Biometrics 64, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Generalized Phase I-II Designs to Increase Long Term Therapeutic Success Rate

Peter F. Thall

Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX

email: rex@mdanderson.org

and

Yong Zang

Department of Biostatistics and Health Data Science, Indiana University, IN

and

Ying Yuan

Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX

Summary: Early phase clinical trial designs assume implicitly that the dose of an experimental agent that is optimal based on short-term outcomes will maximize the agent's long-term therapeutic success rate. This assumption may not be true. A dose selected in an early phase trial may give suboptimal progression free survival or overall survival time, which reduces future patient benefit and the power of a randomized comparative trial. To address this problem, we propose a family of Bayesian generalized phase I-II designs that first use a conventional design based on short-term response and toxicity to identify a set of candidate doses, rather than one dose. Additional patients then are randomized among the candidates, with dose-specific sample sizes determined adaptively to obtain a given level of reliability. Each patient is followed for a fixed time period and a final dose is selected to maximize the estimated long term therapeutic success rate. A simulation study comparing the proposed design to conventional phase I-II designs is presented.

Key words: Bayesian Design; Cell Therapy; Dose Finding; Phase I-II Clinical Trial

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

A phase I-II clinical trial of an experimental agent in oncology applies outcome-adaptive rules based on early response and toxicity to monitor safety and effectiveness, choose doses for successive patient cohorts, and pick a final optimal dose (Braun, 2002; Bekele et al., 2005; Zhang et al., 2006; Mandrekar et al., 2010; Yuan et al., 2016; Zang and Lee., 2017). Most phase I-II designs use binary response, Y_R, and toxicity, Y_T, although the outcomes may be ordinal (Thall et al., 2012) or event time variables (Jin et al., 2014; Zhang et al., 2021; Zhang and Zang, 2021), and some designs use more than two outcomes (Thall et al., 2013; Lee et al., 2019; Lee at al., 2020). In practice, $\boldsymbol{Y} = (Y_R, Y_T)$, must be evaluated over a time period, $[0,\ t_1]$, short enough to allow outcome-adaptive decisions to be made feasibly.

Early phase designs rely on the implicit assumption that, if a dose d is optimal based on a criterion defined in terms of $\boldsymbol{Y}$, then d also is likely to maximize the therapeutic success rate over a longer follow up period, $[0,\ t_2]$, where t_2 is substantively larger than t_1. Denoting progression-free survival (PFS) time time starting from the follow up time t_1 when $\boldsymbol{Y}$ is evaluated by Z, long-term therapeutic success may be defined as the event $(Z > t_2 - t_1)$. Distinguishing between this event and early response, $(Y_R = 1)$, is important because a common problem in oncology is that an early response may not be durable, since a patient who has a response may relapse quickly. In general, given t_1 and t_2, response durability may be defined as the conditional probability $\Pr(Z > t_2 - t_1 \mid d, Y_R = 1, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the model parameter vector. Response durability has been discussed for radiation oncology (Tseng et al., 2015), donor lymphocyte infusion following relapse after allogeneic bone marrow transplantation (Dazzi et al., 2000), and many other areas of oncology. Phase I-II protocols often include provisions to follow each patient up to a longer time t_2 to obtain data on Z in order to estimate response durability. This is the case in our motivating trial, described below in section 4, where the goal is to optimize the dose of chimeric antigen

receptor (CAR) engineered cord blood derived natural killer (NK) cells for treating severe hematological malignancies, with $t_1 = 1$ month for dose finding and $t_2 = 6$ months to evaluate response durability.

In this paper, we focus on the common problem of how to choose an optimal dose of a new agent. We propose a design that uses the distributions of both short-term $\boldsymbol{Y}$ evaluated over $[0,\ t_1]$ and Z evaluated over $[0,\ t_2]$ to optimize dose. Let $\phi(d, \boldsymbol{\theta})$ be an objective function, defined in terms of the early outcome distribution $p(\boldsymbol{Y} \mid d, \boldsymbol{\theta})$, that is used as a criterion to optimize dose in a phase I-II trial. Denote the experimental agent by X, and let $X(d)$ denote X administered at dose d, since the effects of $X(d)$ and $X(d')$ on $\boldsymbol{Y}$ or Z may be different for $d \neq d'$. Denote a long-term success criterion by $\xi(d, \boldsymbol{\theta})$, which may be $S_Z(t_2 - t_1 \mid d, \boldsymbol{\theta})$ or the mean $E(Z \mid d, \boldsymbol{\theta})$. The implicit assumption underlying phase I-II trials is that, if a selected dose $d^{sel,\phi}$ maximizes an estimate of $\phi(d, \boldsymbol{\theta})$ at the end of phase I-II, then $d^{sel,\phi}$ also maximizes $\xi(d, \boldsymbol{\theta})$. The validity of this assumption depends on how the distributions $p(\boldsymbol{Y} \mid d)$ and $p(Z \mid d)$ vary with d, and associations between Z and $\boldsymbol{Y}$, which depend on how these outcomes are defined in a given clinical setting. Conventionally, using $\phi(d, \boldsymbol{\theta})$ in place of $\xi(d, \boldsymbol{\theta})$ to optimize d is motivated mainly by the desire for logistical convenience when conducting a sequentially adaptive dose-finding trial.

In many settings, $d^{sel,\phi}$ does not maximize $\xi(d, \boldsymbol{\theta})$. It can be shown easily by example that, for assumed true values $\phi^{true}(d_j, \boldsymbol{\theta})$ and $\xi^{true}(d_j, \boldsymbol{\theta})$ for doses $d_1, \cdots, d_J$, the true optimal doses $d^{opt,\phi}$ and $d^{opt,\xi}$ under the two criteria may differ. Even if Z depends on $\boldsymbol{Y}$ as well as d, an optimal phase I-II dose $d^{sel,\phi}$ based on $\phi(d, \boldsymbol{\theta})$ still may be suboptimal in terms of $\xi(d, \boldsymbol{\theta})$. For example, with targeted agents, immunotherapies, or cellular therapies, differences between how $p(\boldsymbol{Y} \mid d, \boldsymbol{\theta})$ and $p(Z \mid d, \boldsymbol{\theta})$ vary with d may be due to direct biological effects of $X(d)$ on Z that vary with d but are not mediated by Y_R or Y_T.

The fact that $\phi(d, \boldsymbol{\theta})$ and $\xi(d, \boldsymbol{\theta})$ are different criteria based on different outcomes may

have very undesirable practical consequences. If the dose $d^{opt,\xi}$ that is truly optimal in terms of $\xi(d, \boldsymbol{\theta})$ is different from $d^{sel,\phi}$ then, on average, $d^{opt,\xi}$ may give substantially larger Z than $\phi(d, \boldsymbol{\theta})$. In this case, a randomized trial of $X(d^{sel,\phi})$ versus a control, C, with outcome Z is studying the suboptimal version $X(d^{sel,\phi})$ of X. This reduces the probability that the trial will yield a positive result, compared to what would have been obtained if $X(d^{sel,\xi})$ had been used. It also reduces patient benefit in terms of Z. Another possibility is that a completed phase III trial may conclude that $X(d^{sel,\phi})$ is superior to C, when in fact $X(d^{sel,\phi})$ is inferior to $X(d')$ in terms of $\xi(d, \boldsymbol{\theta})$ for some dose $d' \neq d^{sel,\phi}$. Future patients treated with $X(d^{sel,\phi})$ based on the trial's results will have stochastically smaller Z than they would have had if d' had been chosen prior to phase III. Thus, choosing a dose in phase I-II that is optimal in terms of $\phi(d, \boldsymbol{\theta})$ but suboptimal in terms of $\xi(d, \boldsymbol{\theta})$ reduces the potential long-term effectiveness of X before a phase III trial is begun. Moreover, if a truly superior dose different from $d^{sel,\phi}$ is discarded erroneously prior to phase III, it is impossible to determine reliably whether this was done or, if so, how severely values of Z for future patients may have been reduced.

Relationships between short-term and long-term outcomes are a major issue in treatment evaluation. This has been discussed extensively, often with regard to use of an early outcome as a surrogate for a long-term outcome. Common examples are early response and PFS time, and the times to progression and death (Anderson et al., 1983; Simon and Makuch, 1984; Buyse and Piebois, 1996; Fleming and Powers, 2012; Thall, 2020). An example of the disconnect between early and long-term outcomes is a phase I trial in allogeneic stem cell transplantation for acute leukemia (unpublished). The goal was to optimize dose of a new histone deacetylase inhibitor (HDI) added to the standard preparative regimen. Six doses of the HDI were studied using the time-to-event continual reassessment method (TiTE-CRM) (Cheung and Chappell, 2000) with an expansion cohort. Toxicity was defined as graft failure or grade 4 or 5 non-hematologic, non-infectious toxicity, mucositis, or diarrhea within 30

days, using target toxicity probability 0.30. Note that, if response is defined as engraftment at day 30, this composite definition of toxicity includes non-response. The TiTE-CRM design selected the highest dose as the MTD, with final sample sizes (3, 3, 3, 4, 4, 51) at the six doses. Unfortunately, Kaplan-Meier estimates of survival for doses {1, 2, 3, 4, 5} combined versus dose 6, given in Figure 1, show that patients treated with dose 6 had worse survival than patients given one of the five lower doses. This trial illustrates that the disconnect between a selected phase I-II dose and the desire to increase Z also may arise if a phase I trial is used to choose a dose. A MTD from phase I may be suboptimal in terms of $\xi(d, \boldsymbol{\theta})$, simply because $p(Y_T \mid d)$ provides little information about $p(Z \mid d)$. Since it is well established phase I designs have severe flaws compared to phase I-II designs (Yuan et al., 2016; Yan et al., 2018; Gauthier et al., 2019), we will not consider phase I designs further here.

We address the problems described above by proposing a Bayesian generalized phase I-II design, called *Gen I-II*. First, a conventional phase I-II design based on $\boldsymbol{Y}$ is used to screen doses for safety and effectiveness, and identify a set of acceptable candidate doses, rather than selecting one dose. Additional patients are randomized among the candidates and followed to evaluate Z, and a best dose is selected based on long-term success rates. A Gen I-II design is practical if investigators plan to follow patients long enough to evaluate long-term treatment success, and are willing to treat additional patients to improve reliability.

Section 2 presents the Gen I-II design. Dose outcome models are given in section 3, followed by a description of the motivating trial in section 4. Section 5 provides details of a utility based Gen I-II design, and a simulation study is presented in section 6. We close with a brief discussion in section 7.

2. Generalized Phase I-II Design Elements

The Gen I-II design described here has three stages. In stage 1, a conventional phase I-II design is used, with $\boldsymbol{Y}$ evaluated over a short time period $[0,\ t_1]$. Any phase I-II design that

includes an objective function $\phi(d, \boldsymbol{\theta})$ characterizing dose desirability may be used for stage 1. In stage 2, doses are chosen using adaptive randomization (AR) with probabilities defined in terms of $\phi(d, \boldsymbol{\theta})$. At the end of stage 2, a set of acceptable *candidate doses* having estimated $\phi(d, \boldsymbol{\theta})$ close to the maximum estimate is determined. In stage 3, additional patients are randomized among the candidates, and patients are followed for a longer period $[0,\ t_2]$ to obtain data on $[Z \mid d]$ for all d in the candidate set. At the end of stage 3, the candidate dose maximizing the posterior mean of a long-term success criterion function $\xi(d, \boldsymbol{\theta})$ is selected.

To make things concrete for the Gen I-II design that we will consider here, we define the early outcomes, evaluated over the interval $[0,\ t_1]$, to be a binary indicator variable Y_T of toxicity and a three-level ordinal response variable Y_R' taking on the possible values 2 for response (RES), 1 for stable disease (SD), and 0 for progressive disease or death (PD). Denoting the indicator of the event A by $\mathcal{I}[A]$, we define the binary response indicator $Y_R = \mathcal{I}[Y_R' = 2]$. Including the third event SD $= (RES \cup PD)^c$ accommodates settings where PD and RES are not complementary events. The early outcome $\boldsymbol{Y}$ may be defined as either (Y_T, Y_R) or (Y_T, Y_R'), depending on which phase I-II design is used. More generally, other early outcomes may be used for $\boldsymbol{Y}$, including ordinal toxicity, or Y_R' with more than three levels, with appropriate modifications of design parameters.

To connect short-term and long-term outcomes, we define PFS time in a non-standard way, with Z starting at t_1, including discrete $Z = 0$ if $Y_R' = 0$, and continuous $Z > 0$ if $Y_R' \neq 0$. Thus, PD observed during $[0,\ t_1]$ by signs or symptoms or by imaging a solid tumor is an early treatment failure, and long term treatment success cannot occur if $Y_R' = 0$. This implies that $\Pr(Z > t_2 - t_1 \mid Y_R' = 0, d, \boldsymbol{\theta}) = 0$ for any $t_2 > t_1$, and hence

$$\Pr(Z > t_2 - t_1 \mid d, \boldsymbol{\theta}) = \Pr(Z > t_2 - t_1 \mid Y_R' > 0, d, \boldsymbol{\theta})\ \Pr(Y_R' > 0 \mid d, \boldsymbol{\theta}). \tag{1}$$

One may define $\xi(d, \boldsymbol{\theta})$ to equal expression (1) or, if desired, as the mean

$$E(Z \mid d, \boldsymbol{\theta}) = E(Z \mid Y_R' > 0, d, \boldsymbol{\theta})\ \Pr(Y_R' > 0 \mid d, \boldsymbol{\theta}). \tag{2}$$

Denote the Gen I-II stage s sample size by n_s for $s = 1, 2, 3$, and overall sample size $N = n_1 + n_2 + n_3$. Values of n_1 and n_2 are specified at the start of the trial, but n_3 is determined adaptively at the end of stage 2, as described below. Denote $\pi_k(d, \boldsymbol{\theta}) = \Pr(Y_k = 1 \mid d, \boldsymbol{\theta})$ for $k = R, T$. Examples of $\phi(d, \boldsymbol{\theta})$ based on bivariate binary $\boldsymbol{Y} = (Y_R, Y_T)$ include the response probability $\pi_R(d, \boldsymbol{\theta})$ (Thall and Russell, 1998), the odds ratio $[\pi_R(d, \boldsymbol{\theta})/\{1 - \pi_R(d, \boldsymbol{\theta})\}]/[\pi_T(d, \boldsymbol{\theta})/\{1 - \pi_T(d, \boldsymbol{\theta})\}]$ (Yuan et al., 2016), and the trade-off function $f\{\pi_R(d, \boldsymbol{\theta}), \pi_T(d, \boldsymbol{\theta})\}$ used by the EffTox design (Thall and Cook, 2004 ; Thall et al., 2014). If numerical utilities, $U(\boldsymbol{Y})$, of the early outcomes can be elicited, an objective function may be defined as the mean utility, $E\{U(\boldsymbol{Y}) \mid d, \boldsymbol{\theta}\}$. Below, we will use this approach with $\boldsymbol{Y} = (Y_R', Y_T)$. Utility-based designs are given by Thall et al., 2011, Thall et al., 2012, Lee et al., 2020, and many others.

For the i^{th} patient enrolled in a Gen I-II trial, denote the assigned dose by $d_{[i]}$ and let U_i be the independent right censoring time starting from the time t_1 when $Y_{i,R}'$ is evaluated, conditional on $Y_{i,R}' > 0$. The observed time to failure or censoring following t_1 is $Z_i^o = \min\{Z_i, U_i\}$. Let $\delta_i = 1$ if $Z_i^o = Z_i$ and $\delta_i = 0$ if $Z_i^o = U_i < Z_i l$, and denote the data from the first n patients enrolled in the trial by

$$\mathcal{D}_n = \{(Y_{i,R}', Y_{i,T}, Z_i^o, \delta_i, d_{[i]}) : \ i = 1, \cdots, n\}. \tag{3}$$

Stages 1 and 2 of the Gen I-II design include two Bayesian dose acceptability criteria,

$$\Pr\{\pi_R(d, \boldsymbol{\theta}) > \underline{\pi}_R \mid \mathcal{D}_n\} > .10 \ \text{ and } \ \Pr\{\pi_T(d, \boldsymbol{\theta}) < \overline{\pi}_T \mid \mathcal{D}_n\} > .10, \tag{4}$$

where $\underline{\pi}_R$ and $\overline{\pi}_T$ are fixed limits corresponding to the disease and clinical setting. Thus, an acceptable dose must not be unlikely to have response rate at least $\underline{\pi}_R$ or toxicity rate at most $\overline{\pi}_T$. Decision cut-offs other than .10 may be used in (4), if desired, based on preliminary computer simulations. Acceptability rules of this form have been used by Thall and Russell, 1998, Thall and Cook, 2004, Lin et al., 2021, and many others. Denote the set of acceptable doses satisfying (4) by $\mathcal{A}_n$. During stages 1 and 2, no patient is treated with an unacceptable

dose, and if it is determined that no dose is acceptable, i.e. $\mathcal{A}_n$ is empty, the trial is stopped, stage 3 is not conducted, and no dose is selected.

Denote $n_{1,2} = n_1 + n_2$. For each dose $d_j,\ j = 1, \cdots, J$ and $n = 1, \cdots, n_{1,2}$, denote the posterior mean based on the current data $\mathcal{D}_n$ by $\hat{\phi}_{j,n} = E\{\phi(d_j, \boldsymbol{\theta}) \mid \mathcal{D}_n\}$. In stage 1, doses are chosen to maximize $\hat{\phi}_{j,n}$ for n_1 patients. In stage 2, doses are chosen for n_2 patients using AR. Given fixed $0 < \zeta \leqslant 1$, AR probabilities may be defined as

$$r_{j,n} \ = \ \frac{(\hat{\phi}_{j,n})^{\zeta}}{\sum_{l:d_l \in \mathcal{A}_n} (\hat{\phi}_{l,n})^{\zeta}}, \quad \text{for } n = n_1 + 1, \cdots, n_{1,2} \text{ and } d_j \in \mathcal{A}_n.$$

Values of $\zeta < 1$ shrink $r_{j,n}$ toward the fair randomization probability $1/|\mathcal{A}_n|$. AR distributes patients more evenly among the doses during stage 2, giving a more even distribution of patients among the doses in $\mathcal{C}$. This may reduce the additional stage 3 per-dose sample sizes, described below. To facilitate trial conduct in stage 2, the AR probabilities may be updated and assigned for cohorts of $c > 1$, rather than individual patients, with n_2 a multiple of c. At the end of stage 2, for a given fixed $0 < \rho < 1$, the *candidate dose set* is defined to be all $d_j \in \mathcal{A}_{n_{1,2}}$ with posterior mean desirability close to the maximum value,

$$\mathcal{C} = \big\{ d_j \in \mathcal{A}_{n_{1,2}} : \ \hat{\phi}_{j,n_{1,2}} \ \geqslant \ \rho \max_{d_l \in \mathcal{A}_{n_{1,2}}} \hat{\phi}_{l,n_{1,2}} \big\}. \tag{5}$$

To use (5) in practice, rather than fixing a value of ρ arbitrarily the trial should be simulated using several numerical values, such as $\rho = .60, .70, .80$, to determine a ρ that gives a design with good operating characteristics (OCs) under a set of dose-outcome scenarios.

The stage 3 sample size n_3 is determined adaptively using the data $\mathcal{D}_{n_{1,2}}$ and the per-dose sample sizes $\{n_{1,2}(d_j) : \ d_j \in \mathcal{C}\}$ at the end of stage 2. For any $n_{1,2}$ and phase I-II design, the $n_{1,2}(d_j)$'s are random because doses are chosen adaptively in stages 1 and 2. Denote the stage 3 sample size of dose $d_j \in \mathcal{C}$ by $n_3(d_j)$. Thus, $n_3 = \sum_{d_j \in \mathcal{C}} n_3(d_j)$ and the per-dose sample sizes from all three stages are $N(d_j) = n_1(d_j) + n_2(d_j) + n_3(d_j)$. To determine $n_3(d_j)$ adaptively, we choose a fixed overall per dose sample size $N(d) = N(d_j)$ for all j that ensures a desired level of reliability for selecting an optimal dose from $\mathcal{C}$ at the end of the trial. Since $\mathcal{C}$ is a

random set, the value of $N(d)$ may be chosen from several feasible values. such as $N(d) =$ 10, 15, or 20. This may be based on simulations of the trial, for given n_1, n_2, ρ, and assumed true values of the long term success probabilities, $\boldsymbol{\xi}^{true} = (\xi^{true}(d_1), \cdots, \xi^{true}(d_J))$, and short term success probabilities, $\boldsymbol{\phi}^{true} = (\phi^{true}(d_1), \cdots, \phi^{true}(d_J))$. Each $n_3(d_j) = N(d_j) - n_{1,2}(d_j)$ depends on $\mathcal{C}$ and the values of $n_{1,2}(d_j)$ for the candidate doses $d_j \in \mathcal{C}$. For example, if $J = 4$, $\mathcal{C} = \{d_3, d_4\}$, $n_{1,2}(d_3) = 12$, and $n_{1,2}(d_4) = 6$, then $N(d) = 20$ requires $n_3(d_3) = 8$ and $n_3(d_4)$ $= 14$. In stage 3, a total of 22 additional patients would be randomized between d_3 and d_4, restricted to obtain overall per-dose sample sizes of 20.

For the final dose selection, we require that each $d \in \mathcal{C}$ must satisfy the additional long-term success probability acceptability requirement

$$\Pr\{\xi(d, \boldsymbol{\theta}) > \underline{\xi} \mid \mathcal{D}_N\} > .10, \tag{6}$$

where fixed $\underline{\xi}$ is the lowest acceptable value of $\xi(d_j, \boldsymbol{\theta})$. Denote the final set of acceptable doses in $\mathcal{C}$ by $\mathcal{A}_N^{\xi}$. The futility requirement (6) avoids selecting the best dose from a set of candidate doses that all are unlikely to have a long term therapeutic success rate that is at least $\underline{\xi}$. For example, $\underline{\xi}$ may be the historical mean of ξ with standard therapy. The final selected optimal dose in $\mathcal{A}_N^{\xi}$ has largest posterior mean long term success probability,

$$d_N^{sel,\xi} = \underset{d_j \in \mathcal{A}_N^{\xi}}{\operatorname{argmax}} E\{\xi(d_j, \boldsymbol{\theta}) \mid \mathcal{D}_N)\}. \tag{7}$$

If desired, one may use the alternative definition

$$d_N^{sel,\xi} = \underset{d_j \in \mathcal{A}_N^{\xi}}{\operatorname{argmax}} \Pr\big[\xi(d_j, \boldsymbol{\theta}) = \max_{d_r \in \mathcal{A}_N^{\xi}} \{\xi(d_r, \boldsymbol{\theta})\} \mid \mathcal{D}_N)\big]. \tag{8}$$

Figure 2 provides a schematic for Gen I-II design conduct. The design parameters include all values required to specify the stage 1 and 2 phase I-II design and the objective function $\phi(d, \boldsymbol{\theta})$, including t_1, n_1, n_2, cohort size c, acceptability limits $\underline{\pi}_E$ and $\overline{\pi}_T$, and the exponent ζ used to define the AR probabilities. For stage 3, one must specify the long-term follow up time t_2, ρ, $\underline{\xi}$, and the overall per-dose sample size $N(d)$ required for each $d_j \in \mathcal{C}$. For the Bayesian model, one must specify hyperparameters $\tilde{\boldsymbol{\theta}}_1$ of the noninformative prior $p(\boldsymbol{\theta}_1 \mid \tilde{\boldsymbol{\theta}}_1)$

in the model for $p(\boldsymbol{Y} \mid d, \boldsymbol{\theta}_1)$, and hyperparameters $\tilde{\boldsymbol{\theta}}_2$ of the noninformative prior $p(\boldsymbol{\theta}_2 \mid \tilde{\boldsymbol{\theta}}_2)$ in the conditional failure time distribution $p(Z \mid d, \boldsymbol{Y}, \boldsymbol{\theta}_2)$.

3. Dose Outcome Models

Here, we will assume the following simple model for the early outcome $\boldsymbol{Y} = (Y_R', Y_T)$. For each $j = 1, \cdots, J$, denote $p_{a,b}(d_j) = \Pr(Y_R' = a, Y_T = b \mid d_j, \boldsymbol{\theta})$ for $a = 0, 1, 2$ and $b = 0, 1$ and $\boldsymbol{p}(d_j) = (p_{0,1}(d_j), \cdots, p_{1,2}(d_j))$. Thus, $\sum_{a=0}^{2} \sum_{b=0}^{1} p_{a,b}(d_j) = 1$ and $\boldsymbol{\theta}_1 = (\boldsymbol{p}(d_1), \cdots, \boldsymbol{p}(d_J))$. For each dose d_j and interim sample size n, we assume that the six-dimensional count vector

$$\boldsymbol{X}_n(d_j) = \sum_{i=1}^{n} \big(I[(\boldsymbol{Y} = (0,0)], \cdots, I[\boldsymbol{Y} = (2,1)] \big)\ I[d_{[i]} = d_j]$$

is multinomial with parameters $n(d_j)$ and $\boldsymbol{p}(d_j)$, and that $\boldsymbol{p}(d_j)$ follows a non-informative Dirichlet prior with parameter 1/6 in each cell, which has effective sample size 1. While this model does not borrow strength between doses, it facilitates posterior computation because, by conjugacy, $\boldsymbol{p}(d_j) \mid \boldsymbol{X}_n(d_j)$ is Dirichlet with parameters (1/6, $\cdots$,1/6) + $\boldsymbol{X}_n(d_j)$ for each d_j. The early outcome objective function is defined as the mean utility

$$\phi(d_j, \boldsymbol{\theta}_1) = \overline{U}(d_j, \boldsymbol{\theta}_1) = \sum_{a=0}^{2} \sum_{b=0}^{1} U(a,b) p_{a,b}(d_j) \quad \text{for} \quad j = 1, \cdots, J. \tag{9}$$

To illustrate the Gen I-II design, we will focus on this model and utility structure for $\boldsymbol{Y}$. This Multinomial-Dirichlet model and formulation of $\phi(d_j, \boldsymbol{\theta}_1)$ may be extended quite easily, and may accommodate any discrete $\boldsymbol{Y}$, including ordinal Y_T or Y_R' indexing outcomes such as {PD, SD, partial response, complete response}.

For the conditional distribution of Z, due to the limited sample size a flexible but parsimonious parametric model is needed. Accordingly, we assume that $[Z \mid Y_R' > 0]$ follows a Weibull distribution with probability density function

$$f_Z(z \mid Y_R' > 0, Y_T, d_j, \boldsymbol{\theta}_2) = \frac{\alpha}{\lambda} \left(\frac{z}{\lambda} \right)^{\alpha - 1} \exp\{-(z/\lambda)^{\alpha}\}, \qquad z > 0,$$

where $\alpha > 0$ is the shape parameter and the rate parameter λ is the following function of

the early outcomes and dose,

$$\log\{\lambda(Y_R', Y_T, d_j, \boldsymbol{\theta})\} = \beta_0 + \beta_R\, \mathcal{I}[Y_R' = 2] + \beta_T\, Y_T + \gamma_j\, \mathcal{I}[j > 1]. \tag{10}$$

Thus, $\boldsymbol{\theta}_2 = (\alpha, \beta_0, \beta_R, \beta_T, \gamma_1, \cdots, \gamma_J)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Non-informative $N(0, 10^2)$ priors are assumed for elements of $\boldsymbol{\theta}_2$, and a $Gamma(0.01, 0.01)$ prior is assumed for α. The joint likelihood from n patients with data $\mathcal{D}_n$ is the product

$$\mathcal{L}(\mathcal{D}_n, \boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} \pi_{\boldsymbol{Y}}(\boldsymbol{Y}_i \mid d_{[i]}, \boldsymbol{\theta}) \Big[\big\{ f_Z(Z_i^o \mid \boldsymbol{Y}_i, d_{[i]}, \boldsymbol{\theta} \big\}^{\delta_i} \big\{ S_Z(Z_i^o \mid \boldsymbol{Y}_i, d_{[i]}, \boldsymbol{\theta}) \big\}^{(1-\delta_i)} \Big]^{I[Y_{i,R}' > 0]},$$

and the posterior is $p(\boldsymbol{\theta} \mid \mathcal{D}_n) \;\propto\; \mathcal{L}(\mathcal{D}_n, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$.

4. Motivating Trial

The Gen I-II design was motivated by a trial of CD70 CAR-NK cells as targeted immunotherapy for patients with recurrent or treatment-resistant B-cell hematologic malignancies. Treatment consists of lymphodepleting chemotherapy for three days followed by CAR-NK cell infusion. This therapy is motivated by reverse signalling of CD70 that enhances the anti-disease activity of the CAR-NK cells against B-cell malignancies (Alsayed et al., 2017 ; Liu et al., 2020). The goal is to determine an optimal dose among the four values $5.0 \times (10^6, 10^7, 10^8, 10^9)$ cells, with standardized values $(d_1, d_2, d_3, d_4) = (1,\ 2,\ 3,\ 4)$. The outcomes are the indicators Y_R of response, defined as complete remission at day $t_1 = 30$, and Y_T of grade 3 or 4 non-hematologic toxicity or cytokine release syndrome within 30 days. The early outcome dose acceptability limits in (4) are $\overline{\pi}_T = .30$ and $\underline{\pi}_R = .50$. The EffTox phase I-II design (Thall et al., 2014) is used to conduct stages 1 and 2 of the Gen I-II trial, with a maximum of 48 patients treated in cohorts of size 3 each, $n_1 = 15$ treated by choosing doses to optimize the estimated efficacy-toxicity trade-off function $\phi(d, \boldsymbol{\theta})$ and $n_2 = 33$ patients' doses chosen by AR. The first cohort is treated at d_1, no untried dose may be skipped when escalating, and no unacceptable doses may be used to treat a patient. The

candidate dose set $\mathcal{C}$ is defined as all acceptable d at the end of stage 2 satisfying

$$E\{\phi(d,\boldsymbol{\theta}) \mid \mathcal{D}_{48}\} \geqslant .70 \times E\{\phi(d^{sel},\boldsymbol{\theta}) \mid \mathcal{D}_{48}\}. \tag{11}$$

For long term therapeutic success, Z = PFS time is defined starting at one month if $Y_R = 1$, monitored for up to five more months, so $t_2 = 6$, and the long term success probability is $\xi(d,\boldsymbol{\theta}) = \Pr(Z > 5 \mid d, Y_R = 1, \boldsymbol{\theta}_2)\pi_R(d,\boldsymbol{\theta}_1)$. The fixed lower limit on the long term success probability acceptability criterion is $\underline{\xi} = .40$.

5. A Utility Based Gen I-II Design

Because the patients in the CD70 CAR-NK cell trial have active disease at enrollment, to define Y_R' for the 1-month evaluation, PD is defined as worsening of disease, and RES as complete remission, with $SD = (PD \cup RES)^c$. To obtain a utility, as usually done we first fixed the two boundary values $U(1,0) = 0$ for the worst possible outcome and $U(0,2) = 100$ for the best possible outcome, and then determined the remaining four intermediate values, subject to the admissibility constraints $U(a,0) \leqslant U(a,1) \leqslant U(a,2)$ for $a = 0, 1$, and $U(1,b) \leqslant U(0,b)$ for $b = 0$, 1, or 2. Table 1 gives the utility that we will use for the simulations. The short term outcome objective function is defined as the mean utility (9). During stages 1 and 2, given interim data $\mathcal{D}_n$, the posterior mean utility is $u(d_j, \mathcal{D}_n) = E\{\overline{U}(d_j,\boldsymbol{\theta}) \mid \mathcal{D}_n\}$.

In stages 1 and 2, the admissible dose set $\mathcal{A}_n$ is updated after each cohort's outcomes $\boldsymbol{Y} = (Y_R', Y_T)$ have been evaluated at follow up time $t_1 = 1$. For a Gen I-II trial with $c = 3$ and $n_1 = 15$, stage 1 is conducted as follows.

1. Treat the first cohort of patients at the lowest dose d_1.

2. For each new cohort, update the posterior distribution and compute the admissible dose set $\mathcal{A}_n$ and $u(d_j, \mathcal{D}_n)$ for each $j = 1, 2, 3, 4$.

3. If $\mathcal{A}_n$ is empty, stop the trial and select no dose.

4. If $\mathcal{A}_n$ is not empty, treat the next cohort of patients at the dose in $\mathcal{A}_n$ maximizing $u(d_j, \mathcal{D}_n)$, subject to the constraint that an untried dose may not be skipped when escalating.

5. If $\mathcal{A}_n$ is not empty and the current dose d_j is the highest dose that has been tried so far and it satisfies the safety rule $\Pr\{\pi_T(d_j, \boldsymbol{\theta}) < \overline{\pi}_T \mid \mathcal{D}_n\} > .10$, escalate one dose level. This rule supersedes rule 4.

6. Repeat steps 1-5 until five cohorts of patients have been treated and their values of $\boldsymbol{Y}$ evaluated.

Rule 5 is included because, due to its simplicity, the assumed model cannot estimate $\boldsymbol{p}(d_j)$ for untried d_j. This rule reduces the chance getting stuck at a locally optimal dose since it facilitates exploring untried doses. In stage 2, for $n_2 = 33$, another 11 cohorts of patients are randomized sequentially among the doses in $\mathcal{A}_{n_{1,2}}$ using the AR probabilities $r_{j,n}$ with shrinkage parameter $\zeta = 0.5$. The admissible dose set $\mathcal{A}_n$ is updated after each cohort's outcomes $\boldsymbol{Y}$ have been evaluated at $t_1 = 1$ month of follow up. In stage 3, additional patients are randomized among the doses in the candidate set $\mathcal{C}$ with $\rho = 0.7$. The per-dose stage 3 sample sizes are chosen adaptively to ensure that a three-stage total of $N(d_j) = 15$ patients are treated at each candidate dose. Thus, the stage 3 sample sizes $n_3(d_j) = N(d_j) - n_{1,2}(d_j)$ are random. Long term treatment success is the event $[Z > 5]$ that a patient is alive without disease in remission at $t_2 = 6$ months, and $\xi(d_j, \boldsymbol{\theta}) = S_Z(5 \mid d_j, \boldsymbol{\theta})$. Denoting the final overall sample size by N, a dose $d^{opt,\xi}$ is chosen at the end of stage 3 to maximize the posterior mean $\mathrm{E}\{S_Z(5 \mid d_j, \boldsymbol{\theta}) \mid \mathcal{D}_N\}$. We use JAGS to run Markov chain Monte Carlo (MCMC) simulation to generate posterior samples of $\boldsymbol{\theta}$ and estimate all posterior values. R code for implementing this design is available from https://github.com/yongzang2020.

6. Simulation study

We conducted a simulation study to investigate the OCs of the utility based Gen I-II design, using the CD70 CAR-NK cell trial as the basis for constructing simulation settings. We specified the scenarios with various patterns of $\phi(d_j)^{true}$, $\xi(d_j)^{true}$, and outcome distributions. Figure 3 shows the assumed true dose-response curves $\pi_T^{true}(d_j)$, $\pi_R^{true}(d_j) = \Pr^{true}(RES \mid d_j)$, and $\xi^{true}(d_j) = \Pr^{true}(Z > 5 \mid Y_R' > 0)$. As comparators, we used two conventional phase I-II designs, called Conv 1 and Conv 2, based on $\boldsymbol{Y}$. The Conv 1 design consists of stages 1 and stage 1 of the Gen I-II design, with an optimal dose selected to maximize the posterior mean utility $u(d_j, \mathcal{D}_n)$. The Conv 2 design is almost identical to the Conv 1 design, except that more patients are randomized in stage II to match the sample size of the Gen I-II design. We simulated 5,000 trials under each scenario using each of the three designs.

We employed a latent bivariate normal distribution to generate (Y_R', Y_T). Specifically, we assumed $\mathbf{W} = (W_{R'}, W_T)$ following a bivariate normal distribution with mean (0,0), variances 1, and correlations .20, and defined (Y_R', Y_T) as follows:

$$Y_T = \begin{cases} 0 & \text{if } W_T < \kappa_T(d_j) \\ 1 & \text{if } W_T \geqslant \kappa_T(d_j) \end{cases}, \quad Y_{R'} = \begin{cases} 0 & \text{if } W_{R'} < \kappa_{R'1}(d_j) \\ 1 & \text{if } \kappa_{R'1}(d_j) \leqslant W_{R'} < \kappa_{R'2}(d_j) \\ 2 & \text{if } W_{R'} > \kappa_{R'2}(d_j) \end{cases}.$$

To generate PFS time, we assumed $[Z \mid Y_{R'} > 0, Y_T, d_j]$ followed a piecewise exponential distribution with probability density function

$$f_Z(Z|Y_{R'}, Y_T, d_j) = \exp\{-Z/\tilde{\lambda}(Z)\}, \quad 0 < Z \leqslant 5,$$

with piecewise log hazard

$$\log\tilde{\lambda}(Z) = \begin{cases} \tilde{\beta}_{01} + \tilde{\beta}_{R1} I[Y_{R'} = 2] + \tilde{\beta}_{T1} Y_T + \tilde{\gamma}_{j1} I[j > 1] & \text{if } \ 0 < Z \leqslant 2.5 \\ \tilde{\beta}_{02} + \tilde{\beta}_{R2} I[Y_{R'} = 2] + \tilde{\beta}_{T2} Y_T + \tilde{\gamma}_{j2} I[j > 1] & \text{if } \ 2.5 < Z \leqslant 5 \end{cases}$$

The parameters $\kappa_T(d_j)$, $\kappa_{R'1}(d_j)$, $\kappa_{R'2}(d_j)$, $\tilde{\beta}_{0l}$, $\tilde{\beta}_{Rl}$, $\tilde{\beta}_{Tl}$ and $\tilde{\gamma}_{j1}, \tilde{\gamma}_{j2}$ were derived under each scenario to match the pre-determined $\phi(d_j)^{true}$ and $\xi(d_j)^{true}$ values for $j = 1, \cdots, 4$.

Table 2 summaries the OCs of the Gen I-II, Conv 1, and Conv 2 designs, including dose

selection percentages, average number of patients at each dose, and average overall sample size. The percentage under dose 0 is the percentage of trials terminated early with no dose is selected. A summary statistic to evaluate performance by comparing the selected optimal dose to the truly optimal dose is defined as $R(d^{sel,\xi}) = \xi(d^{sel,\xi})/\xi(d^{opt,\xi})$. R has domain [0, 1], and $R(d^{opt,\xi}) = 1$ corresponds to selecting the best possible dose in terms of long term treatment success. Using $R(d^{opt,\xi})$ rather than only the empirical probability of selecting $d^{opt,\xi}$ to quantify how well a method behaves is useful in scenarios where two or more doses have $\xi(d)^{true}$ close to $\xi(d^{opt,\xi})$, so choosing a nearly optimal dose is a good decision.

Scenarios 1 and 2 are "null" cases where no dose has both acceptable $\pi_T^{true}(d_j)$ and acceptable $\xi(d)^{true}$. In scenario 1, the Gen I-II design terminates the trial early 93.5% of the time compared to 56.2% and 57.7% for the Conv 1 and Conv 2 designs. In scenario 2, the Gen I-II design terminates the trial early 83.8% of the time. In contrast, because d_1 and d_2 both have substantially high $\pi_R^{true}(d_j)$ and the conventional designs ignore Z, they both have only about a 3% chance of stopping early, and a 73% chance of selecting d_1 as optimal. Scenario 2 illustrates the advantage that the Gen I-II design includes an admissibility requirement in terms of $\xi(d)^{true}$, while the conventional designs do not, and consequently they have a high risk of selecting a dose with a low long term success rate. In scenarios 1 and 2, because no dose is truly optimal, R is undefined.

Scenarios 3 and 4 are cases where there are multiple nearly optimal doses in terms of the mean utilities $\overline{U}^{true}(d_j)$ based on $\boldsymbol{Y}$, but only one dose is optimal in terms of the long-term criterion $\xi^{true}(d_j)$. In scenario 3, d_3 and d_4 have similar mean utilities around 75, while d_4 is truly optimal with the highest $\xi^{true}(d_4) = 0.7$, compared to $\xi^{true}(d_3) = 0.5$. In scenario 4, d_2, d_3 and d_4 all have similar mean utilities $\overline{U}^{true}(d_j)$ around 81, but d_3 is truly optimal with $\xi^{true}(d_3) = 0.65$, compared to $\xi^{true}(d_2) = .45$ and $\xi^{true}(d_4) = .50$. The Gen I-II design has a 69.1% probability of correctly selecting d_3 in scenario 3, whereas the Conv 1 and Conv 2

designs have 31% and 32% chances of selecting d_3, and are about as likely to select dose d_2 or d_4, again because both conventional designs ignore Z.

Scenarios 5 and 6 are cases where the truly optimal dose in terms of $\xi^{true}(d_j)$ and the dose with highest mean utility $\overline{U}^{true}(d_j)$ differ. In scenario 5, d_4 is truly optimal with the highest $\xi^{true}(d_4) = 0.65$, whereas d_3 has the highest mean utility $\overline{U}^{true}(d_3) = 82.3$. Similarly, in scenario 6, d_2 is truly optimal with the highest $\xi^{true}(d_2) = 0.7$, whereas d_4 has highest mean utility $\overline{U}^{true}(d_4) = 77.8$. The Conv 1 and Conv 2 designs both have over a 60% chance of incorrectly selecting d_3 as optimal in scenario 5, and around a 50% chance of incorrectly selecting d_4 as optimal in scenario 6, and both have below 15% and around 25% chances of correct optimal dose selection in scenarios 5 and 6. In contrast, the Gen I-II design has correct dose selection rates 59.1% in scenario 5 and 68.4% in scenario 6. In scenarios 7 and 8, the truly optimal dose and the dose with highest mean utility value are identical. The Gen I-II still outperforms the Conv 1 and Conv 2 designs, with a 25% higher correct optimal dose selection percentage in scenario 7. In scenario 8, the Gen-II design and the Conv 2 have similar correct optimal dose selection probabilities of 56.9% and 54.0%.

In summary, the Gen I-II design outperforms the conventional phase I-II designs quite substantially, especially in terms of optimal dose selection probabilities. The Gen I-II design also always yields the highest R values across all the scenarios, and the difference is at least 10% in all scenarios except scenario 8. The three designs have similar patient allocation distributions, essentially because for all designs allocate patients based on short-term outcomes, while Z is only used in the final optimal dose selection of the Gen I-II design.

We conducted sensitivity analyses to explore several aspects of the Gen I-II design. Table 3 summarizes the effects of including AR in stage 2. The design "with AR" is the Gen I-II design that allocates $n_1 = 15$ patients for the stage 1 and $n_2 = 33$ patients for stage 2 with AR; "without AR" is a modified Gen I-II design that combines stages 1 and 2, does not

include AR, and allocates 48 patients with dose-finding done to maximize $u(d_j, , \mathcal{D}_n)$ for all cohorts. Table 3 shows that, compared to the original "with AR" version of Gen I-II, the "without AR" version results in substantially inflated sample sizes, with only a mild gain of $\leqslant 5\%$ in true optimal dose selection percentage. This implies that AR is an important component of the Gen I-II design.

In Table 2 we fixed ρ used in candidate doses selection to be $\rho = 0.7$. Table 4 summarizes the OCs of the Gen I-II design for different values of ρ, ranging from 0.6 to 0.9. The results indicate that larger ρ is more favorable under the null scenarios 1 and 2, while smaller ρ is more favorable when a truly optimal dose exists, in scenarios 3 to 8. For an overall comparison across all the scenarios, we recommend $\rho = 0.7$ to obtain a generally good performance.

In Table 5 shows the effects of changing the patient allocation between stages 2 and 3. Given the values $n_2 = 33; N(d) = 15$ used in Table 2, we consider the two alternates $n_2 = 39; N(d) = 9$ and $n_2 = 27; N(d) = 21$. The results show that the default allocation and the allocation of $n_2 = 39, N(d) = 9$ have very similar performances, and both give better OCs compared to the last allocation $n_2 = 27; N(d) = 21$, in terms of the trade between correct true optimal dose selection percentage and sample size.

7. Discussion

While the basic idea underlying the Gen I-II design is very simple, it addresses a severe problem with conventional early phase dose finding methods. The Gen I-II design is modular in that $\boldsymbol{Y}$ can be any early outcome used for dose finding, and any phase I-II design and early criterion $\phi(d, \boldsymbol{\theta})$ can be used for stages 1 and 2. Thus, it can be tailored to accommodate a particular clinical setting at hand. While we have investigated a utility-based Gen I-II design based on a very simple Bayesian model, its large advantages over conventional competitors shown by our simulations indicate it is likely that other Gen I-II designs will provide a

similarly large benefit. This is because optimizing dose using a criterion based on a long-term outcome explicitly addresses the problem that conventional phase I-II designs ignore.

The Gen I-II design is practical in settings where investigators plan to follow phase I-II patients long enough to assess response durability. This is done quite commonly in oncology due to the fact that disease progression following early response remains a central problem. Consequently, once Gen -II designs and software are developed to accommodate a reasonable set of cases arising in practice, the methodology should provide a substantive advance over conventional phase I-II designs.

References

Al Sayed MF, Ruckstuhl CA, Hilmenyuk T, et al. (2017). CD70 reverse signaling enhances NK cell function and immunosurveillance in CD27-expressing B-cell malignancies. *Blood* **130**, 297-309.

Anderson JR, Cain KC and Gelber RD (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology* **1**, 710-719.

Bekele Nebiyou and Shen Yu (2005). A Bayesian Approach to Jointly Modeling Toxicity and Biomarker Expression in a Phase I/II Dose-Finding Trial. *Biometrics* **61**, 343-354.

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

Buyse M and Piedbois P (1996). One the relationship between response to treatment and survival time. *Statistics in Medicine* **15**, 2797-2812.

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

Dazzi F, Szydlo RM, Cross NC and et al. (2000). Durability of responses following donor lymphocyte infusions for patients who relapse after allogeneic stem cell transplantation for chronic myeloid leukemia. *Blood* **15**, 2712-2716.

Fleming TR and Powers JH (2012). Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine* **31**, 2973-2984.

Gauthier J, Yuan Y and Thall PF (2019). Bayesian phase 1/2 trial designs and cellular immunotherapies: a practical primer. *Cell and Gene Therapy Insights* **5**, 1483-1495.

Jin IH, Liu S, Thall PF and Yuan Y (2014). Using data augmentation to facilitate conduct of phase I–II clinical trials with delayed outcomes. *Journal of the American Statistical Association* **109**, 525-536.

Lin R, Thall PF and Yuan Y (2021). A phase I-II basket trial design to optimize dose-schedule regimes based on delayed outcome. *Biostatistics* **16**, 179-202.

Liu E, Marin D, Banerjee P, Macapinlac H, Thall PF, Rezvani K et al. (2020). IL-15 armored CAR-transduced NK cells against CD19 positive B cell tumors. *New England Journal of Medicin* **382**, 545-553.

Lee J, Thall PF and Rezvani K (2019). Optimizing natural killer cell doses for heterogeneous cancer patients on the basis of multiple event times. *Journal of the Royal Statistical Society, Series C* **68**, 462-474.

Lee J, Thall PF and Msaouel P (2020). A phase I-II design based on periodic and continuous monitoring of ordinal disease severity and the times to toxicity and death. *Statistics in Medicine* **39**, 2035-2050.

Mandrekar S, Qin R and Sargent D (2010). Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges. *Statistics in Medicine* **29**, 1077-1083.

Simon RM and Makuch RW (1984). A non-parametric graphical representation of the relationship between survival and the occurrence of an event: Application to responder versus non-responder bias. *Statistics in Medicine* **3**, 35-44.

Thall and 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 PF and Cook JD (2004). Dose-Finding Based on Efficacy–Toxicity Trade-Offs. *Biometrics* **60**, 684-693.

Thall PF, Szabo A, Nguyen HQ, Amlie-Lefond CM and Zaidat OO (2011). Optimizing the concentration and bolus of a drug delivered by continuous infusion. *Biometrics* **67**, 1638-1646.

Thall PF and Nguyen HQ (2012). Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of biopharmaceutical statistics* **22**, 785-801.

Thall PF, Nguyen HQ, Braun TM and Qazilbash MH (2013). Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes. *Biometrics* **69**, 673-682.

Thall PF, Herrick R, Nguyen HQ, Venier JJ and Norris JC (2014). Using effective sample size for prior calibration in Bayesian phase I-II dose-finding. *Clinical Trials* **11**, 251-264.

Thall PF (2020). Statistical Remedies for Medical Researchers. *Springer*.

Tseng YD, Chen YH, Catalano PJ and Ng A (2015). Rates and durability of response to salvage radiation therapy among patients with refractory or relapsed aggressive non-hodgkin lymphoma. *International J Radiation Oncology, Biology, Physics* **91**, 223-231.

Yan F, Thall PF, Lu KH, Gilbert MR and Yuan Y (2018). Phase I-II clinical trial design: A state-of-the-art paradigm for dose finding with novel agents. *Annals of Oncology* **29**, 694-699.

Yuan Y, Nguyen HQ, Thall PF (2016). Bayesian Designs for Phase I-II Clinical Trials. *Chapman & Hall/CRC: New York*.

Zang Y and Lee JJ (2017). A robust two-stage design identifying the optimal biological dose for phase I/II clinical trials. *Statistics in Medicine*, **36**, 27-42.

Zhang Y, Cao S, Zhang C, Jin IH and Zang Y (2021). A Bayesian adaptive phase I/II clinical trial design with late-onset competing risk outcomes. *Biometrics*, in press.

Zhang Y and Zang Y (2021). CWL: A conditional weighted likelihood method to account for the delayed joint toxicityefficacy outcomes for phase I/II clinical trials. *Statistical Methods in Medical Research*, in press.

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

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

Supporting Information

Web Appendix A, referenced in Section **??**, is available with this paper at the Biometrics website on Wiley Online Library.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

Appendix

Title of appendix

Put your short appendix here. Remember, longer appendices are possible when presented as Supplementary Web Material. Please review and follow the journal policy for this material, available under Instructions for Authors at `http://www.biometrics.tibs.org`.

Figure 1. Kaplan-Meier plot of overall survival (OS) for the acute leukemia phase I trial by dose group, defined as Low (doses 1-5 combined) or High (dose 6).

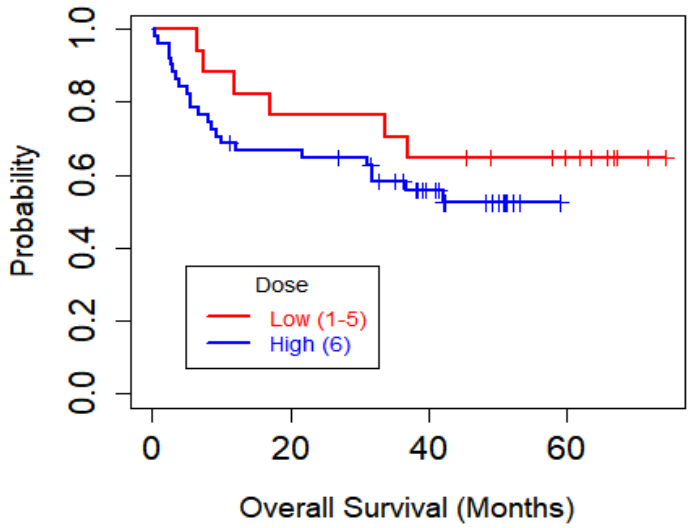

Figure 2. Schematic for the Gen I-II design.

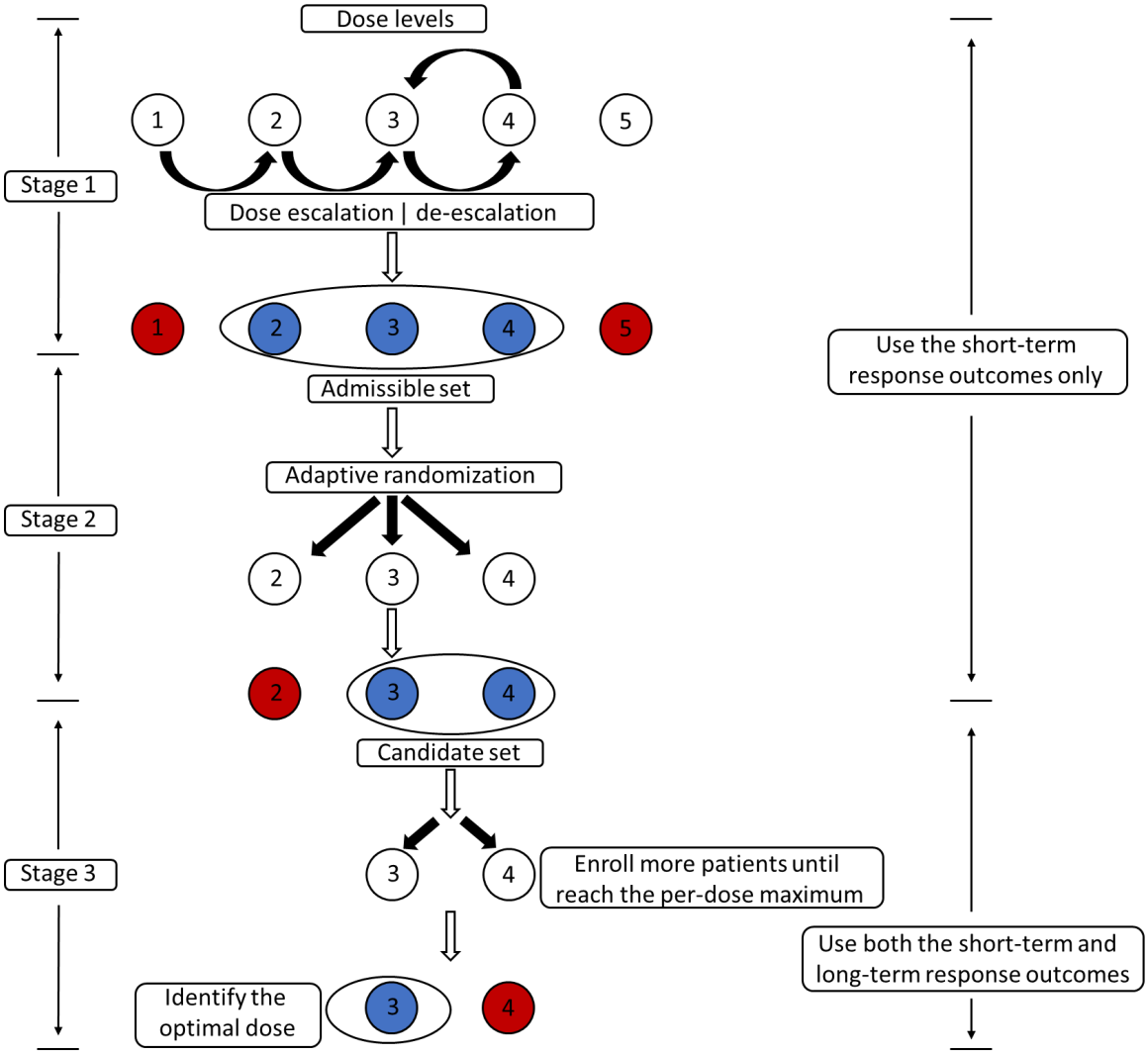

Figure 3. Dose-outcome curves for the 8 scenarios in the simulation study.The dotted, dashed, and solid lines are the $\pi_T^{true}(d_j)$, $\pi_R^{true}(d_j)$, and PFS curves, respectively.

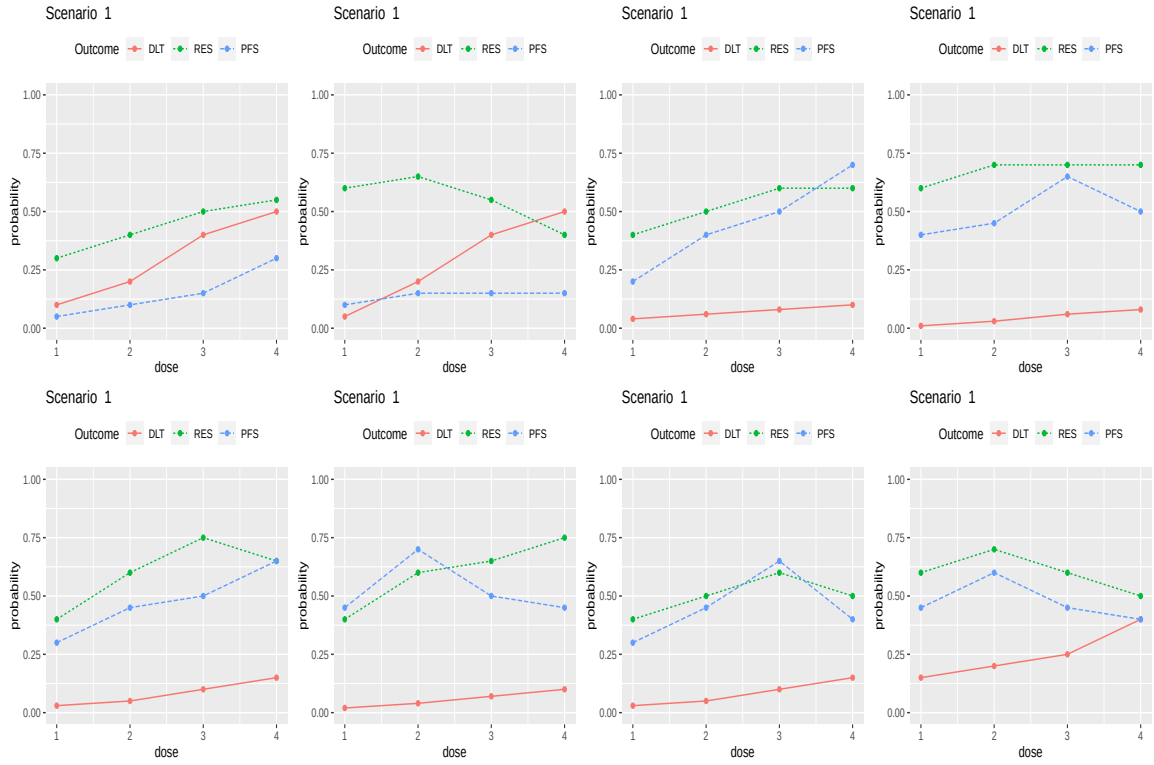

Table 1
Numerical utilities for the outcome $\boldsymbol{Y} = (Y_T, Y_R')$.

		Y_R'		
		$2 = RES$	$1 = SD$	$0 = PD$
Y_T	$0 =$ No DLT	100	50	20
	$1 =$ DLT	60	30	0

Table 2
Dose selection% and mean number of patients treated at each dose level, mean sample size, and $R = \xi(d^{sel,\xi})/\xi(d^{opt,\xi})$ *under the Gen I-II, Conv 1 and Conv 2 designs.* $\xi^{true}(d_j) = \Pr^{true}(Z > 5 \mid Y'_R > 0 \mid d_j)$

Designs		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 1				
	$\pi_T^{true}(d_j)$		0.1	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.3	0.4	0.5	0.55		
	$\overline{U}^{true}(d_j)$		59.1	62.5	60.9	59.9		
	$\xi^{true}(d_j)$		0.05	0.1	0.15	0.3		
Gen I-II	Selection (%)	93.5	0	0.5	2.6	3.3	35.6	NA
	Patients (#)		9.5	13.5	8.8	3.7		
Conv 1	Selection (%)	56.2	6.5	22.6	12.4	2.2	34.9	NA
	Patients (#)		9.4	13.2	8.7	3.6		
Conv 2	Selection (%)	57.7	5.6	22.1	12.4	2.2	35.4	NA
	Patients (#)		9.4	13.4	8.8	3.7		
				Scenario 2				
	$\pi_T^{true}(d_j)$		0.05	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.6	0.65	0.55	0.4		
	$\overline{U}^{true}(d_j)$		76.7	68.0	54.0	40.4		
	$\xi^{true}(d_j)$		0.1	0.15	0.15	0.15		
Gen I-II	Selection (%)	83.8	2.3	9.6	4.1	0.2	48.2	NA
	Patients (#)		21.5	16.4	7.4	2.8		
Conv 1	Selection (%)	3.2	73.3	22.4	1.1	0	46.9	NA
	Patients (#)		21.2	15.9	7.0	2.8		
Conv 2	Selection (%)	3.1	73.2	22.8	0.9	0	48.2	NA
	Patients (#)		21.4	16.6	7.3	2.9		
				Scenario 3				
	$\pi_T^{true}(d_j)$		0.04	0.06	0.08	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.6		
	$\overline{U}^{true}(d_j)$		61.2	67.0	74.2	75.0		
	$\xi^{true}(d_j)$		0.2	0.4	0.5	0.7		
Gen I-II	Selection (%)	2.9	0.3	7.0	19.3	70.5	52.7	91.0
	Patients (#)		10.6	13.0	14.9	14.3		
Conv 1	Selection (%)	2.2	4.6	13.5	39.7	39.9	47.4	79.1
	Patients (#)		9.6	11.7	13.3	12.7		
Conv 2	Selection (%)	2.2	3.6	13.2	40.2	40.8	52.8	79.9
	Patients (#)		10.5	13.1	14.9	14.3		
				Scenario 4				
	$\pi_T^{true}(d_j)$		0.01	0.03	0.06	0.08		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.7	0.7		
	$\overline{U}^{true}(d_j)$		72.1	80.9	81.3	80.6		
	$\xi^{true}(d_j)$		0.4	0.45	0.65	0.5		
Gen I-II	Selection (%)	0.1	4.3	9.2	69.1	17.2	58.1	91.5
	Patients (#)		13.9	15.1	14.8	14.4		
Conv 1	Selection (%)	0	8.7	33.0	31.1	27.7	48.0	80.1
	Patients (#)		11.5	12.7	12.1	11.6		
Conv 2	Selection (%)	0.1	7.0	33.2	32.3	27.3	58.1	80.6
	Patients (#)		13.8	15.3	14.8	14.1		

Table 2 (continued)

Designs		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 5				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.75	0.65		
	$\overline{U}^{true}(d_j)$		63.1	75.2	82.3	72.7		
	$\xi^{true}(d_j)$		0.3	0.45	0.5	0.65		
Gen I-II	Selection (%)	1.2	1.4	14.2	24.1	59.1	54.3	89.1
	Patients (#)		10.5	14.7	15.8	13.3		
Conv 1	Selection (%)	1.0	2.2	22.8	61.3	12.8	47.7	77.1
	Patients (#)		9.1	13.0	14.1	11.4		
Conv 2	Selection (%)	1.1	1.6	20.1	64.6	12.6	54.3	77.5
	Patients (#)		10.2	14.9	16.2	13.1		
				Scenario 6				
	$\pi_T^{true}(d_j)$		0.02	0.04	0.07	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.65	0.75		
	$\overline{U}^{true}(d_j)$		54.4	72.6	72.5	77.8		
	$\xi^{true}(d_j)$		0.45	0.7	0.5	0.45		
Gen I-II	Selection (%)	0.6	5.4	68.4	15.7	9.8	54.3	90.0
	Patients (#)		10.0	14.4	14.7	15.2		
Conv 1	Selection (%)	0.9	1.4	25.1	25.3	47.2	47.7	75.2
	Patients (#)		8.8	12.9	12.8	13.2		
Conv 2	Selection (%)	1.1	1.0	23.9	23.6	50.5	54.2	74.6
	Patients (#)		9.9	14.6	14.5	15.1		
				Scenario 7				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.5		
	$\overline{U}^{true}(d_j)$		66.1	70.3	73.5	68.6		
	$\xi^{true}(d_j)$		0.3	0.45	0.65	0.4		
Gen I-II	Selection (%)	3.3	2.0	15.0	70.9	8.8	52.2	90.6
	Patients (#)		11.2	13.7	15.1	12.1		
Conv 1	Selection (%)	2.8	11.4	26.2	42.9	16.6	47.4	78.9
	Patients (#)		10.1	12.5	14.0	10.7		
Conv 2	Selection (%)	2.8	10.2	26.2	44.6	16.3	52.2	79.7
	Patients (#)		11.1	13.9	15.3	11.8		
				Scenario 8				
	$\pi_T^{true}(d_j)$		0.15	0.2	0.25	0.4		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.6	0.5		
	$\overline{U}^{true}(d_j)$		71.8	77.8	68.5	56.4		
	$\xi^{true}(d_j)$		0.45	0.6	0.45	0.4		
Gen I-II	Selection (%)	9.1	23.4	56.9	9.0	1.6	47.7	90.5
	Patients (#)		17.7	14.8	9.9	5.3		
Conv 1	Selection (%)	9.1	29.3	52.9	8.2	0.5	44.2	89.5
	Patients (#)		16.5	14.1	9.0	4.6		
Conv 2	Selection (%)	9.0	28.3	54.0	8.4	0.3	47.7	89.8
	Patients (#)		17.4	15.2	10.0	5.1		

Table 3
Sensitivity analysis for the AR stage in the Gen I-II design.

Designs		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 1				
	$\pi_T^{true}(d_j)$		0.1	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.3	0.4	0.5	0.55		
	$\overline{U}^{true}(d_j)$		59.1	62.5	60.9	59.9		
	$\xi^{true}(d_j)$		0.05	0.1	0.15	0.3		
With AR	Selection (%)	93.5	0	0.5	2.6	3.3	35.6	NA
	Patients (#)		9.5	13.5	8.8	3.7		
Without AR	Selection (%)	94.8	0	0.5	1.8	2.9	41.4	NA
	Patients (#)		10.8	16.2	10.1	4.2		
				Scenario 2				
	$\pi_T^{true}(d_j)$		0.05	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.6	0.65	0.55	0.4		
	$\overline{U}^{true}(d_j)$		76.7	68.0	54.0	40.4		
	$\xi^{true}(d_j)$		0.1	0.15	0.15	0.15		
With AR	Selection (%)	83.8	2.3	9.6	4.1	0.2	48.2	NA
	Patients (#)		21.5	16.4	7.4	2.8		
Without AR	Selection (%)	86.7	0.8	8.1	3.8	0.5	60.4	NA
	Patients (#)		32.7	18.2	6.9	2.7		
				Scenario 3				
	$\pi_T^{true}(d_j)$		0.04	0.06	0.08	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.6		
	$\overline{U}^{true}(d_j)$		61.2	67.0	74.2	75.0		
	$\xi^{true}(d_j)$		0.2	0.4	0.5	0.7		
With AR	Selection (%)	2.9	0.3	7.0	19.3	70.5	52.7	91.0
	Patients (#)		10.6	13.0	14.9	14.3		
Without AR	Selection (%)	3.0	0.1	5.4	16.4	75.0	69.7	92.7
	Patients (#)		11.6	15.2	21.4	21.5		
				Scenario 4				
	$\pi_T^{true}(d_j)$		0.01	0.03	0.06	0.08		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.7	0.7		
	$\overline{U}^{true}(d_j)$		72.1	80.9	81.3	80.6		
	$\xi^{true}(d_j)$		0.4	0.45	0.65	0.5		
With AR	Selection (%)	0.1	4.3	9.2	69.1	17.2	58.1	91.5
	Patients (#)		13.9	15.1	14.8	14.4		
Without AR	Selection (%)	0	4.0	8.7	70.1	17.1	75.9	91.8
	Patients (#)		14.6	21.0	20.8	19.6		
				Scenario 5				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.75	0.65		
	$\overline{U}^{true}(d_j)$		63.1	75.2	82.3	72.7		
	$\xi^{true}(d_j)$		0.3	0.45	0.5	0.65		
With AR	Selection (%)	1.2	1.4	14.2	24.1	59.1	54.3	89.1
	Patients (#)		10.5	14.7	15.8	13.3		
Without AR	Selection (%)	1.1	1.3	15.0	23.6	59.0	71.3	89.0
	Patients (#)		10.4	18.9	27.0	15.1		

Table 3 (continued)

				Scenario 6				
	$\pi_T^{true}(d_j)$		0.02	0.04	0.07	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.65	0.75		
	$\overline{U}^{true}(d_j)$		54.4	72.6	72.5	77.8		
	$\xi^{true}(d_j)$		0.45	0.7	0.5	0.45		
With AR	Selection (%)	0.6	5.4	68.4	15.7	9.8	54.3	90.0
	Patients (#)		10.0	14.4	14.7	15.2		
Without AR	Selection (%)	0.7	5.5	70.7	14.5	8.6	70.2	90.8
	Patients (#)		9.4	19.4	18.5	22.9		
				Scenario 7				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.5		
	$\overline{U}^{true}(d_j)$		66.1	70.3	73.5	68.6		
	$\xi^{true}(d_j)$		0.3	0.45	0.65	0.4		
With AR	Selection (%)	3.3	2.0	15.0	70.9	8.8	52.2	90.6
	Patients (#)		11.2	13.7	15.1	12.1		
Without AR	Selection (%)	2.8	1.8	13.9	74.6	7.0	69.9	91.9
	Patients (#)		13.7	18.8	21.9	15.5		
				Scenario 8				
	$\pi_T^{true}(d_j)$		0.15	0.2	0.25	0.4		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.6	0.5		
	$\overline{U}^{true}(d_j)$		71.8	77.8	68.5	56.4		
	$\xi^{true}(d_j)$		0.45	0.6	0.45	0.4		
With AR	Selection (%)	9.1	23.4	56.9	9.0	1.6	47.7	90.5
	Patients (#)		17.7	14.8	9.9	5.3		
Without AR	Selection (%)	8.0	20.2	62.5	7.8	1.6	62.7	91.8
	Patients (#)		21.2	24.3	11.8	5.4		

Table 4
Sensitivity analysis for ρ for the candidate set.

ρ		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 1				
	$\pi_T^{true}(d_j)$		0.1	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.3	0.4	0.5	0.55		
	$\overline{U}^{true}(d_j)$		59.1	62.5	60.9	59.9		
	$\xi^{true}(d_j)$		0.05	0.1	0.15	0.3		
0.6	Selection (%)	93.6	0	0.5	2.6	3.3	35.4	NA
	Patients (#)		9.5	13.5	8.7	3.6		
0.7	Selection (%)	93.5	0	0.5	2.6	3.3	35.6	NA
	Patients (#)		9.5	13.5	8.8	3.7		
0.8	Selection (%)	93.9	0	0.5	2.0	3.5	35.0	NA
	Patients (#)		9.4	13.4	8.6	3.6		
0.9	Selection (%)	95.0	0	0.5	2.2	2.3	35.2	NA
	Patients (#)		9.4	13.5	8.7	3.6		
				Scenario 2				
	$\pi_T^{true}(d_j)$		0.05	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.6	0.65	0.55	0.4		
	$\overline{U}^{true}(d_j)$		76.7	68.0	54.0	40.4		
	$\xi^{true}(d_j)$		0.1	0.15	0.15	0.15		
0.6	Selection (%)	82.3	2.2	9.8	5.1	0.6	48.5	NA
	Patients (#)		21.2	16.6	7.8	3.0		
0.7	Selection (%)	83.8	2.3	9.6	4.1	0.2	48.2	NA
	Patients (#)		21.5	16.4	7.4	2.8		
0.8	Selection (%)	86.4	2.4	8.5	2.5	0.2	47.9	NA
	Patients (#)		21.7	16.2	7.1	2.8		
0.9	Selection (%)	90.0	2.4	6.5	1.0	0	47.6	NA
	Patients (#)		21.3	16.4	7.1	2.8		
				Scenario 3				
	$\pi_T^{true}(d_j)$		0.04	0.06	0.08	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.6		
	$\overline{U}^{true}(d_j)$		61.2	67.0	74.2	75.0		
	$\xi^{true}(d_j)$		0.2	0.4	0.5	0.7		
0.6	Selection (%)	3.1	0.4	6.9	19.3	70.4	53.2	91.0
	Patients (#)		11.0	13.2	14.9	14.1		
0.7	Selection (%)	2.9	0.3	7.0	19.3	70.5	52.7	91.0
	Patients (#)		10.6	13.0	14.9	14.3		
0.8	Selection (%)	3.3	0.3	6.9	20.5	69.0	51.6	90.7
	Patients (#)		10.2	12.8	14.6	14.0		
0.9	Selection (%)	3.2	0.7	8.5	28.9	58.6	50.0	87.1
	Patients (#)		9.8	12.2	14.3	13.6		
				Scenario 4				
	$\pi_T^{true}(d_j)$		0.01	0.03	0.06	0.08		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.7	0.7		
	$\overline{U}^{true}(d_j)$		72.1	80.9	81.3	80.6		
	$\xi^{true}(d_j)$		0.4	0.45	0.65	0.5		
0.6	Selection (%)	0.1	4.8	9.6	69.0	16.4	58.6	91.3
	Patients (#)		14.1	15.2	14.8	14.5		
0.7	Selection (%)	0.1	4.3	9.2	69.1	17.2	58.1	91.5
	Patients (#)		13.9	15.1	14.8	14.4		
0.8	Selection (%)	0	4.2	12.3	65.1	18.3	56.5	90.3
	Patients (#)		13.0	14.8	14.5	14.1		
0.9	Selection (%)	0.2	5.7	17.6	54.5	22.0	53.2	87.2
	Patients (#)		12.1	14.1	13.8	13.2		

Table 4 (continued)

ρ		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 5				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.75	0.65		
	$\overline{U}^{true}(d_j)$		63.1	75.2	82.3	72.7		
	$\xi^{true}(d_j)$		0.3	0.45	0.5	0.65		
0.6	Selection (%)	1.4	1.4	13.3	22.9	61.0	54.8	89.6
	Patients (#)		10.7	14.7	15.8	13.6		
0.7	Selection (%)	1.2	1.4	14.2	24.1	59.1	54.3	89.1
	Patients (#)		10.5	14.7	15.8	13.3		
0.8	Selection (%)	1.1	12.8	16.3	30.9	50.5	52.8	86.8
	Patients (#)		9.8	14.4	15.7	12.8		
0.9	Selection (%)	1.6	11.2	18.4	48.8	30.0	50.6	81.9
	Patients (#)		9.3	13.8	15.6	12.0		
				Scenario 6				
	$\pi_T^{true}(d_j)$		0.02	0.04	0.07	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.65	0.75		
	$\overline{U}^{true}(d_j)$		54.4	72.6	72.5	77.8		
	$\xi^{true}(d_j)$		0.45	0.7	0.5	0.45		
0.6	Selection (%)	0.8	5.0	71.4	14.3	8.6	55.2	91.0
	Patients (#)		10.4	14.7	14.8	15.3		
0.7	Selection (%)	0.6	5.4	68.4	15.7	9.8	54.3	90.0
	Patients (#)		10.0	14.4	14.7	15.2		
0.8	Selection (%)	0.6	4.2	64.5	17.0	13.6	52.6	88.7
	Patients (#)		9.2	14.3	14.2	14.9		
0.9	Selection (%)	0.9	1.9	45.8	23.0	28.4	50.8	82.4
	Patients (#)		8.9	13.6	13.7	14.5		
				Scenario 7				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.5		
	$\overline{U}^{true}(d_j)$		66.1	70.3	73.5	68.6		
	$\xi^{true}(d_j)$		0.3	0.45	0.65	0.4		
0.6	Selection (%)	3.0	2.1	14.9	72.4	7.6	52.2	91.1
	Patients (#)		11.1	13.9	15.3	12.0		
0.7	Selection (%)	3.3	2.0	15.0	70.9	8.8	52.2	90.6
	Patients (#)		11.2	13.7	15.1	12.1		
0.8	Selection (%)	3.1	2.3	15.8	70.4	8.4	51.5	90.4
	Patients (#)		10.8	13.5	15.3	11.8		
0.9	Selection (%)	3.6	4.3	20.6	60.6	11.0	50.0	86.7
	Patients (#)		10.5	13.4	14.8	11.3		
				Scenario 8				
	$\pi_T^{true}(d_j)$		0.15	0.2	0.25	0.4		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.6	0.5		
	$\overline{U}^{true}(d_j)$		71.8	77.8	68.5	56.4		
	$\xi^{true}(d_j)$		0.45	0.6	0.45	0.4		
0.6	Selection (%)	8.6	22.5	58.1	9.2	1.6	48.2	90.7
	Patients (#)		17.8	14.9	10.2	5.4		
0.7	Selection (%)	9.1	23.4	56.9	9.0	1.6	47.7	90.5
	Patients (#)		17.7	14.8	9.9	5.3		
0.8	Selection (%)	8.9	22.5	59.2	8.3	1.1	47.2	91.1
	Patients (#)		17.5	15.0	9.7	4.9		
0.9	Selection (%)	9.1	23.3	59.4	7.7	0.5	46.1	91.2
	Patients (#)		17.1	14.8	9.5	4.7		

Table 5
Sensitivity analysis for sample size allocation between stage 2 and stage 3, with stage I sample size fixed at $n_1 = 15$.

n_2; $N(d)$		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 1				
	$\pi_T^{true}(d_j)$		0.1	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.3	0.4	0.5	0.55		
	$\overline{U}^{true}(d_j)$		59.1	62.5	60.9	59.9		
	$\xi^{true}(d_j)$		0.05	0.1	0.15	0.3		
$n_2 = 39$; $N(d) = 9$	Selection (%)	94.9	0	0.5	1.8	2.8	37.4	NA
	Patients (#)		9.6	14.6	9.6	3.6		
$n_2 = 33$; $N(d) = 15$	Selection (%)	93.5	0	0.5	2.6	3.3	35.6	NA
	Patients (#)		9.5	13.5	8.8	3.7		
$n_2 = 27$; $N(d) = 21$	Selection (%)	94.1	0	0.6	2.1	3.2	35.8	NA
	Patients (#)		9.8	13.2	8.9	4.0		
				Scenario 2				
	$\pi_T^{true}(d_j)$		0.05	0.2	0.4	0.5		
	$\pi_R^{true}(d_j)$		0.6	0.65	0.55	0.4		
	$\overline{U}^{true}(d_j)$		76.7	68.0	54.0	40.4		
	$\xi^{true}(d_j)$		0.1	0.15	0.15	0.15		
$n_2 = 39$; $N(d) = 9$	Selection (%)	85.5	1.7	8.4	4.1	0.3	52.8	NA
	Patients (#)		23.9	18.5	7.6	2.8		
$n_2 = 33$; $N(d) = 15$	Selection (%)	83.8	2.3	9.6	4.1	0.2	48.2	NA
	Patients (#)		21.5	16.4	7.4	2.8		
$n_2 = 27$; $N(d) = 21$	Selection (%)	88.6	1.4	7.6	2.1	0.3	50.8	NA
	Patients (#)		21.7	17.8	8.4	3.0		
				Scenario 3				
	$\pi_T^{true}(d_j)$		0.04	0.06	0.08	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.6		
	$\overline{U}^{true}(d_j)$		61.2	67.0	74.2	75.0		
	$\xi^{true}(d_j)$		0.2	0.4	0.5	0.7		
$n_2 = 39$; $N(d) = 9$	Selection (%)	3.1	0.3	6.8	19.8	70.0	53.2	90.9
	Patients (#)		10.3	13.0	15.4	14.5		
$n_2 = 33$; $N(d) = 15$	Selection (%)	2.9	0.3	7.0	19.3	70.5	52.7	91.0
	Patients (#)		10.6	13.0	14.9	14.3		
$n_2 = 27$; $N(d) = 21$	Selection (%)	2.7	0.4	5.1	16.9	74.9	65.2	92.5
	Patients (#)		12.9	16.0	18.4	17.9		
				Scenario 4				
	$\pi_T^{true}(d_j)$		0.01	0.03	0.06	0.08		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.7	0.7		
	$\overline{U}^{true}(d_j)$		72.1	80.9	81.3	80.6		
	$\xi^{true}(d_j)$		0.4	0.45	0.65	0.5		
$n_2 = 39$; $N(d) = 9$	Selection (%)	0.1	5.0	10.6	66.8	17.6	54.2	90.7
	Patients (#)		12.9	14.3	13.8	13.2		
$n_2 = 33$; $N(d) = 15$	Selection (%)	0.1	4.3	9.2	69.1	17.2	58.1	91.5
	Patients (#)		13.9	15.1	14.8	14.4		
$n_2 = 27$; $N(d) = 21$	Selection (%)	0.1	3.2	8.1	73.9	14.6	77.0	92.8
	Patients (#)		17.6	20.0	19.9	19.5		

Table 5 (continued)

n_2; $N(d)$		Dose levels 0	1	2	3	4	Sample size	R (%)
				Scenario 5				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.75	0.65		
	$\overline{U}^{true}(d_j)$		63.1	75.2	82.3	72.7		
	$\xi^{true}(d_j)$		0.3	0.45	0.5	0.65		
$n_2 = 39$; $N(d) = 9$	Selection (%)	1.2	1.4	14.7	22.7	60.0	53.7	89.2
	Patients (#)		9.7	14.9	16.2	12.9		
$n_2 = 33$; $N(d) = 15$	Selection (%)	1.2	1.4	14.2	24.1	59.1	54.3	89.1
	Patients (#)		10.5	14.7	15.8	13.3		
$n_2 = 27$; $N(d) = 21$	Selection (%)	1.1	1.1	13.6	22.9	61.3	68.9	89.7
	Patients (#)		13.0	18.9	19.9	17.1		
				Scenario 6				
	$\pi_T^{true}(d_j)$		0.02	0.04	0.07	0.1		
	$\pi_R^{true}(d_j)$		0.4	0.6	0.65	0.75		
	$\overline{U}^{true}(d_j)$		54.4	72.6	72.5	77.8		
	$\xi^{true}(d_j)$		0.45	0.7	0.5	0.45		
$n_2 = 39$; $N(d) = 9$	Selection (%)	0.7	5.6	70.3	14.6	8.8	53.9	90.6
	Patients (#)		9.8	14.6	14.4	15.0		
$n_2 = 33$; $N(d) = 15$	Selection (%)	0.6	5.4	68.4	15.7	9.8	54.3	90.0
	Patients (#)		10.0	14.4	14.7	15.2		
$n_2 = 27$; $N(d) = 21$	Selection (%)	0.7	4.3	72.8	14.1	8.1	67.7	91.5
	Patients (#)		11.3	18.4	18.7	19.3		
				Scenario 7				
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	0.15		
	$\pi_R^{true}(d_j)$		0.4	0.5	0.6	0.5		
	$\overline{U}^{true}(d_j)$		66.1	70.3	73.5	68.6		
	$\xi^{true}(d_j)$		0.3	0.45	0.65	0.4		
$n_2 = 39$; $N(d) = 9$	Selection (%)	3.7	2.0	14.1	71.2	9.1	53.2	90.8
	Patients (#)		11.1	14.1	16.0	12.0		
$n_2 = 33$; $N(d) = 15$	Selection (%)	3.3	2.0	15.0	70.9	8.8	52.2	90.6
	Patients (#)		11.2	13.7	15.1	12.1		
$n_2 = 27$; $N(d) = 21$	Selection (%)	2.4	1.5	13.3	75.9	7.0	63.9	92.3
	Patients (#)		13.3	16.7	18.6	15.2		
				Scenario 8				
	$\pi_T^{true}(d_j)$		0.15	0.2	0.25	0.4		
	$\pi_R^{true}(d_j)$		0.6	0.7	0.6	0.5		
	$\overline{U}^{true}(d_j)$		71.8	77.8	68.5	56.4		
	$\xi^{true}(d_j)$		0.45	0.6	0.45	0.4		
$n_2 = 39$; $N(d) = 9$	Selection (%)	8.7	22.6	58.0	9.0	1.6	50.0	90.7
	Patients (#)		18.9	16.1	10.0	5.0		
$n_2 = 33$; $N(d) = 15$	Selection (%)	9.1	23.4	56.9	9.0	1.6	47.7	90.5
	Patients (#)		17.7	14.8	9.9	5.3		
$n_2 = 27$; $N(d) = 21$	Selection (%)	8.1	20.9	61.2	8.7	1.1	56.6	91.5
	Patients (#)		19.8	17.4	12.9	6.4		