

The probability of $Y_E = 1$ with the known Y_T value is

$$\begin{aligned}
\text{pr}(Y_E = 1 | X_E > V, Y_T) &= \frac{\text{pr}(Y_E = 1)\text{pr}(Y_T, X_E > V | Y_E = 1)}{\sum_{a=0}^{1} \text{pr}(Y_E = a)\text{pr}(Y_T, X_E > V | Y_E = a)} \\
&= \frac{\text{pr}(Y_E = 1)\text{pr}(Y_T | Y_E = 1)\text{pr}(X_E > V | Y_T, Y_E = 1)}{\sum_{a=0}^{1} \text{pr}(Y_E = a)\text{pr}(Y_T | Y_E = a)\text{pr}(X_E > V | Y_T, Y_E = a)} \\
&= \frac{\text{pr}(Y_T, Y_E = 1)\text{pr}(X_E > V | Y_E = 1)}{\sum_{a=0}^{1} \text{pr}(Y_T, Y_E = a)\text{pr}(X_E > V | Y_E = a)} \\
&= \begin{cases} \frac{\pi_{1,1}\text{pr}(X_E > V | Y_E = 1)}{\pi_{1,1}\text{pr}(X_E > V | Y_E = 1) + \pi_{01}} = \frac{\pi_{1,1} S_{10}}{\pi_{1,1} S_{10} + \pi_{0,1}} & \text{if } Y_T = 1, \\ \frac{\pi_{1,0}\text{pr}(X_E > V | Y_E = 1)}{\pi_{1,0}\text{pr}(X_E > V | Y_E = 1) + \pi_{0,0}} = \frac{\pi_{1,0} S_{10}}{\pi_{1,0} S_{10} + \pi_{0,0}} & \text{if } Y_T = 0. \end{cases}
\end{aligned}$$

where S_{ab} and $\pi_{a,b}$ are defined by equation (2) and (3), respectively.

The probability of $Y_T = 1$ with the known Y_E value is

$$\begin{aligned}
\text{pr}(Y_T = 1 | X_T > V, Y_E) &= \frac{\text{pr}(Y_T = 1)\text{pr}(Y_E, X_T > V | Y_T = 1)}{\sum_{b=0}^{1} \text{pr}(Y_T = b)\text{pr}(Y_E, X_T > V | Y_T = b)} \\
&= \frac{\text{pr}(Y_T = 1)\text{pr}(Y_E | Y_T = 1)\text{pr}(X_T > V | Y_E, Y_T = 1)}{\sum_{b=0}^{1} \text{pr}(Y_T = b)\text{pr}(Y_E | Y_T = b)\text{pr}(X_T > V | Y_E, Y_T = b)} \\
&= \frac{\text{pr}(Y_E, Y_T = 1)\text{pr}(X_T > V | Y_T = 1)}{\sum_{b=0}^{1} \text{pr}(Y_E, Y_T = b)\text{pr}(X_T > V | Y_T = b)} \\
&= \begin{cases} \frac{\pi_{1,1}\text{pr}(X_T > V | Y_T = 1)}{\pi_{1,1}\text{pr}(X_T > V | Y_T = 1) + \pi_{1,0}} = \frac{\pi_{1,1} S_{01}}{\pi_{1,1} S_{01} + \pi_{1,0}} & \text{if } Y_E = 1, \\ \frac{\pi_{0,1}\text{pr}(X_T > V | Y_T = 1)}{\pi_{0,1}\text{pr}(X_T > V | Y_T = 1) + \pi_{0,0}} = \frac{\pi_{0,1} S_{01}}{\pi_{0,1} S_{01} + \pi_{0,0}} & \text{if } Y_E = 0. \end{cases}
\end{aligned}$$

where S_{ab} and $\pi_{a,b}$ are defined by equation (2) and (3), respectively.

When Y_T and Y_E are unknown, the probability of (Y_E, Y_T) is

$$\begin{aligned}
&\text{pr}(Y_E = 1, Y_T = 1 | X_E > V, X_T > V) \\
&\quad = \frac{\text{pr}(X_E > V, X_T > V | Y_E = 1, Y_T = 1)\text{pr}(Y_E = 1, Y_T = 1)}{\sum_{a=0}^{1}\sum_{b=0}^{1}\text{pr}(X_E > V, X_T > V | Y_E = a, Y_T = b)\text{pr}(Y_E = a, Y_T = b)} \\
&\quad = \frac{\pi_{1,1}S_{11}}{\sum_{a=0}^{1}\sum_{b=0}^{1}\pi_{a,b}S_{ab}}, \\
&\text{pr}(Y_E = 1, Y_T = 0 | X_E > V, X_T > V) \\
&\quad = \frac{\text{pr}(X_E > V, X_T > V | Y_E = 1, Y_T = 0)\text{pr}(Y_E = 1, Y_T = 0)}{\sum_{a=0}^{1}\sum_{b=0}^{1}\text{pr}(X_E > V, X_T > V | Y_E = a, Y_T = b)\text{pr}(Y_E = a, Y_T = b)} \\
&\quad = \frac{\pi_{1,0}S_{10}}{\sum_{a=0}^{1}\sum_{b=0}^{1}\pi_{a,b}S_{ab}}, \\
&\text{pr}(Y_E = 0, Y_T = 1 | X_E > V, X_T > V) \\
&\quad = \frac{\text{pr}(X_E > V, X_T > V | Y_E = 0, Y_T = 1)\text{pr}(Y_E = 0, Y_T = 1)}{\sum_{a=0}^{1}\sum_{b=0}^{1}\text{pr}(X_E > V, X_T > V | Y_E = a, Y_T = b)\text{pr}(Y_E = a, Y_T = b)} \\
&\quad = \frac{\pi_{0,1}S_{01}}{\sum_{a=0}^{1}\sum_{b=0}^{1}\pi_{a,b}S_{ab}}.
\end{aligned}$$

where where S_{ab} and $\pi_{a,b}$ are defined by equation (2) and (3), respectively, and $S_{00} = 1$.

Acknowledgments

This research was supported by NIH/NCI Cancer Center Support Grant CA016672 36. Yuan and Jin acknowledge support from NIH grant R01 CA154591. Peter Thall's research was supported by NIH/NCI grant R01 CA 83932.

References

Babb, J., A. Rogatko, and S. Zacks (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine 17*, 1103–1120.

Bekele, B. N., Y. Ji, Y. Shen, and P. F. Thall (2008). Monitoring late-onset toxicities in phase I trials using predicted risks. *Biostatistics 9*, 442–457.

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

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

Braun, T. M. (2006). Generalizing the TiTE-CRM to adapt for early- and late-onset toxicities. *Statistics in Medicine 25*, 2071–2083.

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

Cheung, Y. and R. Chappell (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics 56*, 1177–1182.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 141–151.

Conaway, M., S. Dunbar, and S. Peddada (2004). Design for single- or multiple-agent phase I trials. *Biometrics 60*, 661–669.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics 38*, 143–151.

Gehan, E. A. (1969). Estimating survival functions from the life table. *Journal of Chronic Diseases 21*, 629–644.

Gelman, A., A. Jakulin, M. G. Pittau, and Y. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*, 1360–1383.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics 44*, 455–472.

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

Murtaugh, P. and I. Fisher (1990). Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics, Part A - Theory and Methods 19*, 2003–2020.

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

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

Simon, R. M. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials 10*, 1–10.

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

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association 82*, 528–550.

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

Thall, P., J. Lee, C.-H. Tseng, and E. E.H. (1999). Accrual strategies for phase I trials with delayed patient outcome. *Statistics in Medicine 18*, 1155–1169.

Thall, P. F. and K. E. Russell (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics 54*, 251–264.

Thall, P. F. and R. M. Simon (1994). Practical bayesian guidelines for phase IIB clinical trials. *Biometrics 50*, 337–349.

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

Yin, G., Y. Li, and Y. Ji (2006). Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics 62*, 777–787.

Yuan, Y. and G. Yin (2011a). Bayesian phase I/II adaptively randomized oncology trials with combined drugs. *The Annals of Applied Statistics 5*, 924–942.

Yuan, Y. and G. Yin (2011b). Robust EM continual reassessment method in oncology dose finding. *Journal of the American Statistical Association 106*, 818–831.

Zhang, W., D. J. Sargent, and S. Mandrekar (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine 25*, 2365–2383.

Table 1: Simulation study comparing the four phase I-II designs in Case 1. The probability pairs given in parentheses for each dose level are $(\pi_E(d_r)^{true},\ \pi_T(d_r)^{true})$.

Method		Dose Level 1	2	3	4	5	None	N_E/N_T	Duration (Weeks)
	Scenario 1	(0.05, 0.03)	(0.10, 0.05)	(0.20, 0.07)	(0.25, 0.08)	(0.35, 0.10)			
	δ^{true}	0.52	0.53	0.58	0.61	0.67			
LO-EffTox	% selected	0.0	0.1	2.6	2.5	91.7	3.1	11.8/3.8	36.9
	# patients	6.15	6.80	6.94	4.96	22.80			
One Down	% selected	0.0	0.7	6.0	16.1	76.6	0.6	5.7/2.4	37.0
	# patients	18.89	13.69	10.42	4.46	0.50			
Look Ahead	% selected	0.0	0.2	2.4	2.5	88.7	6.2	13.1/4.0	69.7
	# patients	3.40	3.84	4.30	3.97	30.90	(52.07)		
Complete Case	% selected	0.0	0.2	2.4	3.4	91.8	2.2	9.8/3.3	37.6
	# patients	11.18	8.11	6.76	5.43	16.40			
	Scenario 2	(0.02, 0.10)	(0.10, 0.15)	(0.40, 0.20)	(0.45, 0.30)	(0.50, 0.60)			
	δ^{true}	0.43	0.43	0.59	0.52	0.32			
LO-EffTox	% selected	0.0	11.9	61.6	20.5	5.0	1.0	15.5/13.1	36.9
	# patients	5.90	9.31	13.79	9.60	9.14			
One Down	% selected	0.0	1.0	5.1	24.3	69.1	0.5	8.3/ 7.6	37.2
	# patients	18.73	13.51	10.49	4.67	0.51			
Look Ahead	% selected	0.0	8.0	67.5	19.0	4.6	0.9	16.7/12.5	74.2
	# patients	3.39	7.14	20.58	9.95	6.67	(57.54)		
Complete Case	% selected	0.0	8.9	44.3	27.4	18.5	0.9	13.1/11.9	37.0
	# patients	11.27	10.19	9.86	8.56	7.95			
	Scenario 3	(0.30, 0.10)	(0.35, 0.20)	(0.45, 0.40)	(0.50, 0.60)	(0.55, 0.65)			
	δ^{true}	0.63	0.55	0.43	0.32	0.31			
LO-EffTox	% selected	75.0	22.5	2.3	0.2	0.0	0.0	16.2/8.9	37.4
	# patients	26.21	14.21	5.11	1.88	0.59			
One Down	% selected	77.6	14.4	4.7	2.6	0.0	0.7	15.5/7.3	37.1
	# patients	31.30	12.79	3.34	0.33	0.00			
Look Ahead	% selected	63.8	32.6	3.3	0.1	0.0	0.2	16.4/8.9	62.4
	# patients	24.79	16.39	5.32	1.23	0.23	(39.65)		
Complete Case	% selected	78.5	17.2	2.0	1.2	0.0	1.1	15.7/ 8.1	37.0
	# patients	28.99	12.44	4.61	1.27	0.30			
	Scenario 4	(0.18, 0.20)	(0.28, 0.24)	(0.55, 0.28)	(0.74, 0.31)	(0.79, 0.33)			
	δ^{true}	0.44	0.46	0.62	0.76	0.78			
LO-EffTox	% selected	3.4	3.0	7.1	19.5	65.2	1.8	25.5/13.1	36.8
	# patients	9.63	6.87	7.77	9.98	13.08			
One Down	% selected	9.1	5.8	9.2	29.3	40.4	6.2	13.9/10.6	35.7
	# patients	23.22	12.06	7.57	2.76	0.23			
Look Ahead	% selected	8.5	6.1	11.1	21.0	51.8	1.5	26.5/13.2	70.8
	# patients	9.45	5.48	6.77	9.85	15.83	(52.81)		
Complete Case	% selected	6.9	2.8	8.2	18.5	57.6	6.0	20.3/11.8	35.7
	# patients	15.70	7.79	8.02	7.67	6.73			

Method		Dose Level 1	2	3	4	5	None	N_E/N_T	Duration (Weeks)
	Scenario 5	(0.20, 0.10)	(0.50, 0.19)	(0.52, 0.23)	(0.54, 0.44)	(0.56, 0.54)			
	δ^{true}	0.55	0.69	0.66	0.45	0.38			
LO-EffTox	% selected	13.9	46.8	31.4	6.8	0.6	0.5	21.1/10.2	37.3
	# patients	11.35	16.31	14.22	4.63	1.33			
One Down	% selected	10.8	22.6	39.3	22.1	4.0	1.2	17.5/ 7.7	37.2
	# patients	21.61	16.20	8.29	1.42	0.03			
Look Ahead	% selected	10.7	48.3	37.7	3.1	0.1	0.1	21.9/10.4	76.1
	# patients	8.23	18.61	15.77	4.36	0.98	(60.26)		
Complete Case	% selected	13.2	40.7	34.4	9.7	1.1	0.9	19.6/ 9.4	37.2
	# patients	15.67	15.23	11.96	3.97	0.85			
	Scenario 6	(0.20, 0.10)	(0.50, 0.19)	(0.52, 0.34)	(0.54, 0.44)	(0.56, 0.54)			
	δ^{true}	0.55	0.69	0.53	0.45	0.38			
LO-EffTox	% selected	20.0	60.6	17.5	1.6	0.2	0.1	20.4/10.8	37.5
	# patients	13.32	19.09	11.05	3.38	1.13			
One Down	% selected	14.9	37.5	35.1	9.7	1.6	1.2	17.3/ 8.4	36.9
	# patients	21.78	17.25	7.68	0.83	0.01			
Look Ahead	% selected	16.1	65.4	17.1	1.2	0.1	0.1	21.2/10.8	73.8
	# patients	10.18	23.41	10.76	2.87	0.74	(56.73)		
Complete Case	% selected	21.7	50.3	21.5	5.3	0.4	0.8	18.9/ 9.7	37.3
	# patients	17.69	17.31	9.65	2.50	0.56			
	Scenario 7	(0.02, 0.10)	(0.05, 0.25)	(0.30, 0.30)	(0.40, 0.55)	(0.50, 0.70)			
	δ^{true}	0.43	0.34	0.42	0.31	0.27			
LO-EffTox	% selected	0.0	3.5	60.4	8.3	1.7	26.1	11.0/15.5	33.4
	# patients	6.27	9.08	16.71	6.28	5.91			
One Down	% selected	10.9	4.3	10.9	29.7	37.4	6.8	13.9/10.7	35.5
	# patients	23.48	11.90	7.48	2.62	0.19			
Look Ahead	% selected	9.6	8.3	22.1	47.2	11.8	1.0	25.4/14.9	74.3
	# patients	10.00	6.26	9.31	15.72	6.38	(57.84)		
Complete Case	% selected	0.0	3.0	49.3	11.4	6.6	29.7	8.8/13.8	34.1
	# patients	11.27	12.21	11.94	5.28	4.44			
	Scenario 8	(0.02, 0.10)	(0.05, 0.25)	(0.35, 0.55)	(0.40, 0.60)	(0.50, 0.70)			
	δ^{true}	0.43	0.34	0.29	0.28	0.27			
LO-EffTox	% selected	0.0	1.8	17.0	0.8	0.1	80.3	8.0/14.5	25.7
	# patients	6.09	10.12	13.10	4.13	2.40			
One Down	% selected	0.0	1.8	33.9	14.8	7.2	42.3	5.4/11.9	34.3
	# patients	18.71	14.62	9.86	2.28	0.07			
Look Ahead	% selected	0.0	1.8	18.7	0.5	0.1	78.9	8.3/13.6	59.7
	# patients	3.33	7.93	16.40	3.73	1.26	(50.76)		
Complete Case	% selected	0.0	2.3	22.1	2.9	1.2	71.5	7.3/13.9	29.6
	# patients	11.25	12.78	10.99	4.07	1.91			

Table 2: Simulation study comparing the five phase I-II designs in Case 2. The probability pairs given in parentheses for each dose level are $(\pi_E(d_r)^{true},\ \pi_T(d_r)^{true})$.

Method		Dose Level 1	2	3	4	5	None	N_E/N_T	Duration (Days)
	Scenario 1	(0.05, 0.03)	(0.10, 0.05)	(0.20, 0.07)	(0.25, 0.08)	(0.35, 0.10)			
	δ^{true}	0.52	0.53	0.58	0.61	0.67			
LO-EffTox	% selected	0.0	0.4	1.1	2.2	90.9	5.4	12.7/3.9	237
	# patients	3.58	4.25	5.99	4.91	27.75			
One Down	% selected	0.0	4.8	94.2	1.0	0.0	0.0	2.9/1.6	246
	# patients	38.74	9.23	0.03	0.00	0.00			
Look Ahead	% selected	0.0	0.0	1.6	1.9	91.0	5.5	13.3/4.0	399
	# patients	3.25	3.49	4.54	6.11	29.26	(49.2)		
Complete Case	% selected	0.0	2.7	7.3	6.3	80.7	3.0	6.2/2.4	244
	# patients	28.96	3.64	3.91	3.35	7.97			
TiTE-CRM	% selected	0.0	0.0	0.0	0.0	99.9	0.1	14.2/4.3	240
	# patients	3.22	3.23	3.18	3.40	34.92			
	Scenario 2	(0.02, 0.10)	(0.10, 0.15)	(0.40, 0.20)	(0.45, 0.30)	(0.50, 0.60)			
	δ^{true}	0.43	0.43	0.59	0.52	0.32			
LO-EffTox	% selected	0.0	9.4	62.5	21.4	5.4	1.3	16.4/12.6	244
	# patients	3.35	8.30	18.77	10.09	7.03			
One Down	% selected	0.0	1.8	96.3	1.4	0.0	0.5	1.6/5.2	246
	# patients	38.72	9.17	0.04	0.00	0.00			
Look Ahead	% selected	0.1	8.9	56.9	27.9	5.7	0.5	17.3/13.2	436
	# patients	3.27	6.46	17.54	13.38	7.24	(58.8)		
Complete Case	% selected	0.0	5.3	29.2	32.0	33.2	0.3	7.9/9.1	246
	# patients	29.11	3.65	5.10	5.09	5.04			
TiTE-CRM	% selected	0.0	0.7	14.0	70.8	13.3	1.2	18.3/15.7	238
	# patients	3.99	4.39	8.27	18.97	11.89			
	Scenario 3	(0.30, 0.10)	(0.35, 0.20)	(0.45, 0.40)	(0.50, 0.60)	(0.55, 0.65)			
	δ^{true}	0.63	0.55	0.43	0.32	0.31			
LO-EffTox	% selected	68.1	27.9	3.4	0.4	0.0	0.2	16.4/9.2	247
	# patients	26.19	13.53	5.82	1.99	0.46			
One Down	% selected	44.9	8.0	47.0	0.1	0.0	0.0	14.8/5.6	246
	# patients	38.98	8.99	0.02	0.00	0.00			
Look Ahead	% selected	58.3	32.2	8.7	0.7	0.0	0.1	16.7/9.7	420
	# patients	22.21	16.28	7.54	1.69	0.23	(54.0)		
Complete Case	% selected	85.1	11.3	2.0	1.5	0.1	0.0	14.9/6.2	247
	# patients	39.77	5.24	2.39	0.57	0.03			
TiTE-CRM	% selected	1.1	30.1	62.3	5.1	0.0	1.4	20.1/17.4	238
	# patients	5.32	12.13	18.08	8.98	2.90			

Method		Dose Level 1	2	3	4	5	None	N_E/N_T	Duration (Days)
	Scenario 4	(0.18, 0.20)	(0.28, 0.24)	(0.55, 0.28)	(0.74, 0.31)	(0.79, 0.33)			
	δ^{true}	0.44	0.46	0.62	0.76	0.78			
LO-EffTox	% selected	3.5	3.8	10.4	20.9	61.0	0.4	27.0/13.5	246
	# patients	9.02	5.84	7.97	11.51	13.52			
One Down	% selected	17.0	7.5	73.1	1.2	0.0	1.2	9.6/9.9	245
	# patients	38.64	9.09	0.05	0.00	0.00			
Look Ahead	% selected	6.4	6.1	14.8	23.1	48.5	1.1	27.1/13.4	445
	# patients	7.58	6.17	9.38	10.55	13.93	(62.0)		
Complete Case	% selected	11.4	6.1	7.8	20.1	54.4	0.2	15.3/11.0	246
	# patients	31.64	4.12	3.79	3.82	4.60			
TiTE-CRM	% selected	1.1	5.6	19.1	28.0	40.2	6.0	25.9/12.8	231
	# patients	6.37	6.89	9.16	10.50	12.62			
	Scenario 6	(0.20, 0.10)	(0.50, 0.19)	(0.52, 0.34)	(0.54, 0.44)	(0.56, 0.54)			
	δ^{true}	0.55	0.69	0.53	0.45	0.38			
LO-EffTox	% selected	20.3	63.4	13.9	1.8	0.4	0.2	20.8/10.9	246
	# patients	11.86	20.78	10.87	3.45	0.96			
One Down	% selected	8.0	12.4	79.2	0.4	0.0	0.0	12.4/5.6	247
	# patients	38.73	9.24	0.03	0.00	0.00			
Look Ahead	% selected	16.7	52.2	26.3	3.8	0.9	0.1	21.6/11.5	402
	# patients	9.10	20.03	13.71	4.26	0.86	(48.6)		
Complete Case	% selected	27.8	40.2	22.8	7.1	2.0	0.1	14.4/7.6	246
	# patients	32.84	7.63	5.38	1.82	0.33			
TiTE-CRM	% selected	0.0	11.8	51.6	30.2	4.6	1.8	23.5/16.7	237
	# patients	4.47	7.42	14.68	13.57	7.10			
	Scenario 8	(0.02, 0.10)	(0.05, 0.25)	(0.35, 0.55)	(0.40, 0.60)	(0.50, 0.70)			
	δ^{true}	0.43	0.34	0.29	0.28	0.27			
LO-EffTox	% selected	0.0	1.9	8.1	0.4	0.1	89.5	7.9/13.6	117
	# patients	3.36	7.21	13.99	4.18	1.79			
One Down	% selected	0.0	4.2	90.1	1.4	0.0	4.3	1.3/6.1	247
	# patients	38.62	9.31	0.05	0.00	0.00			
Look Ahead	% selected	0.0	1.6	18.8	1.1	0.1	78.4	8.2/13.4	367
	# patients	3.22	8.10	16.11	3.79	1.17	(53.5)		
Complete Case	% selected	0.0	8.2	33.5	8.7	3.9	45.7	5.0/10.6	204
	# patients	28.81	4.85	7.18	2.84	1.15			
TiTE-CRM	% selected	5.8	58.7	31.7	0.9	0.0	2.9	9.1/17.5	236
	# patients	8.00	17.23	14.29	5.65	1.69			

Figure 1: Desirability Contours: The target contour is represented by a solid line. Other contours on which all (π_E, π_T) have the same desirability are shown as dashed lines.

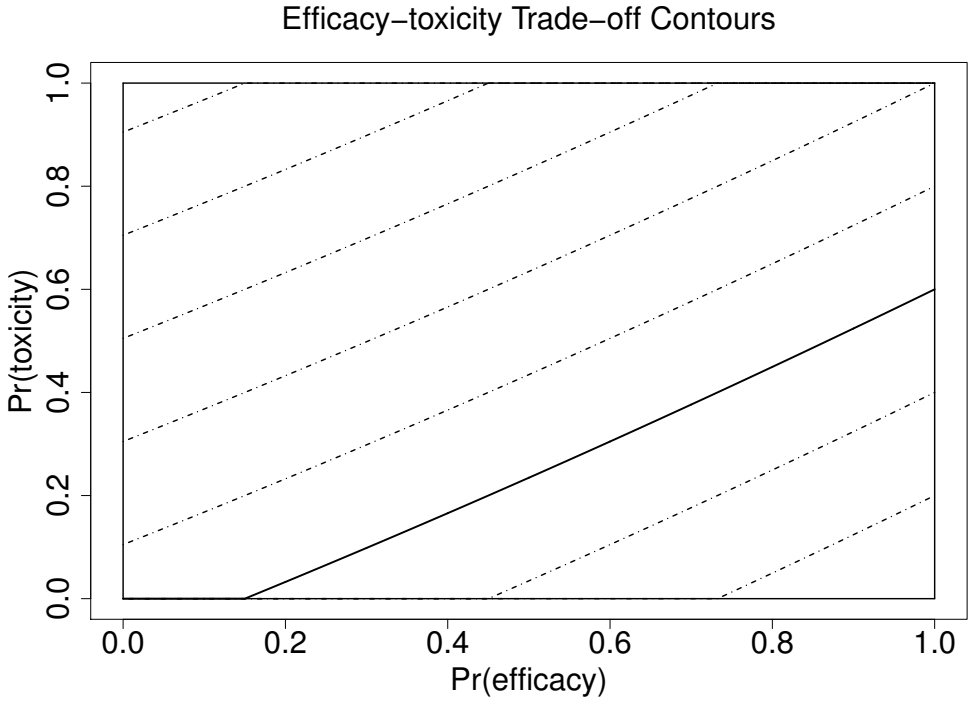


Figure 2: Illustration of scenarios: The solid lines with circles represent true probabilities of efficacy and the dashed lines with triangles represent true probabilities of toxicity. The horizontal lines represent $\bar{\pi}_T$ and $\underline{\pi}_E$.

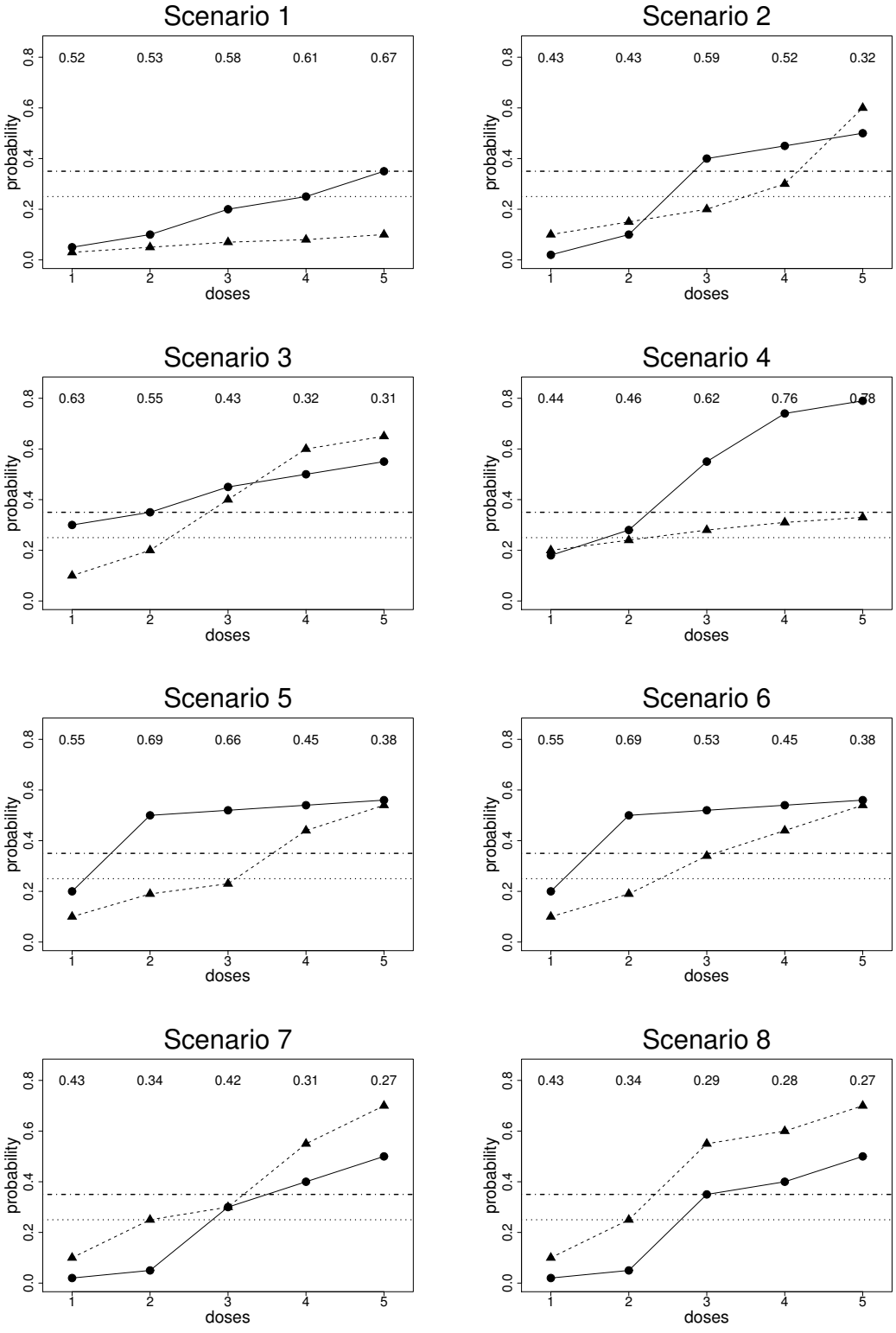


Figure 3: Comparison of the Late Onset Eff-Tox, One Level Down, Look Ahead, and Complete Case methods. The X-axis is $\bar{\delta}$, the desirability weighted selection percentage, and the Y-axis is N_E/N_T. Simulation scenarios are identified by integers.

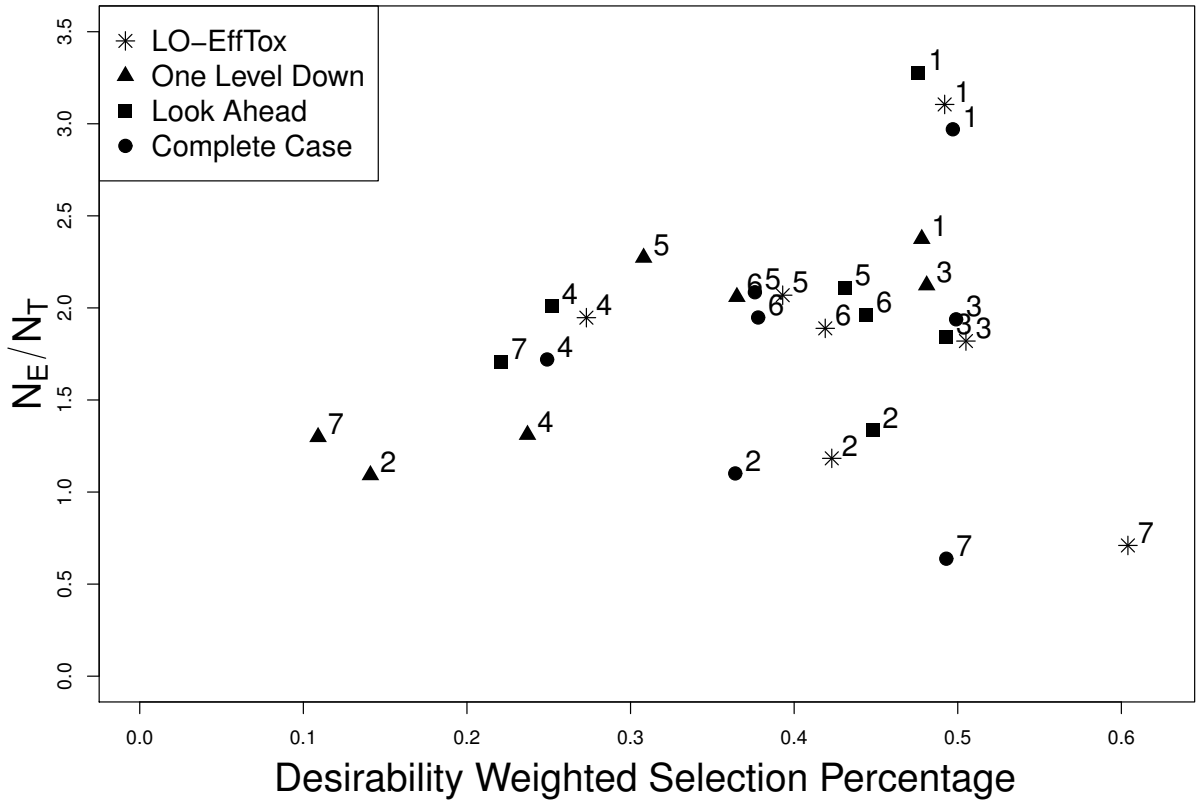
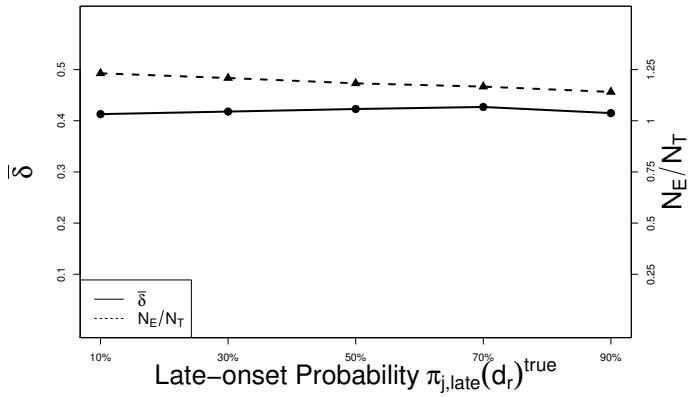
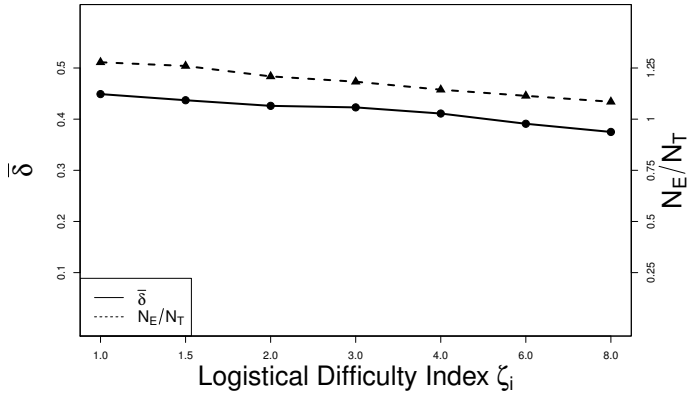


Figure 4: Sensitivity analyses under Scenario 2. Solid lines represent the desirability weighted selection percentages $\bar{\delta}$. Dashed lines represent the means of N_E/N_T.

(a)



(b)



(c)