Statistics and Its Interface Volume 8 (2015) 217–226

A Bayesian phase I/II clinical trial design in the presence of informative dropouts

Beibei Guo, Yong Zang, and Ying Yuan*

A phase I/II trial design utilizes both toxicity and efficacy outcomes to make the decision of dose assignment for patients. Because assessing the efficacy endpoint often requires a relatively long follow-up time, phase I/II trials are more susceptible to the missing data problem caused by informative dropouts that are correlated with treatment efficacy and toxicity. In addition, patient outcomes may not be scored quickly enough to apply decision rules that choose treatments or doses for newly accrued patients. To address these issues, we propose a Bayesian phase I/II design that jointly models efficacy, toxicity, and dropout as time-to-event data. Correlations among the three time-to-event outcomes are taken into account by a shared frailty. This joint model strategy accounts for the informative dropouts and has an additional advantage of accommodating a high accrual rate without suspending patient enrollment when toxicity or efficacy outcomes require a long follow-up. Under the Bayesian paradigm, we continuously update the posterior estimate of the model and assign incoming patients to the most desirable dose based on an efficacy-toxicity trade-off utility. Simulation studies show that the proposed design has good operating characteristics with a high probability of selecting the target dose and assigning the most patients to the target dose.

Keywords and phrases: Bayesian adaptive design, Missing data, Nonignorable dropout, Dose finding, Trade-off.

1. INTRODUCTION

Phase I clinical trials aim to identify the maximum tolerated dose (MTD) of an investigational drug [1, 2, 3, 4, 5, 6, 7], while phase II clinical trials aim to examine the potential efficacy of a new drug based on the MTD obtained from the phase I trials [8, 9, 10, 11, 12, 13]. Traditionally, phase I and phase II trials are conducted separately to assess toxicity and efficacy independently; however, this conventional approach has several drawbacks. For example, the dose-toxicity function is estimated unreliably in phase I due to its small sample size. In addition, informal dose adjustments often are made in phase II if excessive toxicity is observed, which invalidates the assumed properties of any efficacy-based design. Lastly, the conventional approach ignores the inherent trade-off between efficacy and toxicity that characterizes the way physicians make therapeutic decisions.

Recently, combined phase I/II trial designs have drawn increasing attention. This class of designs merges two separate phases of a trial into one phase and assesses toxicity and efficacy simultaneously. Compared to the conventional phase-I-followed-by-phase-II paradigm, phase I/II designs are more efficient in using the information in the data, and appropriately reflect the realistic trade-off between efficacy and toxicity that physicians consider when making therapeutic decisions in clinical practice. Thall and Russell [14] proposed a Bayesian phase I/II design that characterized patient outcome using a trinary ordinal variable to account for both efficacy and toxicity. Braun [15] generalized the continual reassessment method [2] to accommodate efficacy and toxicity simultaneously. Thall and Cook [16] developed the EffTox design based on trade-offs between the probabilities of treatment efficacy and toxicity. In the EffTox design, toxicity and efficacy are modeled jointly as a bivariate variable and the doses are selected for successive patient cohorts based on a set of efficacy-toxicity trade-off contours that partition the two-dimensional outcome probability domain. Bekele and Shen [17] proposed a joint model between a binary toxicity outcome and a continuous biomarker expression outcome by introducing latent Gaussian variables in a probit model. Yin, Li and Ji [18] proposed using the odds ratio of efficacy and toxicity as the measure of desirability for phase I/II trials. Yuan and Yin [19] developed a Bayesian phase I/II time-to-event (TTE) design to accommodate delayed (or late-onset) toxicity and efficacy by jointly modeling them as time-to-event outcomes. Yuan and Yin [20] proposed a phase I/II design for drug combination trials using adaptive randomization.

Because assessing the efficacy endpoint often requires a relatively long follow-up time, the duration of phase I/II trials is often longer than that of phase I trials. As a result, phase I/II trials are more susceptible to the missing toxicity and efficacy outcomes caused by patient dropouts. Such missing data pose a major logistical impediment to implementing the trials because the decision rules of most existing phase I/II trial designs require the availability of the outcomes of the patients who have been enrolled in the trial in order to apply the design rules to choose treatments

*Corresponding author.

or doses for newly accruing patients. Moreover, the missing data caused by dropouts are often informative or nonignorable in the sense that the dropout probability of a patient depends on his/her toxicity and efficacy outcomes. For example, patients who experience higher toxicity and lower efficacy are more likely to drop out of the trial than patients who experience lower toxicity and higher efficacy. In this case, the simple approach of ignoring the missing data results in biased estimates of toxicity and efficacy, thereby causing inappropriate dose assignment and selection.

Our motivating example is an acute leukemia clinical trial, which investigates six doses of suberoylanilide hydroxamic acid (SAHA), i.e., 200, 400, 600, 800, 1,000, 1,200mg, combined with a fixed dose of fludarabine (10mg/m^2). SAHA is a histone deacetylase inhibitor that acts by inhibiting cell growth and inducing apoptosis of cancer cells; and fludarabine is a chemotherapy drug that inhibits tumor growth by interfering with ribonucleotide reductase and DNA polymerase. Toxicity is evaluated during the 45 days after the initiation of the treatment. Toxicity is defined as a clinically significant non-hematologic adverse event or abnormal laboratory value assessed as unrelated to disease progression, intercurrent illness, or concomitant medications based on Common Terminology Criteria for Adverse Events (CTCAE) v4.0. Efficacy requires 90 days to be scored. Efficacy is defined as complete remission (i.e., leukemic blasts in the bone marrow $\leq$ 5% and absolute neutrophil count $\geq$ 1,000/ul) or marrow complete remission (i.e., leukemic blasts in the bone marrow $\leq$ 5% and absolute neutrophil count $<$ 1,000/ul). The total sample size is 60 patients and patients are treated in cohorts of size 3. Due to the relatively long follow-up time, a substantial number of patients are expected to drop out of the trial, and these dropouts are believed to depend on the patient's toxicity and efficacy status. In addition, efficacy is a "delayed" outcome with respect to the expected accrual rate of 3 patients per month. One may expect to accrue 9 patients before efficacy is scored for even the first patient, so applying any adaptive rule to choose a dose for patients 4, 5, and 6, (cohort 2), based on the data from patients 1, 2, and 3, (cohort 1), is not possible without delaying the start of therapy for the second cohort.

We propose a Bayesian phase I/II clinical trial design that accommodates informative dropouts and delayed outcomes. We treat toxicity, efficacy and dropout as time-to-event outcomes and jointly model them using proportional hazards models with a shared frailty. At each decision-making time, patients who drop out of the trial without experiencing toxicity and/or efficacy are considered as informative censoring. Based on the observed data, we adaptively assign patients to the dose with the highest posterior mean of the utility that trades off between toxicity and efficacy. Simulation studies show that the proposed design has good operating characteristics and selects the target dose with a high probability.

The remainder of this article is organized as follows. Section 2 presents the joint frailty model for the times to efficacy, toxicity, and dropout; and the dose-finding algorithm. Section 3 examines the operating characteristics of our new design through simulation studies. We provide concluding remarks in Section 4.

2. METHOD

2.1 Probability model

Consider a phase I/II trial with J doses, $d_1 < d_2 < \ldots < d_J$, under investigation. A total of N patients are sequentially enrolled and treated in the trial. Each patient will be followed for fixed periods of T_1 and T_2 to evaluate efficacy and toxicity, respectively. Conventionally, efficacy and toxicity are defined as binary outcomes that take a value of 1 (or 0) depending on whether the corresponding event is (or is not) observed within the follow-up periods. The existing phase I/II designs based on binary outcomes typically require that efficacy and toxicity outcomes be immediately ascertainable without any dropouts.

In order to handle dropouts and potential delayed outcomes, we treat efficacy and toxicity as time-to-event outcomes. Because dropouts are often associated with toxicity and efficacy and censor these outcomes, the dropout problem here can also be regarded as an informative censoring problem. One way to handle informative censoring is to jointly model the times to toxicity and efficacy and the censoring (i.e., dropout) process. Let t_1, t_2 and t_3 denote the times to efficacy, toxicity and dropout, respectively, and let $h_1(t_1|Z)$, $h_2(t_2|Z)$, and $h_3(t_3|Z)$ denote the corresponding hazard functions given dosage $Z \in (d_1, \ldots, d_J)$. For the ith patient administered dosage Z_i, we jointly model the times to efficacy, toxicity and dropout using proportional hazards models [21] as follows,

$$
\begin{aligned}
(1) \quad h_1(t_1|Z_i,\theta_i) &= \lambda_1(t_1)\exp(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2) \\
(2) \quad h_2(t_2|Z_i,\theta_i) &= \lambda_2(t_2)\exp(\alpha_2\theta_i + \beta_2 Z_i) \\
(3) \quad h_3(t_3|Z_i,\theta_i) &= \lambda_3(t_3)\exp(\theta_i + \beta_3 Z_i),
\end{aligned}
$$

where $\lambda_1(t_1)$, $\lambda_2(t_2)$, and $\lambda_3(t_3)$ are baseline hazard functions; and β_1, β_2, β_3, and γ are regression parameters characterizing the dose effects. In the hazard function for efficacy $h_1(t_1|Z_i,\theta_i)$, we include a quadratic term γZ_i^2 to accommodate possibly non-monotone dose-efficacy curves, e.g., for biological agents. The common frailty θ_i shared by the three hazard functions is used to capture the potential correlations among the times to efficacy, toxicity and dropout. We assume that θ_i follows a normal distribution with mean 0 and variance σ^2, i.e., $\theta_i \sim N(0, \sigma^2)$. To allow for flexibility such that the correlations between the different endpoints can be either positive, negative, or 0, we introduce parameters α_1 and α_2 in equations (1) and (2). A positive (or negative) value of α_1 indicates a positive (or negative) correlation between efficacy and dropout, and a positive (or negative) value of α_2 indicates a positive (or negative) correlation between toxicity and dropout. When $\alpha_1 = \alpha_2 = 0$, the three time-to-event outcomes are independent. As the sample size of phase I/II trials is typically small, we take

a parsimonious parametric approach by assuming that the baseline hazards follow exponential distributions with constant hazards, i.e., $\lambda_k(t_k) = \lambda_k$, for $k = 1, 2$, and 3.

Under the proposed time-to-event model, the efficacy and toxicity rates of dose Z (at the end of the follow-up periods) are given by cumulative distribution functions $F_1(T_1|Z) = \text{pr}(t_1 \le T_1|Z)$ and $F_2(T_2|Z) = \text{pr}(t_2 \le T_2|Z)$, respectively. To measure the desirability of a dose, we define the following utility as a trade-off between toxicity and efficacy

$$U(Z) = F_1(T_1|Z) - \frac{1}{w}F_2(T_2|Z),$$

where constant $w > 0$ can be interpreted as the additional percentages of toxicity that patients are willing to tolerate in exchange for one percentage increase of efficacy. For example, $w = 1.5$ means that patients are willing to tolerate additional 1.5% of toxicity in exchange for 1% increase of efficacy. In practice, the value of w should be elicited from clinicians or patients. For example, we can ask clinicians to provide two pairs of equivalently desirable efficacy and toxicity probabilities, say (p_{11}, p_{12}) and (p_{21}, p_{22}), based on which we determine the value of w as

$$w = \frac{p_{12} - p_{22}}{p_{11} - p_{21}}.$$

The goal of our design is to find the dose with the highest desirability, i.e., the target dose with the highest value of $U(Z)$, which also satisfies certain minimal safety and efficacy requirements.

Let t_{1i}, t_{2i} and t_{3i} denote the times to efficacy, toxicity and dropout, respectively, for the ith patient, and η_i denote the time to administrative censoring. Define the actual observed time $y_{ki} = \min(t_{ki}, t_{3i}, \eta_i)$, for $k = 1, 2$, $y_{3i} = \min(t_{3i}, \eta_i)$, and censoring indicator $\delta_{ki} = I(t_{ki} \le \min(t_{3i}, \eta_i))$ for $k = 1, 2$, $\delta_{3i} = I(t_{3i} \le \eta_i)$. Note that dropout (i.e., t_3) can censor toxicity and efficacy (i.e., t_1 and t_2), but not vice versa. The likelihood for the ith patient with data $D_i = (y_{ki}, \delta_{ki})$ is

$$\begin{aligned} L(D_i|\Theta) &= \left\{\lambda_1 \text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2)\right\}^{\delta_{1i}} \\ &\times \text{exp}(-\lambda_1 y_{1i} \text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2)) \\ &\times \left\{\lambda_2 \text{exp}(\alpha_2\theta_i + \beta_2 Z_i)\right\}^{\delta_{2i}} \\ &\times \text{exp}(-\lambda_2 y_{2i} \text{exp}(\alpha_2\theta_i + \beta_2 Z_i) \\ &\times \left\{\lambda_3 \text{exp}(\theta_i + \beta_3 Z_i)\right\}^{\delta_{3i}} \\ &\times \text{exp}(-\lambda_3 y_{3i} \text{exp}(\theta_i + \beta_3 Z_i)), \end{aligned}$$

where $\Theta = (\lambda_1, \beta_1, \gamma, \lambda_2, \beta_2, \lambda_3, \beta_3, \theta_i, \sigma^2, \alpha_1, \alpha_2)$. Let $p(\Theta)$ denote the joint prior distribution for Θ and assume that the parameters are mutually independent *a priori*, then the joint posterior distribution of Θ based on n treated patients is

$$p(\Theta|\text{data}) \propto p(\Theta) \prod_{i=1}^{n} L(D_i|\Theta).$$

2.2 Posterior inference

We assume that the components of Θ are mutually independent *a priori* and follow the prior distributions:

$$\begin{aligned} &\lambda_1 \sim Gamma(a_1, b_1), \quad \lambda_2 \sim Gamma(a_2, b_2), \\ &\lambda_3 \sim Gamma(a_3, b_3), \quad \beta_1 \sim N(0, \tau_1^2), \quad \beta_2 \sim N(0, \tau_2^2), \\ &\beta_3 \sim N(0, \tau_3^2), \quad \gamma \sim N(0, \tau_4^2), \quad \sigma^2 \sim InvGamma(a_4, b_4), \\ &\alpha_1 \sim Unif(-c_1, c_1), \quad \alpha_2 \sim Unif(-c_2, c_2), \end{aligned}$$

where $Gamma(a, b)$ denotes a Gamma distribution with shape parameter a and inverse scale parameter b, $InvGamma(a, b)$ denotes an inverse Gamma distribution with shape parameter a and scale parameter b, and $Unif(-c, c)$ denotes a uniform distribution with support $[-c, c]$. To obtain vague priors, we set $a_1 = a_2 = a_3 = a_4 = 0.1$, $b_1 = b_2 = b_3 = b_4 = 0.1$, $\tau_1^2 = \tau_2^2 = \tau_3^2 = \tau_4^2 = 100$, and $c_1 = c_2 = 5$, such that the posterior distributions of the parameters will be dominated by the observed data.

We sample the posterior distribution of Θ using Gibbs sampler. Let θ generically denote the parameters that we condition upon, and let D denote the data for the n patients who are already in the trial, i.e., $D = \{(y_{ki}, \delta_{ki}), i = 1, \cdots, n\}$. We sequentially sample the elements of Θ from the following full conditional distributions:

1. $[\lambda_1|D, \theta] \sim Gamma(a_1 + \sum_{i=1}^n \delta_{1i}, b_1 + \sum_{i=1}^n y_{1i}\text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2))$
2. $[\beta_1|D, \theta] \propto \text{exp}\{\beta_1 \sum_{i=1}^n Z_i\delta_{1i} - \lambda_1 \sum_{i=1}^n y_{1i}\text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2) - {\beta_1}^2/(2\tau_1^2)\}$
3. $[\gamma|D, \theta] \propto \text{exp}\{\gamma \sum_{i=1}^n Z_i^2\delta_{1i} - \lambda_1 \sum_{i=1}^n y_{1i}\text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2) - \gamma^2/(2\tau_4^2)\}$
4. $[\lambda_2|D, \theta] \sim Gamma(a_2 + \sum_{i=1}^n \delta_{2i}, b_2 + \sum_{i=1}^n y_{2i}\text{exp}(\alpha_2\theta_i + \beta_2 Z_i))$
5. $[\beta_2|D, \theta] \propto \text{exp}\{\beta_2 \sum_{i=1}^n Z_i\delta_{2i} - \lambda_2 \sum_{i=1}^n y_{2i}\text{exp}(\alpha_2\theta_i + \beta_2 Z_i) - {\beta_2}^2/(2\tau_2^2)\}$
6. $[\lambda_3|D, \theta] \sim Gamma(a_3 + \sum_{i=1}^n \delta_{3i}, b_3 + \sum_{i=1}^n y_{3i}\text{exp}(\theta_i + \beta_3 Z_i))$
7. $[\beta_3|D, \theta] \propto \text{exp}\{\beta_3 \sum_{i=1}^n Z_i\delta_{3i} - \lambda_3 \sum_{i=1}^n y_{3i}\text{exp}(\theta_i + \beta_3 Z_i) - {\beta_3}^2/(2\tau_3^2)\}$
8. $[\theta_i|D, \theta] \propto \text{exp}\{\delta_{1i}\alpha_1\theta_i + \delta_{2i}\alpha_2\theta_i + \delta_{3i}\theta_i - \lambda_1 y_{1i}\text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2) - \lambda_2 y_{2i}\text{exp}(\alpha_2\theta_i + \beta_2 Z_i) - \lambda_3 y_{3i}\text{exp}(\theta_i + \beta_3 Z_i) - \frac{\theta_i^2}{2\sigma^2}\}$
9. $[\sigma^2|D, \theta] \sim InvGamma(n/2 + a_4, b_4 + \sum_{i=1}^n \theta_i^2/2)$
10. $[\alpha_1|D, \theta] \propto \text{exp}\{\alpha_1 \sum_{i=1}^n \delta_{1i}\theta_i - \lambda_1 \sum_{i=1}^n (y_{1i}\text{exp}(\alpha_1\theta_i + \beta_1 Z_i + \gamma Z_i^2))\} I(-c_1 \le \alpha_1 \le c_1)$
11. $[\alpha_2|D, \theta] \propto \text{exp}\{\alpha_2 \sum_{i=1}^n \delta_{2i}\theta_i - \lambda_2 \sum_{i=1}^n (y_{2i}\text{exp}(\alpha_2\theta_i + \beta_2 Z_i))\} I(-c_2 \le \alpha_2 \le c_2)$

2.3 Dose-finding algorithm

A challenge for dose-finding trials is that very limited information is available at the beginning of the trial when only a few patients have been treated. This lack of data becomes more severe when patients drop out of the trial. As

a result, at the beginning of the trial, it is difficult to reliably estimate model parameters and make correct decisions on dose assignments. To facilitate the trial process, we propose a two-stage dose-finding algorithm, for which stage I is rule-based while stage II is model-based. The goal of stage I is to collect some preliminary data for later model fitting. We treat patients in cohorts of size 3 and start the trial by treating the first cohort of patients at the lowest dose d_1. The dose-escalation rule for stage I is similar to that of the traditional "3 + 3" design and is described as follows. At the current dose d_j,

1. If 2 out of 3 patients experience toxicity, stage I is completed and the trial moves forward to stage II with starting dose d_{j-1}. If $d_j = d_1$, i.e., d_1 is the lowest dose, the trial is terminated.
2. If 1 out of 3 patients experiences toxicity, stage I is completed and the trial moves forward to stage II with starting dose d_j.
3. If 0 out of 3 patients experiences toxicity, the dose is escalated to d_{j+1}. However, if $d_j = d_J$, i.e., d_j is the highest dose, stage I is completed and the trial moves forward to stage II with starting dose d_j.

At stage I, the data are extremely sparse and we have little knowledge on the toxicity profile of the drug. To be conservative and protect patients from overly toxic doses, we impose two additional safety restrictions: (1) the toxicity of enrolled patients must be fully assessed before we enroll the next cohort of patients; and (2) if a patient drops out before his/her toxicity outcome is observed, we add a new patient to the cohort, replacing the position of the patient who dropped out.

Stage II involves model-based dose finding. Let ϕ_E and ϕ_T be the respective lower efficacy limit and upper toxicity limit as pre-specified by physicians, and let n denote the number of patients who have been enrolled into the trial at the moment of decision making for assigning a dose to a newly enrolled cohort. To safeguard against treating patients at futile or overly toxic doses, we define the admissible dose set $\mathcal{A}$ as a set of doses satisfying both the efficacy requirement,

$$\text{pr}(F_1(T_1|d) > \phi_E|\text{data}) > a_E + b_E n/N, \tag{4}$$

and the toxicity requirement,

$$\text{pr}(F_2(T_2|d) < \phi_T|\text{data}) > a_T + b_T n/N, \tag{5}$$

where a_E, b_E, a_T and b_T are non-negative tuning parameters that can be calibrated by simulation to achieve good design operating characteristics. We set the posterior probability cutoffs (i.e., $a_E + b_E n/N$ and $a_T + b_T n/N$) to depend on the sample size n such that the toxicity and efficacy requirements adaptively become more stringent when more patients are enrolled in the trial. Such a choice is made based on the following considerations: at the beginning of the trial, the estimates of toxicity and efficacy probabilities are highly unreliable, so we should be lenient regarding the admissible requirements; however, when data accumulate and we have more reliable estimates, we should be more stringent regarding the safety and efficacy requirements.

Our model-based dose-finding algorithm can be described as follows. Assume that l cohorts of patients have been enrolled in the trial. Let d_h be the current highest tried dose, and C_T be the dose escalation cutoff based on toxicity. To assign a dose to the incoming $(l+1)$th cohort:

1. We calculate the posterior probability of toxicity of d_h based on the data obtained from the first l cohorts. If $\text{pr}(F_2(T_2|d_h) < \phi_T|\text{data}) > C_T$ and $d_h \neq d_J$, we escalate the dose and assign the $(l+1)$th cohort to d_{h+1}.
2. Otherwise, we assign the $(l+1)$th cohort to the dose from $\mathcal{A}$ with the highest desirability, i.e., the largest value of $U(Z)$. At any time, if $\mathcal{A}$ is empty, we terminate the trial.
3. We continue the above dose assignment process for subsequent cohorts until the sample size is exhausted. We select the dose in $\mathcal{A}$ with the largest value of $U(Z)$ as the final recommended dose.

3. SIMULATION

To assess the performance of our proposed design, we conducted extensive simulation studies. We considered six doses $(0.2, 0.4, 0.6, 0.8, 1.0, 1.2\text{g})$ as in our motivating trial, with a maximum sample size of 60 patients. The toxicity upper bound was $\phi_T = 0.3$ and the efficacy lower bound was $\phi_E = 0.2$. We used weight $\omega = 1$, which was chosen based on the two pairs of equally desirable efficacy-toxicity probabilities, $(0.3, 0.1)$ and $(0.4, 0.2)$, elicited from the physician. The follow-up time was 3 months for evaluating efficacy and 1.5 months for evaluating toxicity (i.e., $T_1 = 3$; $T_2 = 1.5$), and patient accrual followed a Poisson process with a rate of 3 patients per month. We set $\alpha_1 = -1$ and $\alpha_2 = 1$ such that patients who have high/low probability of experiencing toxicity/efficacy are most likely to drop out of the trial. We set $\sigma^2 = 3.5$ to induce a moderate correlation with Kendall's τ of 0.5 between the three time-to-event outcomes. We took the probability cutoffs $C_T = 0.6$, $a_T = 0.29$, $b_T = 0.16$, $a_E = 0.01$, and $b_E = 0.07$. We compared the proposed design with the Bayesian TTE design proposed by Yuan and Yin [19], which jointly models toxicity and efficacy as time-to-event outcomes. The TTE design addresses the issue of delayed outcomes, but does not account for informative dropouts/censoring and thus is subject to estimation bias. To make these two designs comparable and have the same target dose, in the TTE design, we adopted the same utility function and dose-finding algorithm as in the proposed design.

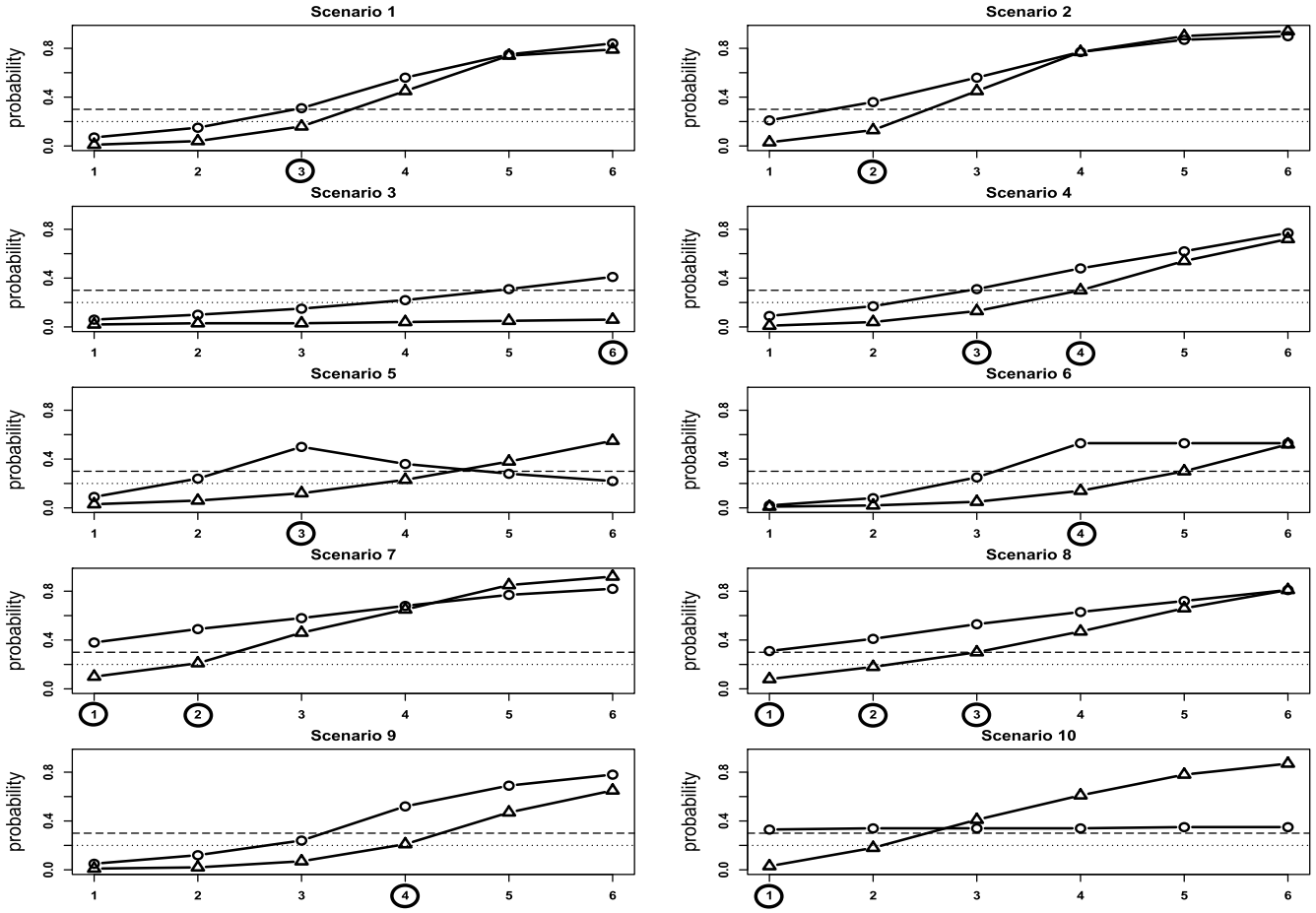

Figure 1. Dose-response curves for the ten scenarios in the simulation studies. The line with circles is the efficacy curve, the line with triangles is the toxicity curve. The efficacy lower bound and toxicity upper bound are indicated by the dotted and broken horizontal lines, respectively. The circled doses are target doses.

We simulated 10 scenarios with different numbers of target doses, locations of the target doses, and true marginal probabilities of toxicity and efficacy (see Figure 1). We designed the ten scenarios to have different levels of dropout rates. The dropout rate is defined as the number of patients with at least one missing outcome divided by the total number of patients. The dropout rate was about 40% for the first five scenarios, and about 30% for the last five scenarios. Under each scenario, we simulated 1,000 trials.

Table 1 summarizes the operating characteristics of our proposed design and the TTE design. Under each scenario, the first row is the true marginal probabilities of efficacy and toxicity at the end of follow-up (3 months for efficacy and 1.5 months for toxicity); the second row gives the utility (U) of each dose; the third and fourth rows show the selection probability and the average number of patients treated (shown in parentheses) at each dose under the two designs. In the first four scenarios, both efficacy and toxicity increase with dose. In scenario 1, the target dose is dose level 3 with the highest desirability $U = 0.15$. The proposed design had a target dose selection percentage that was about 10% higher than that of the TTE design, and also allocated more patients to the target dose. The TTE design tended to be aggressive and selected the overly toxic dose (i.e. dose level 4) 21.9% of the time. This is because the TTE design ignored the fact that dropouts were informative (i.e., patients who have high probabilities of experiencing toxicity were more likely to drop out), and as a result, it underestimated toxicity probabilities and led to aggressive dose escalation. In scenario 2, the target dose is dose level 2. The proposed design outperformed the TTE design in terms of target dose selection percentage and patient allocation. Again, the TTE design selected the higher dose level 3 with a larger probability than the proposed design because it assumed noninformative dropout. In scenario 3, the proposed design and TTE design led to similar selection percentages and patient allocations. Scenario 4 had 2 target doses (i.e., dose levels 3 and 4). The proposed and TTE designs yielded similar total target dose selection percentages, but the proposed design led to more balanced selections than the TTE de-

Table 1. Selection percentage and the number of patients (shown in parentheses) treated at each dose level under the proposed design and TTE design. The bolded numbers are target doses

	Dose level					
Design	1	2	3	4	5	6
			Scenario 1 (dropout rate = 40%)			
(π_E, π_T)	(0.07, 0.01)	(0.15, 0.04)	**(0.31, 0.16)**	(0.56, 0.45)	(0.75, 0.74)	(0.84, 0.79)
U	0.06	0.11	**0.15**	0.11	0.01	0.05
Proposed	3.8 (7.5)	11.6 (7.6)	**77.1 (29.9)**	4.8 (11.7)	0.0 (2.4)	0.0 (0.2)
TTE	2.7 (6.7)	7.1 (6.5)	**66.6 (25.1)**	21.9 (17.8)	0.3 (3.2)	0.0 (0.4)
			Scenario 2 (dropout rate = 40%)			
(π_E, π_T)	(0.21, 0.03)	**(0.36, 0.13)**	(0.56, 0.45)	(0.77, 0.77)	(0.87, 0.90)	(0.90, 0.94)
U	0.18	**0.23**	0.11	0.00	-0.03	-0.04
Proposed	21.5 (15.5)	**73.1 (30.3)**	4.2 (11.2)	0.0 (2.3)	0.0 (0.2)	0.0 (0.0)
TTE	19.2 (14.1)	**66.7 (28.2)**	13.9 (14.6)	0.0 (2.7)	0.0 (0.3)	0.0 (0.0)
			Scenario 3 (dropout rate = 40%)			
(π_E, π_T)	(0.06, 0.02)	(0.1, 0.03)	(0.15, 0.03)	(0.22, 0.04)	(0.31, 0.05)	**(0.41, 0.06)**
U	0.04	0.07	0.12	0.18	0.26	**0.35**
Proposed	1.1 (4.2)	0.2 (3.4)	1.7 (4.0)	4.0 (4.9)	6.4 (5.4)	**85.6 (37.7)**
TTE	0.8 (4.4)	0.3 (3.4)	1.9 (4.0)	3.5 (4.8)	4.5 (4.9)	**87.4 (38.0)**
			Scenario 4 (dropout rate = 40%)			
(π_E, π_T)	(0.09, 0.01)	(0.17, 0.04)	**(0.31, 0.13)**	**(0.48, 0.30)**	(0.62, 0.54)	(0.77, 0.72)
U	0.08	0.13	**0.18**	**0.18**	0.08	0.05
Proposed	6.5 (8.2)	5.9 (5.5)	**40.7 (16.1)**	**45.1 (22.8)**	0.4 (5.6)	0.1 (1.4)
TTE	6.7 (7.2)	4.6 (5.0)	**21.9 (11.8)**	**61.4 (25.7)**	4.3 (8.0)	0.1 (2.1)
			Scenario 5 (dropout rate = 40%)			
(π_E, π_T)	(0.09, 0.03)	(0.24, 0.06)	**(0.50, 0.12)**	(0.36, 0.23)	(0.28, 0.38)	(0.22, 0.55)
U	0.06	0.18	**0.38**	0.13	-0.10	-0.33
Proposed	4.7 (7.7)	15.5 (10.0)	**66.1 (25.1)**	11.7 (9.7)	1.3 (4.5)	0.1 (2.8)
TTE	4.1 (7.0)	11.9 (8.9)	**64.9 (24.0)**	17.0 (11.5)	1.3 (5.0)	0.4 (3.4)
			Scenario 6 (dropout rate = 30%)			
(π_E, π_T)	(0.02, 0.01)	(0.08, 0.02)	(0.25, 0.05)	**(0.53, 0.14)**	(0.53, 0.3)	(0.53, 0.52)
U	0.01	0.06	0.20	**0.39**	0.23	0.01
Proposed	0.5 (4.1)	0.5 (3.5)	11.5 (7.8)	**66.8 (24.1)**	19.3 (14.7)	1.1 (5.7)
TTE	0.6 (3.9)	0.4 (3.4)	7.3 (6.3)	**52.1 (20.9)**	36.4 (18.0)	3.0 (7.5)
			Scenario 7 (dropout rate = 30%)			
(π_E, π_T)	**(0.38, 0.1)**	**(0.49, 0.21)**	(0.58, 0.46)	(0.68, 0.65)	(0.77, 0.85)	(0.82, 0.92)
U	**0.28**	**0.28**	0.12	0.03	-0.08	-0.1
Proposed	**47.0 (24.7)**	**48.3 (22.6)**	2.1 (8.4)	0.2 (2.5)	0.0 (0.6)	0.0 (0.0)
TTE	**40.5 (22.4)**	**46.9 (21.1)**	11.0 (11.8)	0.2 (3.0)	0.1 (0.9)	0.0 (0.1)
			Scenario 8 (dropout rate = 30%)			
(π_E, π_T)	**(0.31, 0.08)**	**(0.41, 0.18)**	**(0.53, 0.30)**	(0.63, 0.47)	(0.72, 0.66)	(0.81, 0.81)
U	**0.23**	**0.23**	**0.23**	0.16	0.06	0.00
Proposed	**37.8 (20.6)**	**30.9 (14.3)**	**28.2 (16.5)**	1.2 (5.4)	0.0 (1.8)	0.0 (0.4)
TTE	**34.2 (19.6)**	**19.2 (11.6)**	**39.1 (17.7)**	5.9 (7.5)	0.2 (2.3)	0.0 (0.6)
			Scenario 9 (dropout rate = 30%)			
(π_E, π_T)	(0.05, 0.01)	(0.12, 0.02)	(0.24, 0.07)	**(0.52, 0.21)**	(0.69, 0.47)	(0.78, 0.65)
U	0.04	0.1	0.17	**0.31**	0.22	0.13
Proposed	1.3 (4.8)	2.0 (4.1)	15.2 (8.9)	**75.6 (29.5)**	4.6 (9.7)	0.0 (2.6)
TTE	0.8 (4.7)	0.8 (3.8)	8.8 (7.3)	**76.2 (28.0)**	11.9 (12.7)	0.3 (3.1)
			Scenario 10 (dropout rate = 30%)			
(π_E, π_T)	**(0.33, 0.03)**	(0.34, 0.18)	(0.34, 0.41)	(0.34, 0.61)	(0.35, 0.78)	(0.35, 0.87)
U	**0.30**	0.16	-0.07	-0.27	-0.43	-0.52
Proposed	**84.5 (40.4)**	13.3 (10.7)	1.4 (5.6)	0.0 (2.3)	0.0 (0.6)	0.0 (0.1)
TTE	**82.6 (39.3)**	15.4 (11.0)	1.4 (6.2)	0.2 (2.5)	0.0 (0.7)	0.0 (0.1)

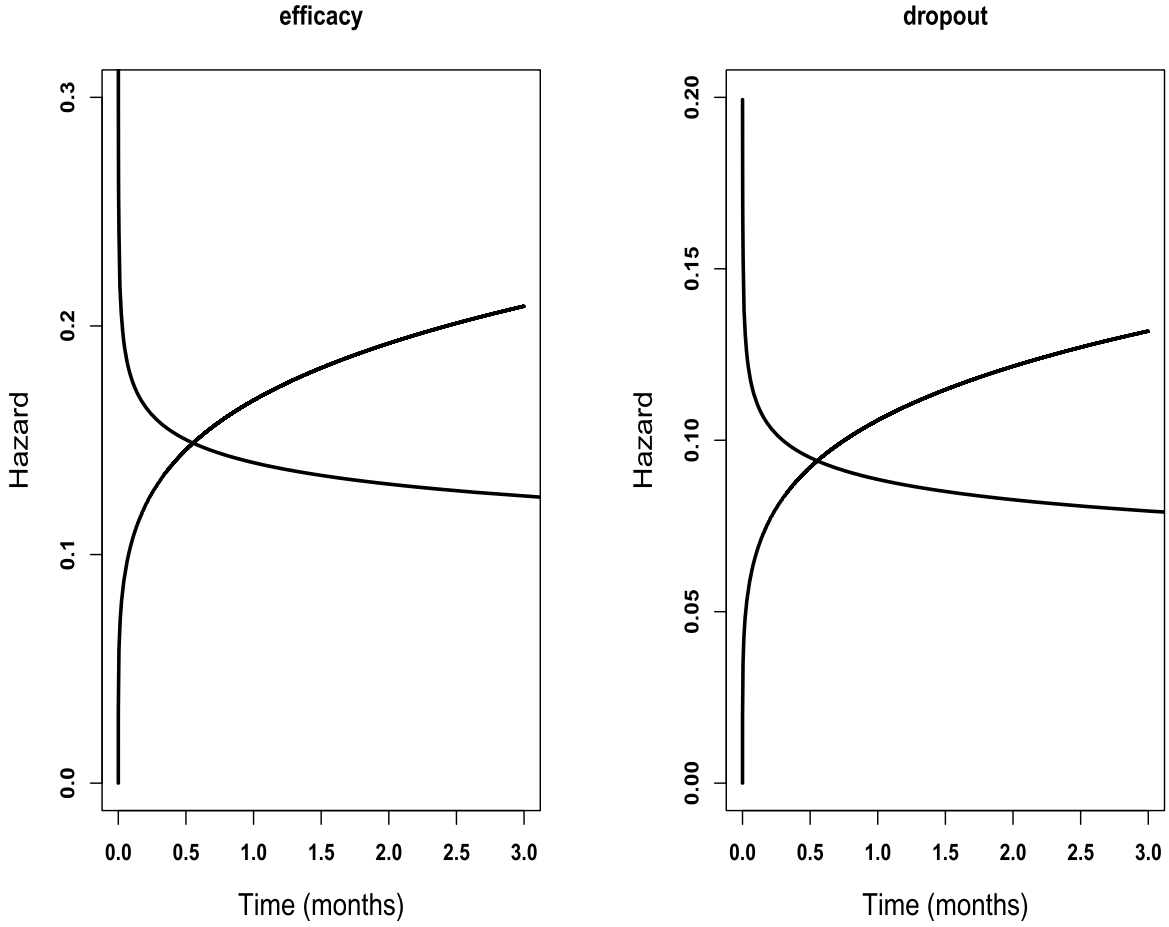

Figure 2. Two shapes (increasing and decreasing) of hazard functions for efficacy (left) and dropout (right) in the sensitivity analyses using a Weibull distribution.

sign (40.7% & 45.1% vs 21.9% & 61.4). Scenario 5 was designed for biological agents with efficacy probabilities first increase and then decrease. The proposed and TTE designs gave similar target dose selection percentages and patient allocations.

Scenarios 6–10 had dropout rates of about 30%. Scenario 6 considered a non-monotonic dose-efficacy relationship, where efficacy probabilities first increase and then plateau. The proposed design had a target dose selection percentage that was 14.7% higher than the TTE design, and also allocated more patients to the target dose. Scenarios 7 and 8 had more than 1 target dose. Scenario 7 had 2 target doses (i.e., dose levels 1 and 2) and scenario 8 had 3 target doses. In both scenarios, the proposed design yielded higher total target dose selection percentages than TTE design, and more balanced selections of the target doses than the TTE design. In scenario 9, the proposed design resulted in similar target dose selection percentage as the TTE design, but the TTE design was more aggressive and selected higher dose (i.e., dose level 5) with a larger probability. In scenario 10, the proposed and TTE designs yielded similar selection percentages and patient allocations.

3.1 Sensitivity analyses

We carried out sensitivity analyses to examine the robustness of our design under two settings. (1) We generated the times to efficacy, toxicity and dropout from the proportional hazards model with Weibull distribution baselines. We assumed that the hazard for the time to toxicity increased with the dose and considered two shapes (increasing and decreasing) of the hazard for times to efficacy and dropout, resulting in a total of 4 simulation settings (see Figure 2). (2) We simulated the times to efficacy, toxicity and dropout from the accelerated failure time model with a log-logistic error. We considered two shapes of the hazard for efficacy (increasing and decreasing) and three increasing shapes of the hazard for toxicity (linear, concave, and convex), resulting in a total of 6 combinations (see Figure 3). For the sensitivity analyses, we matched the marginal probabilities of efficacy, toxicity and dropout with the values in the original scenarios shown in Table 1.

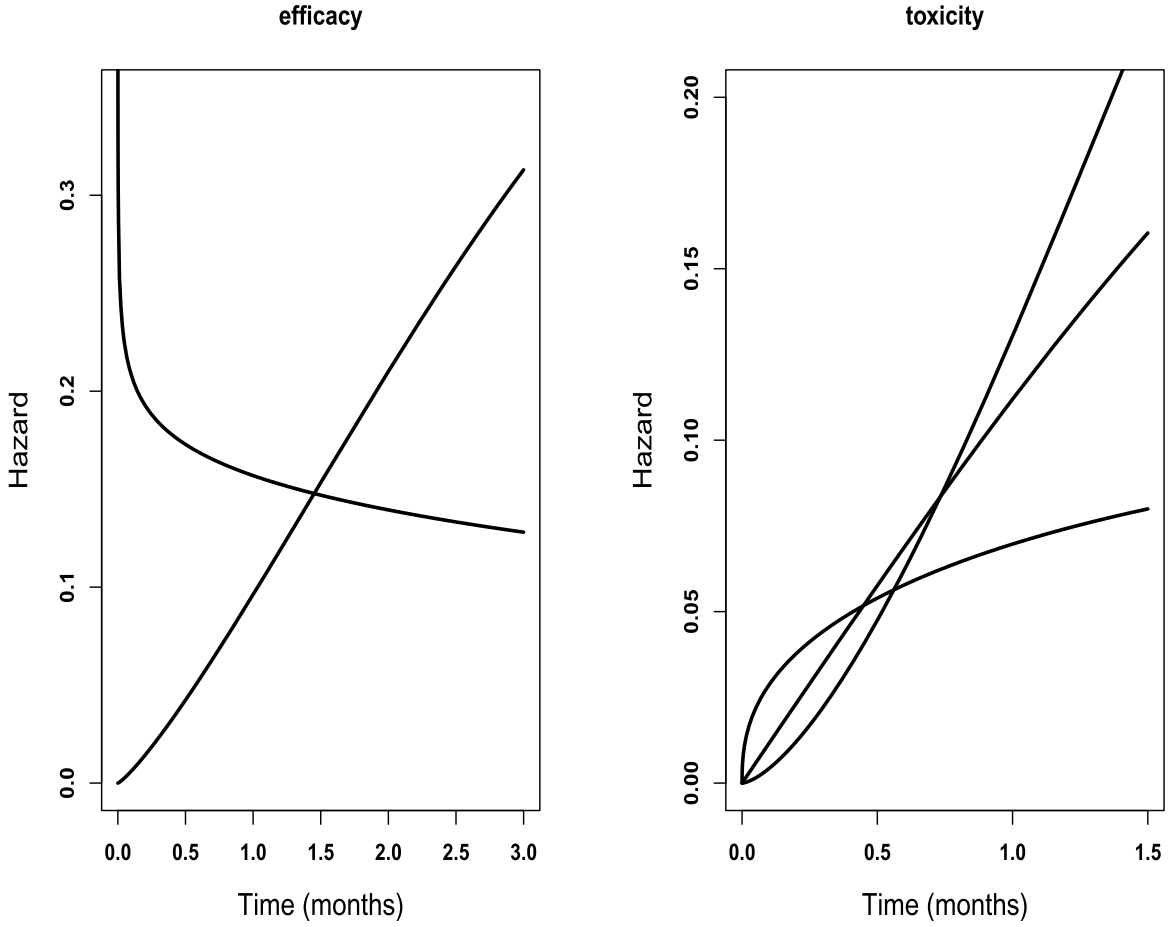

Figure 3. Two shapes (increasing and decreasing) of hazard functions for efficacy (left) and three shapes (linear, convex, concave) of hazard functions for toxicity (right) in the sensitivity analyses using the accelerated failure time model with a log-logistic error.

Table 2 provides the results when the times to the events were generated from the Weibull model or the accelerated failure time model under two representative scenarios, scenario 1 and scenario 9. We can see that, across all simulation settings, our design selected the target dose with the highest probability and assigned the largest number of patients to the target dose. The operating characteristics are comparable to those reported in Table 1, where the time-to-event data were generated from the exponential distribution.

4. CONCLUSIONS

We have proposed a Bayesian phase I/II trial design to handle informative dropouts. We jointly model efficacy, toxicity and patient dropout as time-to-event outcomes. This work was motivated by the practical problem of clinical trials in which patient dropout induces missing data, and as a result, the traditional trial designs, which assume binary efficacy and toxicity outcomes, are inappropriate. Our design incorporates correlations among the three types of time-to-event data by using a shared frailty, as the event of a patient dropping out of the trial often correlates with the individual's toxicity and efficacy outcomes. Our design can also accommodate a high patient accrual rate in the case of late-onset toxicity or efficacy and lead to a much shorter trial duration.

ACKNOWLEDGEMENTS

The authors would like to thank associated editor and reviewers for insightful and constructive comments that substantially improved the paper. Yuan's research was partially supported by Award Number R01 CA154591 and P50 CA098258 from the National Cancer Institute. Zang's research was partially supported by Award Number P50 CA098258 from the National Cancer Institute.

Received 31 January 2014

Table 2. Sensitivity analysis when baseline hazard follows Weibull distribution and when the times to efficacy, toxicity and dropout follow an accelerated failure time model with a log-logistic error. The bolded numbers are target doses

Weibull distribution							
Efficacy	Dropout			Scenario 1			
increasing	decreasing	2.5 (6.9)	6.3 (6.4)	**80.4 (28.5)**	9.1 (14.4)	0.0 (2.9)	0.0 (0.4)
increasing	increasing	3.5 (7.8)	10.4 (7.4)	**76.1 (27.6)**	7.9 (13.7)	0.1 (2.6)	0.0 (0.3)
decreasing	decreasing	2.6 (6.6)	9.0 (6.8)	**79.3 (29.0)**	7.5 (13.9)	0.1 (2.8)	0.0 (0.3)
decreasing	increasing	3.0 (6.5)	5.7 (5.8)	**79.2 (28.9)**	10.4 (14.9)	0.0 (3.1)	0.0 (0.4)
				Scenario 9			
increasing	decreasing	1.6 (4.7)	0.9 (3.9)	13.0 (8.2)	**79.0 (29.2)**	4.6 (11.1)	0.2 (2.7)
increasing	increasing	2.1 (5.1)	1.6 (4.1)	16.0 (9.4)	**74.7 (28.6)**	4.3 (10.0)	0.2 (2.5)
decreasing	decreasing	1.0 (4.8)	1.5 (4.0)	12.6 (8.3)	**77.5 (28.6)**	5.9 (10.9)	0.1 (3.0)
decreasing	increasing	1.4 (4.5)	1.4 (3.9)	10.6 (7.3)	**79.0 (29.2)**	6.4 (11.8)	0.1 (3.1)
Accelerated failure time model							
Efficacy	Toxicity			Scenario 1			
increasing	linear	4.3 (8.9)	11.6 (8.3)	**71.0 (26.0)**	8.8 (12.9)	0.0 (2.4)	0.0 (0.3)
increasing	concave	4.5 (10.3)	16.1 (10.0)	**67.4 (25.8)**	5.0 (10.0)	0.0 (2.0)	0.0 (0.1)
increasing	convex	4.6 (9.1)	9.5 (8.0)	**72.2 (25.6)**	9.7 (13.5)	0.0 (2.3)	0.0 (0.3)
decreasing	linear	2.5 (6.7)	8.5 (6.9)	**75.5 (27.7)**	10.5 (15.0)	0.0 (2.4)	0.0 (0.3)
decreasing	concave	3.7 (7.7)	11.0 (8.1)	**74.1 (28.6)**	7.6 (11.9)	0.0 (2.3)	0.0 (0.2)
decreasing	convex	2.5 (6.6)	8.0 (6.6)	**76.3 (27.7)**	11.1 (15.5)	0.0 (2.7)	0.0 (0.4)
				Scenario 9			
increasing	linear	2.5 (6.0)	1.7 (4.1)	15.3 (9.5)	**74.2 (27.5)**	4.3 (10.0)	0.0 (2.4)
increasing	concave	1.1 (5.9)	2.3 (4.4)	18.7 (10.7)	**72.0 (27.9)**	3.7 (8.6)	0.0 (2.0)
increasing	convex	1.2 (6.1)	1.4 (4.2)	14.2 (8.8)	**75.2 (26.6)**	5.2 (10.6)	0.0 (2.7)
decreasing	linear	0.7 (4.4)	1.6 (3.8)	9.8 (7.3)	**79.5 (28.7)**	7.4 (12.1)	0.0 (3.4)
decreasing	concave	1.9 (4.9)	1.2 (4.1)	13.7 (8.5)	**76.7 (29.3)**	5.0 (10.2)	0.0 (2.6)
decreasing	convex	0.9 (4.3)	0.8 (3.7)	8.8 (6.5)	**80.7 (27.6)**	7.7 (13.9)	0.0 (3.6)

REFERENCES

[1] Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, **45**, 925–937. MR1029610

[2] O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, **46**, 33–48. MR1059105

[3] Whitehead, J. and Brunier, H. (1995). Bayesian decision procedures for dose determining experiments. *Statistics in Medicine*, **14**, 885–893.

[4] Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997). A random walk rule for phase I clinical trials. *Biometrics*, **53**, 745–760.

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

[6] Stylianou, M. and Flournoy, N. (2002). Dose finding using the biased coin up-and-down design and isotonic regression. *Biometrics*, **58**, 171–7. MR1891376

[7] Yin, G. and Yuan. Y. (2009). Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*, **104**, 954–968. MR2750228

[8] Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and follow-up trials of a new chemotherapeutic agent *Journal of Chronic Diseases*, **13**, 346–353.

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

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

[11] Thall, P. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*, **50**, 337–349. MR1294683

[12] Chen T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, **16**, 2701–2711.

[13] Tan, S. B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, **21**, 1991–2012.

[14] Thall, P. and Russell, K. (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Statistics in Medicine*, **27**, 4895–4913.

[15] 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.

[16] Thall, P. and Cook, J. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, **60**, 684–693. MR2089444

[17] Bekele, B. and Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics*, **61**, 344–354. MR2140905

[18] Yin, G., Li, Y., and Ji, Y. (2006). Bayesian dose-finding in phase I/II trials using toxicity and efficacy odds ratio. *Biometrics*, **62**, 777–784. MR2247206

[19] Yuan, Y. and Yin, G. (2009). Bayesian dose finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society, Series C*, **58**, 719–736. MR2750264

[20] Yuan, Y. and Yin, G. (2011). Bayesian phase I/II drug-combination trial design in oncology. *Annals of Applied Statistics*, **5**, 156-174. MR2840181

[21] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistics Society, Series B*, **34**, 187–220. MR0341758

Beibei Guo
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803
USA
E-mail address: beibeiguo@lsu.edu

Yong Zang
Department of Mathematical Sciences
Florida Atlantic University
Boca Raton, FL 33431
USA
E-mail address: zangyong2008@gmail.com

Ying Yuan
Department of Biostatistics
The University of Texas MD Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: yyuan@mdanderson.org