

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

Peter F. Thall<sup>1\*</sup>, Yong Zang<sup>2</sup>, and Ying Yuan<sup>1</sup>

1. Department of Biostatistics, M.D. Anderson Cancer Center
2. Department of Biostatistics and Health Data Science; Center for Computational Biology  
and Bioinformatics, Indiana University

\*For correspondence: Peter F. Thall, Department of Biostatistics, M.D. Anderson Cancer  
Center, Houston, TX. Email: rex@mdanderson.org

## Abstract

Designs for early phase dose finding clinical trials typically are either phase I based on toxicity, or phase I-II based on toxicity and efficacy. These designs rely on the implicit assumption that the dose of an experimental agent chosen using these short-term outcomes will maximize the agent's long-term therapeutic success rate. In many clinical settings, this assumption is not true. A dose selected in an early phase oncology trial may give suboptimal progression free survival or overall survival time, often due to a high rate of relapse following response. To address this problem, a new family of Bayesian generalized phase I-II designs is proposed. First, a conventional phase I-II design based on short-term outcomes is used to identify a set of candidate doses, rather than selecting one dose. Additional patients then are randomized among the candidates, patients are followed for a predefined longer time period, and a final dose is selected to maximize the long term therapeutic success rate, defined in terms of duration of response. Dose-specific sample sizes in the randomization are determined adaptively to obtain a desired level of selection reliability. The design was motivated by a

phase I-II trial to find an optimal dose of natural killer cells as targeted immunotherapy for recurrent or treatment-resistant B-cell hematologic malignancies. A simulation study shows that, under a range of scenarios in the context of this trial, the proposed design has much better performance than two conventional phase I-II designs.

KEY WORDS: Bayesian Design; Cell Therapy; Dose Finding; Phase I-II Clinical Trial

# 1 Introduction

## 1.1 Conventional Dose Finding Designs

Adoptive T-cell therapy uses immune cells engineered to attack a specific disease target. Cell therapy has been used to treat leukemia and lymphoma [1, 2], and diseases such as type 1 diabetes, Parkinson’s disease, and Alzheimer’s disease. The dose-finding design proposed in this paper was motivated by an early phase trial of CD70 CAR natural killer (NK) cells as targeted immunotherapy for recurrent or treatment-resistant B-cell hematologic malignancies following frontline treatment with chemotherapy or an allogeneic stem cell transplant. The therapeutic aims are to achieve a disease remission, and reduce the rate of subsequent disease recurrence. Treatment begins with three days of chemotherapy to debulk the patient’s disease, followed by one infusion of NK cells at a selected dose. The primary scientific goal is to optimize dose among the four values  $5.0 \times (10^6, 10^7, 10^8, 10^9)$  cells.

Initially, a conventional phase I-II design for this trial was based on indicators  $Y_R$  of response at one month, and  $Y_T$  of grade 3 or 4 toxicity within one month of cell infusion. The plan was to treat up to 48 patients in 16 cohorts of size three, starting at the lowest dose, with the following screening rules. For dose  $d$ , and binary toxicity and response indicators  $Y_T$  and  $Y_R$ , denote  $\pi_R(d, \boldsymbol{\theta}) = \Pr(Y_R = 1 \mid d, \boldsymbol{\theta})$ ,  $\pi_T(d, \boldsymbol{\theta}) = \Pr(Y_T = 1 \mid d, \boldsymbol{\theta})$ , and  $\mathcal{D}_n$  the data from  $n$  patients. Given lower limit .50 on  $\pi_R(d, \boldsymbol{\theta})$  and upper limit .30 on  $\pi_T(d, \boldsymbol{\theta})$ , a dose  $d$  was considered *acceptable* if

$$\Pr\{\pi_R(d, \boldsymbol{\theta}) > .50 \mid \mathcal{D}_n\} > .10 \quad \text{and} \quad \Pr\{\pi_T(d, \boldsymbol{\theta}) < .30 \mid \mathcal{D}_n\} > .10. \quad (1)$$

Bayesian dose acceptability criteria of this form have been used in many phase I-II designs [3, 4, 5]. Subsequently, a more general design was motivated by concerns about response durability, since a patient’s disease may recur soon after a response is achieved. The investi-

gators planned to follow each patient for up to six months to assess whether they were alive with disease in remission at that time. This led to consideration of how data from this later evaluation might be used to help choose an optimal dose. The result is the generalized phase I-II design presented here, which we call *Gen I-II*.

Most early phase trials determine a dose using either a phase I design based on  $Y_T$ , or a phase I-II design based on bivariate binary or ordinal  $(Y_R, Y_T)$  and  $Y_R$ . In oncology, examples of binary response include 50% reduction of a solid tumor or complete remission of acute leukemia. Most of these designs use sequentially outcome-adaptive rules to choose doses for successive patient cohorts. Fair randomization among doses seldom is used, due to concerns that higher doses may be unacceptably toxic. Many phase I designs [6, 7, 8, 9] and phase I-II designs [3, 10, 11, 12, 13, 14, 15, 16] have been proposed. Most of these designs use binary outcomes, although phase I-II designs have been proposed for ordinal outcomes [17], event times [18, 19], or more than two outcomes [5, 20].

A practical requirement in dose finding trials is that the outcomes,  $Y_T$  or  $\mathbf{Y} = (Y_R, Y_T)$ , must be evaluated over a time period,  $[0, t_1]$ , short enough to avoid delaying accrual to evaluate previous patients' outcomes to make adaptive decisions. For designs based on event times, such as the phase I time-to-event continual reassessment method (TITE-CRM) [37] or late-onset efficacy-toxicity phase I-II design [22], follow up intervals must be short enough so that outcome-adaptive decisions can be made without unduly suspending accrual.

## 1.2 A Dose Finding Trial That Failed

A conventional dose finding design may fail if there is a disconnect between short-term response and long term outcomes, such as disease progression or overall survival (OS) time. This problem arose in a trial of allogeneic stem cell transplantation for acute leukemia [23]. to optimize the dose of vorinostat added to a standard preparative regimen. Six doses were

studied using the TiTE-CRM with target toxicity probability 0.30, followed by an expansion cohort at the selected dose. Toxicity was defined as graft failure or grade 4 or 5 non-hematologic, non-infectious toxicity, mucositis, or diarrhea within one month, For response defined as the patient being alive and engrafted at one month, this definition of DLT includes non-response, due to death or graft failure within one month, so this was a phase I-II trial. Since very few DLTs were observed, the TiTE-CRM design rapidly escalated and selected the highest dose as the MTD, where 51 patients were treated, most as an expansion cohort. The final sample sizes were (3, 3, 3, 4, 4, 51) at the six doses. Analysis of the final survival data gave the Kaplan-Meier estimates for doses {1, 2, 3, 4, 5} combined versus dose 6 in Figure 1, showing that patients treated with the selected dose 6 had worse survival than patients given one of the five lower doses. While these results are far from confirmatory, they are very troubling. To design a future trial, the highest dose is undesirable but a lower dose giving superior survival time cannot be determined reliably from the data. Since this trial is unlikely to be repeated, there is no clear path forward. This trial illustrates a disconnect between short term outcomes and dose effects on progression or survival time.

### 1.3 A Flawed Assumption

Phase I and I-II designs assume that, if  $d$  is optimal based on a criterion defined using early outcomes, then  $d$  also maximizes the therapeutic success rate over a long-term follow up period  $[0, t_2]$ , for  $t_2 > t_1$ . To account for this in the Gen I-II design, given fixed  $t_1$  and  $t_2$ , we define duration of response (DOR),  $Z$ , among responders as the time from  $t_1$  to relapse or death. The long-term treatment success criterion is the probability

$$\xi(d, \boldsymbol{\theta}) =_{def} \Pr(Z > t_2 - t_1 \mid d, \boldsymbol{\theta}), \quad (2)$$

where  $\boldsymbol{\theta}$  denotes the model parameter vector. Expression (2) is the probability that a patient who is alive with disease in remission at  $t_1$  is alive without progressive disease (PD) at  $t_2$ .

The Gen I-II design addresses the common problem in oncology that an early response may not be durable, in that a patient who responds by  $t_1$  may relapse before  $t_2$ . Consequently, a dose optimizing  $\phi(d, \boldsymbol{\theta})$  may not optimize  $\xi(d, \boldsymbol{\theta})$ . Response durability has been discussed for radiation oncology [24], donor lymphocyte infusion following relapse after allogeneic bone marrow transplantation [25], and many other areas of oncology. The Gen I-II design is practical in settings where investigators planning a phase I-II trial intend to follow each patient for a longer time  $t_2$ , to obtain data to estimate response durability. Thus, no additional followup is required beyond what already is planned. A dose chosen in an early phase trial may be suboptimal because the distribution of  $(Y_R, Y_T)$  provides little information about the time to progression or death as a function of dose. This problem is more severe with phase I designs. Since phase I designs have severe flaws compared to phase I-II designs [12, 26, 27], we will not consider phase I designs further.

A Gen I-II design uses the distributions of ordinal  $Y_R$  and  $Y_T$  evaluated over  $[0, t_1]$  and  $Z$  evaluated over  $[t_1, t_2]$  among responders to optimize  $d$ . Let  $\phi(d, \boldsymbol{\theta})$  be an objective function, defined in terms of the early outcome distribution  $p(\mathbf{Y} \mid d, \boldsymbol{\theta})$ , used to optimize  $d$  in phase I-II. Examples of  $\phi(d, \boldsymbol{\theta})$  will be given below. Denote the experimental agent by  $X$ , and let  $X(d)$  denote  $X$  administered at  $d$ , to account for the possibility that effects of  $X(d)$  and  $X(d')$  on  $\mathbf{Y}$  or  $Z$  may be different for  $d \neq d'$ . While we use the long-term success criterion  $\xi(d, \boldsymbol{\theta}) = \Pr(Z > t_2 - t_1 \mid d, \boldsymbol{\theta})$ , the mean  $E(Z \mid d, \boldsymbol{\theta})$  may be used. 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  $p(\mathbf{Y} \mid d, \boldsymbol{\theta})$  and  $p(Z \mid d, \boldsymbol{\theta})$  vary with  $d$ , and associations between  $Z$  and  $\mathbf{Y}$ . Conventionally, using  $\phi(d, \boldsymbol{\theta})$  rather than  $\xi(d, \boldsymbol{\theta})$  to optimize  $d$  is motivated primarily by the desire for logistical convenience when conducting a dose-finding trial.

It is easy to show by example that, for assumed true values  $\phi^{true}(d, \boldsymbol{\theta})$  and  $\xi^{true}(d, \boldsymbol{\theta})$ , the optimal doses  $d^{opt,\phi}$  and  $d^{opt,\xi}$  under the two criteria may differ. Several scenarios reflecting this possibility are included in the simulations given below in Section 5. Even if  $Z$  depends on  $\mathbf{Y}$ , an optimal phase I-II dose  $d^{sel,\phi}$  based on  $\phi(d, \boldsymbol{\theta})$  may be suboptimal in terms of  $\xi(d, \boldsymbol{\theta})$ . A causal explanation with targeted agents or cellular therapies is that there may be direct biological effects of  $X(d)$  on  $Z$  not mediated by  $Y_R$ . This is a version of the general problem when dealing with relationships between short-term and long-term outcomes in treatment evaluation, which has been discussed extensively, often with regard to using an early outcome as a surrogate for a long-term outcome. Common examples are response and PFS time, and the times to PD and death [28, 29, 30, 31].

Section 2 presents the general Gen I-II design paradigm. Dose-outcome models are given in section 3. Section 4 describes a utility based version of the Gen I-II design. A simulation study of the Gen I-II design and two conventional phase I-II designs is presented in section 5. We close with a discussion in section 6.

## 2 A Generalized Phase I-II Design Paradigm

### 2.1 Overview of the Design

A Gen I-II design begins with a conventional phase I-II design based on ordinal, possibly binary  $Y_R$  and  $Y_T$ , to screen out unsafe or ineffective doses, and identify a set  $\mathcal{C}$  of acceptable *candidate doses* for later evaluation, rather than selecting one final dose. Additional patients are randomized among the doses in  $\mathcal{C}$  and followed to evaluate the times to progression or death, with a final dose selected to maximize  $\xi(d, \boldsymbol{\theta})$  for  $d \in \mathcal{C}$ . The number of additional patients enrolled after the initial phase I-II portion of the trial is determined adaptively, based on  $\mathcal{C}$  and the numbers of patients treated at its doses, to obtain a high probability

of correctly selecting a dose to maximize  $\xi(d, \boldsymbol{\theta})$  for  $d \in \mathcal{C}$ . The additional patients may be thought of as a *generalized expansion cohort*, but randomized among the doses in  $\mathcal{C}$ , rather than being treated at one dose that may turn out to be suboptimal. The Gen I-II design is a modular paradigm in that any phase I-II design based on bivariate ordinal  $\mathbf{Y}$  may be used, provided that it includes a dose optimality criterion  $\phi(d, \boldsymbol{\theta})$ .

A Gen I-II design has three stages. Stages 1 and 2 together comprise the nominally “phase I-II” portion of the trial. In stage 1, a conventional phase I-II design is based on  $\mathbf{Y}$  evaluated over  $[0, t_1]$ . Any phase I-II design with an objective function  $\phi(d, \boldsymbol{\theta})$  characterizing dose desirability may be used. 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  $\mathcal{C}$  of candidate doses with estimated  $\phi(d, \boldsymbol{\theta})$  close to the maximum estimate is determined. In stage 3, additional patients are randomized among the doses in  $\mathcal{C}$ , and all patients are followed for a longer period  $[0, t_2]$  to obtain data on  $(d, Z)$ . At the end of stage 3, the candidate dose maximizing the posterior mean of  $\xi(d, \boldsymbol{\theta})$  is selected.

## 2.2 Design Construction and Trial Conduct

A Gen I-II design may be constructed in numerous ways, depending on how  $\mathbf{Y}$  and  $\phi(d, \boldsymbol{\theta})$  are defined and the probability models  $p(\mathbf{Y} \mid d, \boldsymbol{\theta})$  and  $p(Z \mid d, \boldsymbol{\theta})$ . To make things concrete for the Gen I-II design that we will use to illustrate the methodology, 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  $Y'_R = 2$  for response (RES),  $Y'_R = 1$  for stable disease (SD), and  $Y'_R = 0$  for progressive disease or death (PD). Denoting the indicator of the event  $A$  by  $I[A]$ , we define the binary response indicator  $Y_R = I[Y'_R = 2]$ . We include the third event  $SD = (RES \cup PD)^c = [Y'_R = 1]$  to accommodate settings where  $RES$  and  $PD$  are not complementary, that is, a patient may not have a



response but this was not due to early PD. Other early outcomes may be used, including ordinal toxicity with three or more levels of severity, or  $Y'_R$  with more than three levels, with appropriate modifications of the Gen I-II design parameters. When either  $Y_R$  or  $Y_T$  has three or more ordinal levels, a binary version of each must be defined in order to specify the dose admissibility criteria (1).

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 to obtain a desired level of final dose selection reliability, as described below. Examples of  $\phi(d, \boldsymbol{\theta})$  based on bivariate binary  $\mathbf{Y} = (Y_R, Y_T)$  include the response probability  $\pi_R(d, \boldsymbol{\theta})$  [33], the odds ratio defined in terms of  $\pi_R(d, \boldsymbol{\theta})$  and  $\pi_T(d, \boldsymbol{\theta})$ , and the trade-off function  $f\{\pi_R(d, \boldsymbol{\theta}), \pi_T(d, \boldsymbol{\theta})\}$  used by the EffTox design [3, 34]. If numerical outcome utilities,  $U(\mathbf{Y})$ , are elicited, then the optimality criterion may be the mean utility  $\phi(d, \boldsymbol{\theta}) = E\{U(\mathbf{Y}) \mid d, \boldsymbol{\theta}\}$  [5, 14, 15, 16, 17, 35, 36]. In our construction of a Gen I-II design, below, we take a utility based approach with  $\mathbf{Y} = (Y'_R, Y_T)$ .

For the  $i^{\text{th}}$  patient enrolled in a Gen I-II trial, denote the assigned dose by  $d_{[i]}$  and let  $V_i$  be the independent right censoring time of  $Z_i$  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 thus  $Z_i^o = \min\{Z_i, V_i\}$ . Let  $\delta_i = 1$  if  $Z_i^o = Z_i$  and  $\delta_i = 0$  if  $Z_i^o = V_i < Z_i$ , so the data from the first  $n$  patients enrolled in the trial is

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

Stages 1 and 2 of the Gen I-II design include dose acceptability criteria of the form given in equation (1), with fixed lower limit  $\underline{\pi}_R$  for response probability and upper limit  $\bar{\pi}_T$  for toxicity probability. Denote the set of acceptable doses satisfying (1) 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, 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  $d_j$ ,  $j = 1, \dots, J$  and  $n = 1, \dots, n_{1,2}$ , denote  $\hat{\phi}_{j,n} = E\{\phi(d_j, \boldsymbol{\theta}) \mid \mathcal{D}_n\}$ . 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 a fixed shrinkage parameter  $0 < \zeta \leq 1$ , AR probabilities may be defined to be proportional to  $(\hat{\phi}_{j,n})^\zeta$ . A formula for the AR probabilities is given in the Supplement. Compared to maximizing  $\hat{\phi}_{j,n}$ , AR distributes patients more evenly among acceptable doses during stage 2, which gives a more even distribution of patients among the candidate doses. While using AR is not a requirement of the Gen I-II design, our simulations will show that AR improves final correct selection probabilities and reduces additional stage 3 per-dose sample sizes.

At the end of stage 2, 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,

$$\mathcal{C} = \{d_j \in \mathcal{A}_{n_{1,2}} : \hat{\phi}_{j,n_{1,2}} \geq \rho \max_{d_l \in \mathcal{A}_{n_{1,2}}} \hat{\phi}_{l,n_{1,2}}\}. \quad (4)$$

The parameter  $\rho$  determines how close the estimated optimality criterion of a dose must be to the maximum for it to be in  $\mathcal{C}$ . Preliminary simulations examining several numerical values, such as  $\rho = .60, .70$ , and  $.80$ , should be used to identify a value giving a design with good OCs. The value  $\rho = .70$  was chosen for the CAR NK cell Gen I-II trial design.

The stage 3 sample size  $n_3$  is determined adaptively using the data  $\mathcal{D}_{n_{1,2}}$  and the per-dose subsample sizes  $\{n_{1,2}(d_j) : d_j \in \mathcal{C}\}$  at the end of stage 2. The  $n_{1,2}(d_j)$  values 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  $n_3(d_j)$ , so that  $n_3 = \sum_{d_j \in \mathcal{C}} n_3(d_j)$  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 is done based on simulations of the trial, for given  $n_1, n_2, \rho,$  and assumed true values of the long term success probabilities,  $\boldsymbol{\xi}^{true} = (\xi^{true}(d_1), \dots, \xi^{true}(d_J))$ , and short term success probabilities,  $\boldsymbol{\phi}^{true} = (\phi^{true}(d_1), \dots, \phi^{true}(d_J))$ . 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  $n_3(d_4) = 14$ . Thus, 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.

To illustrate the per-dose sample size determination process between stages 2 and 3, we use scenario 3 of our simulation study, which is given in detail in Table S1 of the supplementary materials. We consider three different combinations of the total sample size  $n_2$  used in stage 2 and total sample  $N(d)$  at  $d,$  specifically  $(n_2, N(d)) = (39,9), (33,15),$  and  $(27,21)$ . Simulations of the Gen I-II design under scenario 3 using each of these sample size configurations show that  $(n_2, N(d)) = (39,9)$  gives very slightly worse results than  $(33,15)$  in terms of both the optimal dose selection percentages (70.0% vs 70.5%) and average total sample sizes (53.2 vs 52.7). The combination  $(27,21)$  gives the highest optimal dose selection percentage of 74.9%, but larger average total sample size 65.2. The sample size setting  $(n_2, N(d)) = (33, 15)$  was chosen by considering the tradeoff between correct true optimal dose selection percentage and sample size. In this setting, if the investigators were willing to treat an expected total of about  $65 - 48 = 17$  more patients in stage 3, rather than  $53 - 48 = 5$  more, as the price to obtain an improvement from 70.5% to 74.9% in optimal dose selection percentage, then the third pair  $(n_2, N(d)) = (27, 21)$  could be used.

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

$$\Pr\{\xi(d, \boldsymbol{\theta}) > \underline{\xi} \mid \mathcal{D}_N\} > .10, \tag{5}$$

where  $\underline{\xi}$  is a fixed lower limit for  $\xi(d_j, \boldsymbol{\theta})$ . Denote the final set of acceptable doses in  $\mathcal{C}$  by  $\mathcal{A}_N^\xi$ . The futility requirement (5) reduces the chance of selecting a dose from a set of candidates that all are unlikely to have a long term success rate at least  $\underline{\xi}$ . In practice,  $\underline{\xi}$  may be the historical mean of  $\xi$  with standard therapy. The final selected optimal dose in the acceptable dose set  $\mathcal{A}_N^\xi$  is defined to maximize the posterior mean long term success probability,

$$d_N^{sel,\xi} = \operatorname{argmax}_{d_j \in \mathcal{A}_N^\xi} E\{\xi(d_j, \boldsymbol{\theta}) \mid \mathcal{D}_N\}. \quad (6)$$

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

We assume a Bayesian model to exploit the Bayesian paradigm's ability to fully account for uncertainty and provide shrinkage toward the prior for posterior criteria used to make decisions. 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  $p(\mathbf{Y} \mid \mathbf{d}, \boldsymbol{\theta}_1)$ , 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  $p(Z \mid d, \mathbf{Y}, \boldsymbol{\theta}_2)$ .

### 3 Dose Outcome Models

We assume the following Bayesian multinomial-Dirichlet model for the early outcome  $\mathbf{Y} = (Y'_R, Y_T)$ . More elaborate models may be used, but we found that this model gives a design with good properties while avoiding possibly restrictive assumptions. For each dose  $d_1, \dots, d_J$ , and outcome indices  $a = 0, 1, 2$  for  $Y'_R$  and  $b = 0, 1$  for  $Y_T$ , denote the joint probability  $p_{a,b}(d_j) = \Pr(Y'_R = a, Y_T = b \mid d_j)$ , with  $\mathbf{p}(d_j) = (p_{0,1}(d_j), \dots, p_{1,2}(d_j))$ . Thus,

the model parameter vector is  $\boldsymbol{\theta}_1 = (\mathbf{p}(d_1), \dots, \mathbf{p}(d_J))$ . For each dose  $d_j$  and interim sample size  $n$ , we assume that the six-dimensional outcome count vector

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

is multinomial with parameters  $n(d_j)$  and  $\mathbf{p}(d_j)$ , and that  $\mathbf{p}(d_j)$  follows a Dirichlet prior with parameter  $1/6$  in each cell. While this model does not borrow strength between doses, it is robust since it makes no assumptions about dose-response curves, and facilitates posterior computation because  $\mathbf{p}(d_j) \mid \mathbf{X}_n(d_j)$  is Dirichlet with parameters  $(1/6, \dots, 1/6) + \mathbf{X}_n(d_j)$  for each  $d_j$ . The early outcome objective function is the mean utility

$$\phi(d_j, \boldsymbol{\theta}_1) = \bar{U}(d_j, \boldsymbol{\theta}_1) = \sum_{a=0}^2 \sum_{b=0}^1 U(a, b) p_{a,b}(d_j) \quad \text{for } j = 1, \dots, J. \quad (7)$$

This Multinomial-Dirichlet model and definition of  $\phi(d_j, \boldsymbol{\theta}_1)$  may be extended easily to accommodate any discrete bivariate ordinal  $(Y_T, Y'_R)$ .

For the distribution of  $Z$ , due to limited sample size a flexible but parsimonious model is needed. We thus assume that  $Z$  follows a Weibull distribution with pdf

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

where  $\alpha > 0$  is the shape parameter and the rate parameter  $\lambda$  is given by

$$\lambda(Y_T, d_j, \boldsymbol{\theta}_2) = \exp(\beta_0 + \beta_T Y_T + \gamma_j) \quad (8)$$

with  $\gamma_1 = 0$ . We denote  $\boldsymbol{\theta}_2 = (\alpha, \beta_0, \beta_T, \gamma_2, \dots, \gamma_J)$  and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ . Because the Weibull is an accelerated failure time model, the parameters in (8) are effects on  $Z$ , with  $\beta_T$  for toxicity, and  $\gamma_j$  the  $d_j$  versus  $d_1$  effect, for  $j \geq 2$ . There is no  $Y_R$  effect since  $Z$  is defined only

if  $Y_R = 1$ . A different distribution may be used, provided that it includes parameters for  $Y_T$  and  $d_j$ . Non-informative  $N(0, 10^2)$  priors for elements of  $\theta_2$  and a  $Gamma(0.01, 0.01)$  prior for  $\alpha$  are assumed. The likelihood for data  $\mathcal{D}_n$  is given in Supplement.

## 4 A Utility Based Gen I-II Design

Because patients in the 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 compared to its baseline severity, and RES as complete remission. To establish a utility, we first fixed  $U(1, 0) = 0$  for the worst and  $U(0, 2) = 100$  for the best possible outcome, and then determined the four remaining intermediate values, subject to the admissibility constraints  $U(a, 0) \leq U(a, 1) \leq U(a, 2)$  for  $a = 0, 1$ , and  $U(1, b) \leq U(0, b)$  for  $b = 0, 1$ , or  $2$ . This formalizes the idea that either better disease status or absence of toxicity is more desirable. Table 1 gives the utility function used for the simulations. The short term outcome objective function is the mean utility (7). During stages 1 and 2, given interim data  $\mathcal{D}_n$ , the posterior mean utility is  $u(d_j, \mathcal{D}_n) = E\{\bar{U}(d_j, \theta) \mid \mathcal{D}_n\}$ . Stage 1 of the Gen I-II design is conducted as follows.

### Steps for Stage 1

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

Step 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  $j = 1, 2, 3, 4$ .

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

Step 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.

Step 5. If  $\mathcal{A}_n$  is not empty and the current dose  $d_j$  is the highest untried dose and satisfies

$\Pr\{\pi_T(d_j, \boldsymbol{\theta}) < \bar{\pi}_T \mid \mathcal{D}_n\} > .10$ , escalate one dose level. This requirement supersedes Step 4.

6. Repeat steps 1-5 until  $n_1/c$  cohorts have been treated and their values of  $\mathbf{Y}$  evaluated.

We include Step 5 because, due to simplicity of the assumed model,  $\mathbf{p}(d_j)$  cannot be estimated for untried  $d_j$ . Step 5 reduces the chance getting stuck at a locally optimal but globally suboptimal dose, because it provides a way to explore untried doses.

## Stage 2

Adaptively randomize up to  $n_2/c$  additional cohorts of patients among the doses in  $\mathcal{A}_{n_{1,2}}$ . The admissible dose set  $\mathcal{A}_n$  is updated after each cohort's outcomes  $\mathbf{Y}$  have been evaluated. In the trial,  $c = 3$ ,  $n_1 = 15$ , and  $n_2 = 33$ , so stage 1 includes up to five cohorts, stage 2 includes up to 11 cohorts, and  $\zeta = 0.5$  is used to define the AR probabilities.

## Stage 3

Fairly randomize an additional  $n_3 = \sum_{d_j \in \mathcal{C}} n_3(d_j)$  patients among the doses in  $\mathcal{C}$ . The per-dose stage 3 sample sizes are chosen adaptively to ensure that a three-stage total of  $n_{1,2}(d_j) + n_3(d_j) = N(d_j)$  patients are treated at each  $d_j \in \mathcal{C}$ . 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  $[Z > t_2 - t_1]$  that a patient is alive with disease in remission at  $t_2$ , and  $\xi(d_j, \boldsymbol{\theta}) = S_Z(t_2 - t_1 \mid Y'_R > 0, d_j, \boldsymbol{\theta})$ . Denoting final overall sample size by  $N$ , a dose  $d^{opt, \xi}$  is chosen at the end of stage 3 to maximize  $E\{\xi(d_j, \boldsymbol{\theta}) \mid \mathcal{D}_N\}$ . We defined  $\mathcal{C}$  using  $\rho = 0.70$  based on preliminary simulations, the value  $N(d_j) = 15$  was chosen adaptively, and the long term success event was  $[Z > 5]$ . We used JAGS to run Markov chain Monte Carlo to generate posterior samples.

For the different long-term goal of choosing a dose to maximize expected PFS time over  $[0, t_2]$ , a simpler approach would be to randomize a predetermined sample of patients among the doses, follow all patients until treatment failure (progression or death) up to time  $t_2$ , and

use the right-censored PFS time data to choose an optimal dose. To protect patient safety, an additional monitoring rule to shut down excessively toxic doses would be required. A Gen I-II design thus provides a practical approach, based on both  $\mathbf{Y}$  and  $Z$ , that may be considered intermediate between a conventional phase I-II design based on  $\mathbf{Y}$  evaluated over  $[0, t_1]$ , and the simpler approach of randomizing and evaluating PFS time over  $[0, t_2]$ .

## 5 Simulation study

In this section, we present a simulation study to evaluate OCs of the utility based Gen I-II design, using the CAR-NK cell trial to specify design settings. We investigated scenarios with a variety of different patterns for  $\phi(d_j)^{true}$ ,  $\xi(d_j)^{true}$ , and outcome distributions. Figure 3 shows the assumed true dose-outcome curves  $\pi_T^{true}(d_j) = \Pr^{true}(Y_T = 1 \mid d_j)$ ,  $\pi_R^{true}(d_j) = \Pr^{true}(RES \mid d_j)$ , and  $\xi^{true}(d_j) = \Pr^{true}(Z > 5)$ . As comparators, we used two more conventional utility-based phase I-II designs, Conv 1 and Conv 2. The Conv 1 design consists of stages 1 and 2 of the Gen I-II design, and selects an optimal dose to maximize the posterior mean utility  $u(d_j, \mathcal{D}_n)$ . While Conv 1 may appear to be a straw man, since the Gen I-II design is almost certain to outperform it, we include Conv 1 because it is what would be used in practice. The Conv 2 design is nearly identical to Conv 1, with the one difference that more patients are randomized in stage 2 in order to match the Gen I-II design’s sample size, as a more fair comparison. Formulas for distributions used to generate  $(Y'_R, Y_T)$  and  $Z$  in the simulations are given in the Supplementary Material. We simulated 5,000 trials under each scenario using each design.

Table 2 summaries OCs of the Gen I-II, Conv 1, and Conv 2 designs, including dose selection percentages, mean number of patients treated at each dose, and mean overall sample size. The number tabled under “dose 0” is the percentage of trials stopped early with no dose selected. A summary statistic to evaluate performance by comparing the selected optimal



dose to the truly optimal dose is  $R(d^{sel,\xi}) = \xi(d^{sel,\xi})/\xi(d^{opt,\xi})$ , which has domain  $[0, 1]$ , with  $R(d^{sel,\xi}) = 1$  corresponding to always selecting the dose that maximizes long term treatment success probability. 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 acceptably low  $\pi_T^{true}(d_j)$  and acceptably high  $\xi(d)^{true}$ . 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 Conv 1 and Conv 2. In scenario 2, the Gen I-II design terminates the trial early 83.8% of the time. Because  $d_1$  and  $d_2$  both have large  $\pi_R^{true}(d_j)$  and the conventional designs ignore  $Z$ , both conventional designs have only about a 3% chance of stopping early, and a 73% chance of incorrectly selecting  $d_1$  as optimal. Scenario 2 illustrates the advantage that the Gen I-II design includes an admissibility requirement defined in terms of  $\xi(d)^{true}$ , while the conventional designs do not, and consequently they both have a high risk of selecting a dose with a low long term success rate. In scenarios 1 and 2, because no truly optimal dose exists,  $R$  is undefined.

In scenarios 3 and 4, multiple doses are nearly optimal in terms of  $\bar{U}^{true}(d_j)$  based on  $\mathbf{Y}$ , but only one dose is optimal based on the long-term criterion  $\xi^{true}(d_j)$ . In scenario 3,  $d_3$  and  $d_4$  have similar mean utilities near 75, while  $d_4$  is truly optimal with highest  $\xi^{true}(d_4) = 0.70$ , compared to  $\xi^{true}(d_3) = 0.50$ . In scenario 4,  $d_2, d_3$  and  $d_4$  have similar mean utilities  $\bar{U}^{true}(d_j)$  near 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% chance 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  $d_2$  or  $d_4$ , because both conventional designs ignore  $Z$ .

In scenarios 5 and 6, the truly optimal dose in terms of  $\xi^{true}(d_j)$  and the dose with highest mean utility  $\bar{U}^{true}(d_j)$  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  $\bar{U}^{true}(d_3) = 82.3$ . In scenario 6,  $d_2$  is truly

optimal with the highest  $\xi^{true}(d_2) = 0.70$ , whereas  $d_4$  has highest mean utility  $\bar{U}^{true}(d_4) = 77.8$ . The Conv 1 and Conv 2 designs both have over a 60% chance of incorrectly selecting  $d_3$  as optimal in scenario 5, about 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 are identical. The Gen I-II design 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 percentages of 56.9% and 54.0%, respectively. In scenario 9,  $\pi_R(d)^{true}$  is flat and  $\pi_T(d)^{true}$  increases with  $d$ , so the mean utility  $\bar{U}^{true}(d_j)$  decreases monotonically with  $d$ , but the truly optimal dose in terms of  $\xi^{true}(d_j)$  is  $d_3$ . In this case, the Gen-I-II design outperforms both Conv 1 and Conv 2, with about a 35% higher correct selection percentage.

In summary, in all scenarios, the Gen I-II design outperforms the conventional phase I-II designs substantially, with the highest R, that is at least 10% higher in each of scenarios 3 – 7. The three designs have similar patient allocation distributions, essentially because 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 performed additional sensitivity analyses to explore several other aspects of the Gen I-II design. The results are summarized in the online supporting information. Table S1 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 up to 48 patients with dose-finding done to maximize  $u(d_j, \mathcal{D}_n)$  for all cohorts. Table S1 shows that, compared to the original “with AR” version of Gen I-II, the “without AR” version has substantially inflated sample sizes, with at most a mild

gain of  $\leq 5\%$  in true optimal dose selection percentage. This shows that AR is a very useful component of the Gen I-II design that provides a large savings in sample size.

Recall that we fixed  $\rho = 0.70$  for defining a candidate dose set. Table S2 summarizes the OCs of the Gen I-II design for values of  $\rho$  ranging from 0.60 to 0.90. The results indicate that larger  $\rho$  is more favorable under the null scenarios 1 and 2, while smaller  $\rho$  is more favorable when a truly optimal dose exists, in scenarios 3 – 9. In practice,  $\rho$  should be chosen, based on preliminary simulations, to accommodate the application at hand.

Table S3 shows effects of changing patient allocation between stages 2 and 3. Given the values  $(n_2, N(d)) = (33, 15)$  used in Table 2, we considered the two alternative pairs,  $(39, 9)$  and  $(27, 21)$ . The results show that the allocation  $(33, 15)$  and the alternative  $(39, 9)$  give very similar design performances, and that both give better OCs compared to  $(27, 21)$  in terms of the tradeoff between correct true optimal dose selection percentage and sample size.

## 6 Discussion

By using data on duration of response, the Gen I-II design addresses an important problem with conventional phase I-II methods. The design is modular, since  $\mathbf{Y}$  can be any ordinal early outcomes used by a phase I-II design, and any criterion  $\phi(d, \boldsymbol{\theta})$  can be used for stages 1 and 2. Thus, a Gen I-II design can be tailored to accommodate the particular clinical setting at hand. The Gen I-II design is practical in settings where investigators plan to follow patients long enough to assess response duration, which commonly is done in oncology trials. The main additional requirement is the sample of patients randomized among candidate doses in stage 3. Our simulations showed that about 15 more patients per candidate dose gives a reliable design. While we have investigated a utility-based Gen I-II design with a simple Bayesian model, the large advantages over conventional designs in our simulations suggest that other Gen I-II designs also will provide a large benefit over conventional designs.

Guo and Yuan [32] proposed a dose-ranging approach to optimizing dose (DROID) for oncology drug development. The Gen I-II design and DROID design share some high-level design strategies, including identifying an admissible dose set based on short-term endpoints, randomizing patients within the admissible dose set, and using both short-term and long-term endpoints for logistical convenience and identifying an optimal dose. However, they focus on different clinical settings. DROID considers binary toxicity and a continuous surrogate efficacy endpoint, e.g. pharmacodynamics, whereas Gen I-II considers early phase I-II toxicity and efficacy endpoints and a long term event time endpoint. This difference requires very different dose-outcome models. Stages 1 and 2 of a Gen I-II design follow the phase I-II paradigm [12], while DROID identifies both a minimal active dose (MAD) and MTD. The randomized portion of the DROID design, used for inference and decision making, does conventional dose-ranging, whereas stage 3 of the Gen I-II design is similar to a multi-arm randomized trial and identifies the dose with largest response duration.

A *caveat* is that, in settings using survival time rather than response duration to define long term treatment success, a Gen I-II design's behavior will depend on relationships between  $\mathbf{Y}$  and survival time. This is a complex issue involving persistence of biological treatment effects over time, and effects of salvage therapy given at relapse on subsequent survival. How the Gen I-II paradigm behaves compared to conventional phase I-II designs in such settings is an important area for future research. R code for implementing the Gen I-II design is available from <https://github.com/yongzang2020>.

## References

- [1] Neelapu, S.S., Locke, F.L., Bartlett, N.L., et al. (2017) Axicabtagene Ciloleucel CAR T-Cell therapy in refractory large B-Cell lymphoma. *N Engl J Med*, **377**, 2531-2544.
- [2] Liu E., Marin, D., Banerjee, P. et al. (2020) Use of CAR-transduced natural killer cells

- in CD19-positive lymphoid tumors. *N Engl J Med*, **382**, 545-553.
- [3] Thall PF and Cook JD (2004). Dose-Finding Based on Efficacy–Toxicity Trade-Offs. *Biometrics* **60**, 684-693.
- [4] 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.
- [5] Liu, S., Guo, B. and Yuan, Y. (2018). A Bayesian phase I/II design for immunotherapy trials. *Journal of the American Statistical Association*, **113**, 1016-1027.
- [6] O’Quigley J, Pepe M, Fisher L. (1990) Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, **46**, 3348.
- [7] Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine* **17** 1103-1120.
- [8] Liu S, Yuan Y. (2015) Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, **64**, 507523.
- [9] Chu, Y., Pan, H., and Yuan, Y. (2016) Adaptive dose modification for phase I clinical trials. *Statistics in Medicine* **35**, 3497-3508.
- [10] 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.
- [11] Zhang W, Sargent DJ and Mandrekar S (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine* **25**, 2365-2383.
- [12] Yuan Y, Nguyen HQ, Thall PF (2016). Bayesian Designs for Phase I-II Clinical Trials. *Chapman & Hall/CRC: New York*.

- [13] 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.
- [14] Guo, B. and Yuan, Y. (2017). Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *Journal of the American Statistical Association*, **112**, 508-520.
- [15] Zhou Y, Lee JJ, Yuan Y. (2019) A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Statistics in Medicine* **38(28)**:5299-5316.
- [16] Lin R, Zhou Y, Yan F, Li D, Yuan Y. (2020) BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies. *JCO Precision Oncology* **4**: 1393-1402.
- [17] 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.
- [18] 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*, **77**: 796-808.
- [19] Zhang Y and Zang Y (2021). CWL: A conditional weighted likelihood method to account for the delayed joint toxicity-efficacy outcomes for phase I/II clinical trials. *Statistical Methods in Medical Research*, **30**: 892-903.
- [20] 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.

- [37] Cheung YK and Chappell R (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56**, 1177-1182.
- [22] 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.
- [23] Al-Atrash, G., Saberian, C., Bassett, R, Thall, P.F., Kebriaei, P., et al. (2022). Vorinostat combined with busulfan, fludarabine, and clofarabine conditioning regimen for allogeneic hematopoietic cell transplantation in patients with acute leukemia: long-term study outcomes. *Transplantation and Cellular Therapy* **28**, 501.e1 - 501.e7.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] Anderson JR, Cain KC and Gelber RD (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology* **1**, 710-719.

- [29] 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.
- [30] Buyse M and Piedbois P (1996). One the relationship between response to treatment and survival time. *Statistics in Medicine* **15**, 2797-2812.
- [31] Fleming TR and Powers JH (2012). Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine* **31**, 2973-2984.
- [32] Guo B and Yuan Y (2023). DROID: Dose-ranging Approach to Optimizing Dose in Oncology Drug Development. *Biometrics*, to appear.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] Murray, T., Thall, P., Yuan, Y., McAvoy, S. and Gomez, D. (2017) Robust treatment comparison based on utilities of semi-competing risks in non-small-cell lung cancer. *Journal of the American Statistical Association*, **112**, 11-23.
- [37] Cheung YK and Chappell R. (2000) Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56**, 1177-1182.



Table 1: Numerical utilities for the early outcomes  $\mathbf{Y} = (Y_T, Y'_R)$ .

		$Y'_R$		
		$2 = RES$	$1 = SD$	$0 = PD$
$Y_T$	$0 = \text{No DLT}$	100	50	20
	$1 = \text{DLT}$	60	30	0

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).

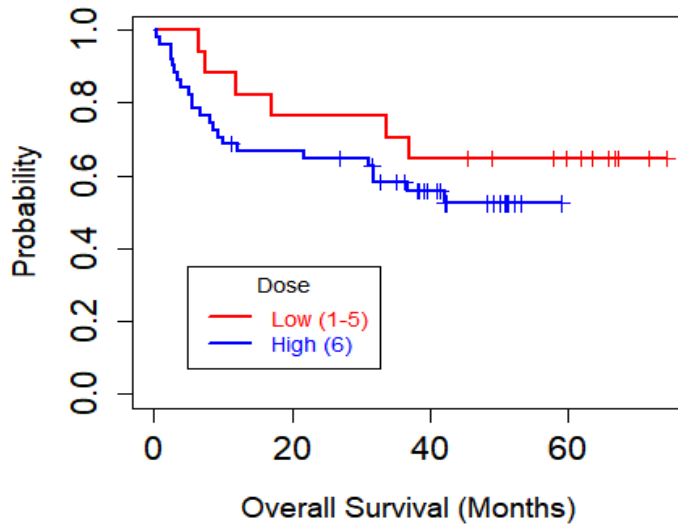


Figure 2: Schematic for the Gen I-II design.

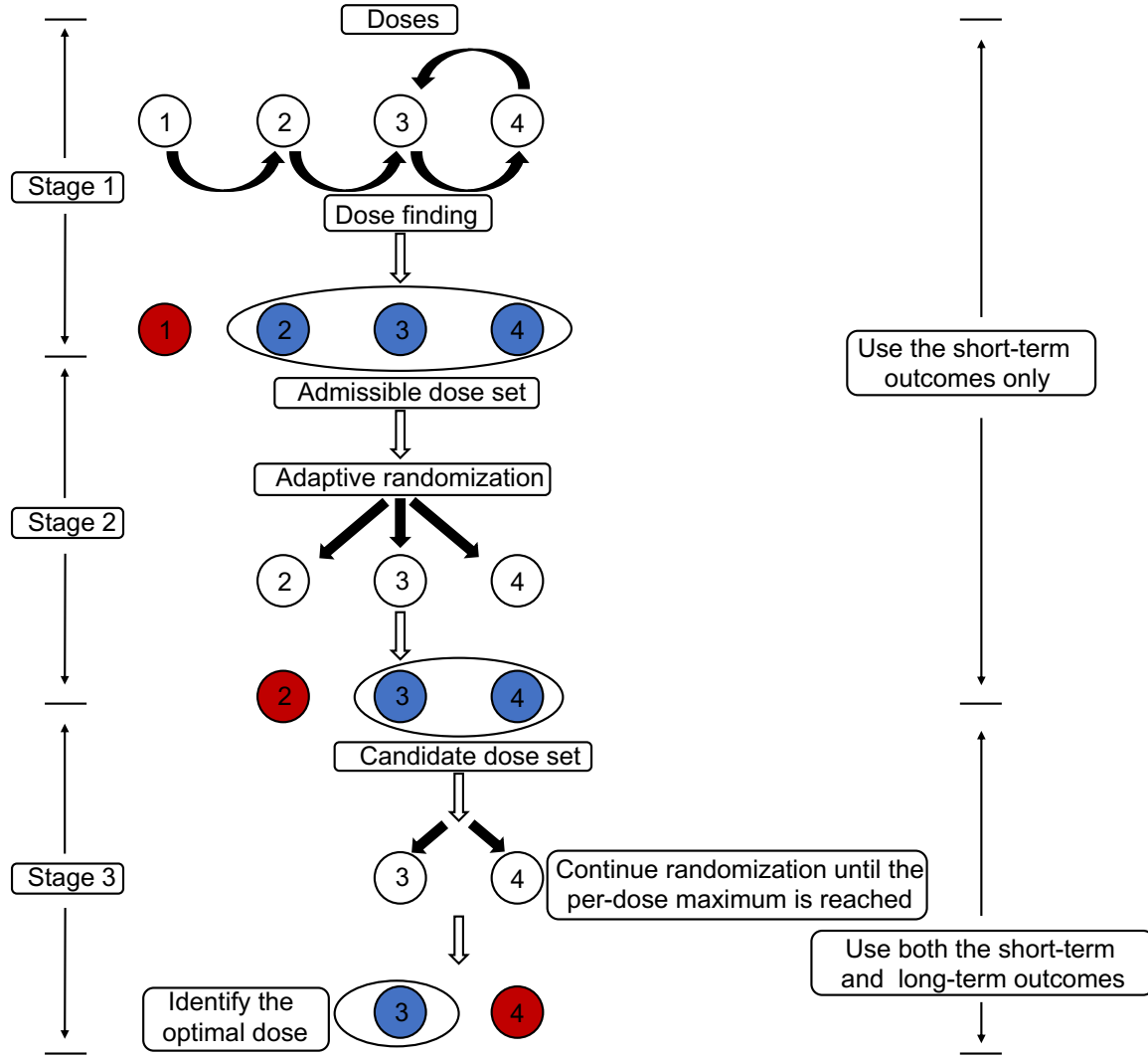


Figure 3: Dose-outcome curves for the scenarios in the simulation study. The red, green, and blue curves are  $\pi_T^{true}(d_j)$ ,  $\pi_R^{true}(d_j)$ , and  $\xi^{true}(d_j) = \Pr^{true}(Z > t_2 - t_1 \mid Z > 0, d_j)$ , respectively. The horizontal lines show the fixed upper limit .30 for  $\pi_T(d_j)$  and fixed lower limit .50 for  $\pi_R(d_j)$  in the dose admissibility rules, and the fixed lower limit 0.40 for  $\xi(d_j)$  for long-term success probability.

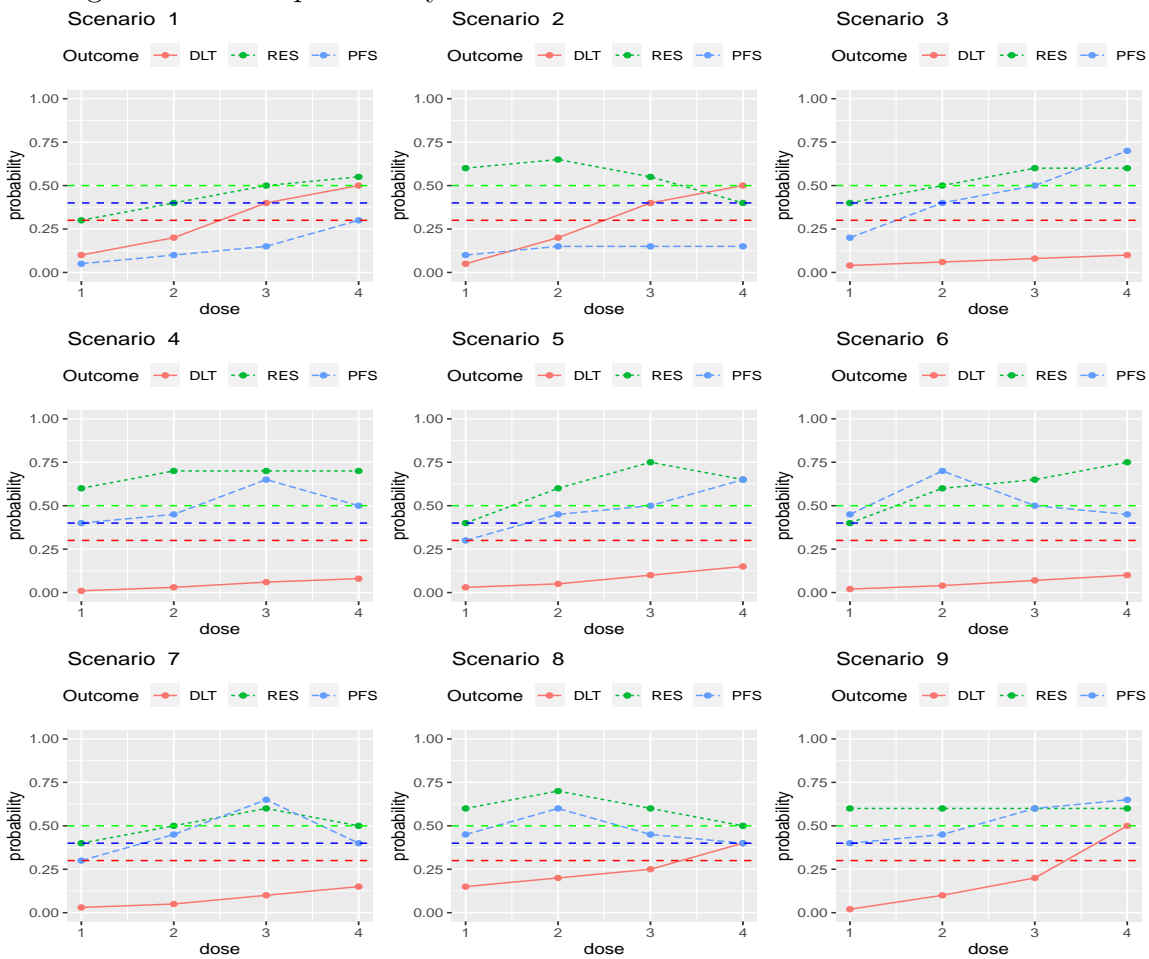


Table 2: Selection % and mean number of patients treated at each dose level, mean sample size, and  $R = \xi(d^{sel,\xi})/\xi(d^{opt,\xi})$  under the Gen I-II, Conv 1 and Conv 2 designs. The true long term success probability at  $d_j$  is  $\xi^{true}(d_j) = \Pr^{true}(Z > 5 \mid Y'_R > 0, d_j)$ . Boldface indicates results for the true optimal decision. The numbers in the brackets indicate the Monte Carlo simulation standard errors.

Designs	Dose levels					Sample size	R %
	0*	1	2	3	4		
<i>Scenario 1</i>							
	$\pi_T^{true}(d_j)$	0.10	0.20	0.40	0.50		
	$\pi_R^{true}(d_j)$	0.30	0.40	0.50	0.55		
	$\bar{U}^{true}(d_j)$	59.1	62.5	60.9	59.9		
	$\xi^{true}(d_j)$	0.05	0.1	0.15	0.3		
Gen I-II	Selection %	<b>93.5(0.3)</b>	0	0.50	2.6	3.3	35.6 NA
	Patients		9.5	13.5	8.8	3.7	
Conv 1	Selection %	<b>56.2(0.7)</b>	6.5	22.6	12.4	2.2	34.9 NA
	Patients		9.4	13.2	8.7	3.6	
Conv 2	Selection %	<b>57.7(0.7)</b>	5.6	22.1	12.4	2.2	35.4 NA
	Patients		9.4	13.4	8.8	3.7	
<i>Scenario 2</i>							
	$\pi_T^{true}(d_j)$	0.05	0.20	0.40	0.50		
	$\pi_R^{true}(d_j)$	0.60	0.65	0.55	0.40		
	$\bar{U}^{true}(d_j)$	76.7	68.0	54.0	40.4		
	$\xi^{true}(d_j)$	0.1	0.15	0.15	0.15		
Gen I-II	Selection %	<b>83.8(0.5)</b>	2.3	9.6	4.1	0.2	48.2 NA
	Patients		21.5	16.4	7.4	2.8	
Conv 1	Selection %	<b>3.2(0.2)</b>	73.3	22.4	1.1	0	46.9 NA
	Patients		21.2	15.9	7.0	2.8	
Conv 2	Selection %	<b>3.1(0.2)</b>	73.2	22.8	0.9	0	48.2 NA
	Patients		21.4	16.6	7.3	2.9	
<i>Scenario 3</i>							
	$\pi_T^{true}(d_j)$	0.04	0.06	0.08	<b>0.10</b>		
	$\pi_R^{true}(d_j)$	0.40	0.50	0.60	<b>0.60</b>		
	$\bar{U}^{true}(d_j)$	61.2	67.0	74.2	<b>75.0</b>		
	$\xi^{true}(d_j)$	0.20	0.40	0.50	<b>0.70</b>		
Gen I-II	Selection %	2.9	0.3	7.0	19.3	<b>70.5(0.6)</b>	52.7 91.0
	Patients		10.6	13.0	14.9	<b>14.3</b>	
Conv 1	Selection %	2.2	4.6	13.5	39.7	<b>39.9(0.7)</b>	47.4 79.1
	Patients		9.6	11.7	13.3	<b>12.7</b>	
Conv 2	Selection %	2.2	3.6	13.2	40.2	<b>40.8(0.7)</b>	52.8 79.9
	Patients		10.5	13.1	14.9	<b>14.3</b>	

\*The number under  $d = 0$  is the percentage of trials terminated early with no dose is selected.

Table 2 (continued)

Designs	Dose levels					Sample size	R %
	0	1	2	3	4		
<i>Scenario 4</i>							
	$\pi_T^{true}(d_j)$		0.01	0.03	<b>0.06</b>	0.08	
	$\pi_R^{true}(d_j)$		0.6	0.7	<b>0.70</b>	0.70	
	$\bar{U}^{true}(d_j)$		72.1	80.9	<b>81.3</b>	80.6	
	$\xi^{true}(d_j)$		0.4	0.45	<b>0.65</b>	0.50	
Gen I-II	Selection %	0.1	4.3	9.2	<b>69.1(0.7)</b>	17.2	58.1
	Patients		13.9	15.1	<b>14.8</b>	14.4	
Conv 1	Selection %	0	8.7	33.0	<b>31.1(0.7)</b>	27.7	48.0
	Patients		11.5	12.7	<b>12.1</b>	11.6	
Conv 2	Selection %	0.1	7.0	33.2	<b>32.3(0.7)</b>	27.3	58.1
	Patients		13.8	15.3	<b>14.8</b>	14.1	
<i>Scenario 5</i>							
	$\pi_T^{true}(d_j)$		0.03	0.05	0.1	<b>0.15</b>	
	$\pi_R^{true}(d_j)$		0.4	0.6	0.75	<b>0.65</b>	
	$\bar{U}^{true}(d_j)$		63.1	75.2	82.3	<b>72.7</b>	
	$\xi^{true}(d_j)$		0.3	0.45	0.5	<b>0.65</b>	
Gen I-II	Selection %	1.2	1.4	14.2	24.1	<b>59.1(0.7)</b>	54.3
	Patients		10.5	14.7	15.8	<b>13.3</b>	
Conv 1	Selection %	1.0	2.2	22.8	61.3	<b>12.8(0.5)</b>	47.7
	Patients		9.1	13.0	14.1	<b>11.4</b>	
Conv 2	Selection %	1.1	1.6	20.1	64.6	<b>12.6(0.5)</b>	54.3
	Patients		10.2	14.9	16.2	<b>13.1</b>	
<i>Scenario 6</i>							
	$\pi_T^{true}(d_j)$		0.02	<b>0.04</b>	0.07	0.10	
	$\pi_R^{true}(d_j)$		0.4	<b>0.60</b>	0.65	0.75	
	$\bar{U}^{true}(d_j)$		54.4	<b>72.6</b>	72.5	77.8	
	$\xi^{true}(d_j)$		0.45	<b>0.70</b>	0.5	0.45	
Gen I-II	Selection %	0.6	5.4	<b>68.4(0.7)</b>	15.7	9.8	54.3
	Patients		10.0	<b>14.4</b>	14.7	15.2	
Conv 1	Selection %	0.9	1.4	<b>25.1(0.6)</b>	25.3	47.2	47.7
	Patients		8.8	<b>12.9</b>	12.8	13.2	
Conv 2	Selection %	1.1	1.0	<b>23.9(0.6)</b>	23.6	50.5	54.2
	Patients		9.9	<b>14.6</b>	14.5	15.1	

Table 2 (continued)

Designs	Dose levels					Sample size	R %
	0	1	2	3	4		
<i>Scenario 7</i>							
	$\pi_T^{true}(d_j)$		0.03	0.05	<b>0.1</b>	0.15	
	$\pi_R^{true}(d_j)$		0.4	0.5	<b>0.6</b>	0.50	
	$\bar{U}^{true}(d_j)$		66.1	70.3	<b>73.5</b>	68.6	
	$\xi^{true}(d_j)$		0.3	0.45	<b>0.65</b>	0.4	
Gen I-II	Selection %	3.3	2.0	15.0	<b>70.9(0.6)</b>	8.8	52.2
	Patients		11.2	13.7	<b>15.1</b>	12.1	
Conv 1	Selection %	2.8	11.4	26.2	<b>42.9(0.7)</b>	16.6	47.4
	Patients		10.1	12.5	<b>14.0</b>	10.7	
Conv 2	Selection %	2.8	10.2	26.2	<b>44.6(0.7)</b>	16.3	52.2
	Patients		11.1	13.9	<b>15.3</b>	11.8	
<i>Scenario 8</i>							
	$\pi_T^{true}(d_j)$		0.15	<b>0.20</b>	0.25	0.40	
	$\pi_R^{true}(d_j)$		0.6	<b>0.70</b>	0.6	0.50	
	$\bar{U}^{true}(d_j)$		71.8	<b>77.8</b>	68.5	56.4	
	$\xi^{true}(d_j)$		0.45	<b>0.60</b>	0.45	0.40	
Gen I-II	Selection %	9.1	23.4	<b>56.9(0.7)</b>	9.0	1.6	47.7
	Patients		17.7	<b>14.8</b>	9.9	5.30	
Conv 1	Selection %	9.1	29.3	<b>52.9(0.7)</b>	8.2	0.5	44.2
	Patients		16.5	<b>14.1</b>	9.0	4.60	
Conv 2	Selection %	9.0	28.3	<b>54.0(0.7)</b>	8.4	0.3	47.7
	Patients		17.4	<b>15.2</b>	10.0	5.10	
<i>Scenario 9</i>							
	$\pi_T^{true}(d_j)$		0.02	0.1	<b>0.2</b>	0.5	
	$\pi_R^{true}(d_j)$		0.6	0.6	<b>0.6</b>	0.6	
	$\bar{U}^{true}(d_j)$		76.3	73.5	<b>70.2</b>	60.4	
	$\xi^{true}(d_j)$		0.4	0.45	<b>0.6</b>	0.65	
Gen I-II	Selection %	1.3	18.2	22.5	<b>51.2(0.7)</b>	6.8	51.6
	Patients		17.2	15.6	<b>12.8</b>	5.9	
Conv 1	Selection %	0.9	50.6	31.0	<b>16.5(0.5)</b>	1.0	47.7
	Patients		16.1	14.4	<b>11.9</b>	5.2	
Conv 2	Selection %	1.1	49.6	31.8	<b>16.9(0.5)</b>	0.6	51.5
	Patients		17.2	15.4	<b>13.1</b>	5.7	

## **Supporting Information**

The technical detail and sensitivity analysis simulation results are available with this paper online.

## **Data availability statement**

Data sharing is not applicable to this paper as all data in this article are computer simulated.

## **Acknowledgments**

The authors thank an associate editor and two referees for their detailed and constructive comments. Peter Thall's research was supported by NIH/NCI grants P30 CA016672, P01 CA148600, and R01CA261978. Yong Zang's research was partially supported by NIH/NCI grants P30 CA082709; R21 CA264257 and the Ralph W. and Grace M. Showalter Research Trust award. Ying Yuan's research was partially supported by by NIH/NCI grants P50 CA098258, P50 CA217685, and P50 CA221707.