A Generalized Phase 1-2-3 Design integrating Dose Optimization with Confirmatory Treatment Comparison

Yong Zang

Department of Biostatistics and Health Data Science;

Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, IN

Peter F. Thall

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

Ying Yuan

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

Correspondence to: Yong Zang, Ph.D., Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, IN (email: zangy@iu.edu); Ying Yuan, Ph.D. (yyuan@mdanderson.org) and Peter Thall, Ph.D. (rex@mdanderson.org), Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX.

Abstract

A generalized phase 1-2-3 design, Gen 1-2-3, that includes all phases of clinical treatment evaluation is proposed. The design extends and modifies the design of Chapple and Thall (2019), denoted as CT. Both designs begin with a phase 1-2 trial including

dose acceptability and optimality criteria, and both select an optimal dose for phase 3. The Gen 1-2-3 design has the following key differences. It uses phase 1-2 criteria to identify a set of candidate doses rather than one dose. In an intermediate stage between phase 1-2 and phase 3, it randomizes additional patients fairly among the candidate doses and an active control treatment arm and uses survival time data to select an optimal dose. It makes a Go / No Go decision of whether or not to conduct phase 3 based on a predictive probability that the optimal dose will provide a substantive improvement over the control in survival time. A simulation study shows that the Gen 1-2-3 design has desirable operating characteristics compared to the CT design and two conventional designs.

Keywords: Bayesian Design; Cell Therapy; Dose Finding; Phase 1-2 Clinical Trial; Phase 1-2-3 Clinical Trial

1 Introduction

Conventionally, clinical evaluation of a potential new anti-disease agent, X, begins by using early toxicity, Y_T, and possibly an early efficacy outcome, Y_E, to select an optimal dose, $\hat{d}^{opt}$. This may be a maximum tolerated dose (MTD) based on Y_T in a phase 1 trial (Storer, 1989, 2001; O'Quigley et al., 1990; Babb et al., 1998; Liu and Yuan, 2015), or a dose based on (Y_E, Y_T) and possibly PK/PD data from a phase 2 or phase 1-2 trial (Ratain, 2014; Zang et al., 2014; Yuan et al., 2016; Yan et al., 2018; Guo and Yuan, 2023). Denoting X administered at dose d by $X(d)$, once $\hat{d}^{opt}$ is chosen, a confirmatory randomized phase 3 trial may be conducted to compare $X(\hat{d}^{opt})$ to a control treatment, C, based on survival or progression-free survival (PFS) time, Y_S. In some settings, once $\hat{d}^{opt}$ is chosen, a Go / No Go decision of whether or not to conduct phase 3 is made by using currently available data to decide whether $X(\hat{d}^{opt})$ is promising.

The convention of first evaluating safety and selecting a dose based on early outcomes,

observed much sooner than Y_S, often with small sample sizes, is motivated by the desire to complete the dose optimization process quickly. Unfortunately, this general strategy has several very undesirable properties. If early phase sample sizes are too small to obtain reliable inferences, then it is not unlikely that a selected dose $\hat{d}^{opt}$ later will be found to be unsafe, ineffective, or both, based on phase 3 or post approval clinical practice data (Shah et al., 2021). If Y_E is ignored, or if (Y_E, Y_T) are used for dose-finding but Y_E has a weak connection to Y_S, then there is a high risk of selecting a dose that is suboptimal in terms of Y_S (Yuan et al., 2016; Yan et al., 2018; Brock et al., 2021; Thall et al., 2023).

Basing inferences and decisions on early outcomes without accounting for Y_S in a clinical trial may lead to poor decisions. This occurred during a randomized trial to compare two preparative regimens, busulfan plus melphalan $(B + M)$ and melphalan alone (M), for autologous stem cell transplantation in multiple myeloma, based on response as the primary outcome. An interim analysis showed that 6/44 (13.6 %) of $B + M$ patients had responses versus 13/32 (40.6%) of M patients, and a futility monitoring rule stopped the trial early. In contrast, the estimated 12-month PFS probabilities were .90 for $B + M$ and .77 for M, and this superiority of $B + M$ over M in terms of PFS persisted after accounting for prognostic covariate effects in a regression analysis (unpublished). The trial was re-designed and completed using PFS time as the primary outcome (Bashir et al., 2019). This trial illustrates potential consequences of the general fact that, while early response may be associated with a better long term outcome, Y_E is not a surrogate for Y_S.

Many hybrid designs have been proposed to improve the reliability and efficiency of the treatment evaluation process. Phase 1-2 designs select a dose or schedule based on both Y_E and Y_T (Thall and Russell, 1998; Braun, 2002; Thall and Cook, 2004; Zhang et al., 2006; Guo and Yuan, 2017; Liu et al., 2018; Yuan et al., 2016; Zhou et al., 2019; Lee et al., 2020; Lin et al., 2020; Zhang et al., 2021). Thall et al. (2023) proposed a generalized phase 1-2 design, Gen 1-2, that optimizes dose based on both (Y_E, Y_T) and remission duration. Many

phase 2-3 designs have been proposed that combine treatment screening with confirmatory evaluation (Thall et al., 1988; Schaid et al., 1990; Inoue et al., 2002; Stallard and Todd, 2003; Korn et al., 2012).

Chapple and Thall (2019) (CT) proposed a phase 1-2-3 design paradigm that includes the entire clinical treatment evaluation process. A phase 1-2-3 design begins by applying a phase 1-2 design based on (Y_E, Y_T), including rules to screen out unsafe or ineffective doses, and adaptive randomization (AR) to reduce the chance of getting stuck at a sub-optimal dose. A best acceptable dose, $\hat{d}_{ET}^{\,opt}$, is selected based on (Y_E, Y_T, d) data, and a group sequential (GS) phase 3 trial based on Y_S then is begun with patients randomized fairly between $X(\hat{d}_{ET}^{\,opt})$ and C. At the first phase 3 GS decision, a re-optimized dose, $\hat{d}_S^{\,opt}$, is selected to maximize estimated mean survival time of $X(d)$ among all acceptable doses. The GS phase 3 trial then is completed with patients randomized between $X(\hat{d}_S^{\,opt})$ and C. CT reported computer simulations showing that this design is greatly superior to the convention of conducting phase 1-2 and using $X(\hat{d}_{ET}^{\,opt})$ in phase 3 without re-optimizing dose.

In this paper, we propose a generalized phase 1-2-3 design, which we call *Gen 1-2-3*, that modifies and extends the CT design. Like the CT design, the Gen 1-2-3 design begins by using a Bayesian phase 1-2 design based on (Y_E, Y_T), and later selects a best dose $\hat{d}^{opt}$ based on survival time. A Gen 1-2-3 design may be considered a hybrid of a Gen 1-2 design and a phase 3 design. The Gen 1-2-3 design has the following key differences from a CT design. A Gen 1-2-3 trial includes either two or three stages. In stage 1, a phase 1-2 design's dose acceptability and optimality criteria based on (Y_E, Y_T, d) are used to assign patients to doses and identify a set of candidate doses, $\mathcal{A}^{can}$, rather than one dose. Similar strategy was considered by Guo and Yuan (2023), who referred to $\mathcal{A}^{can}$ as the recommended phase 2 dose set (RP2S). In stage 2, patients are randomized fairly among $\{X(d):\ d \in \mathcal{A}^{can}\}$ and C and followed to obtain their survival time data. At the end of stage 2, an optimal dose, $\hat{d}^{opt}$, is selected from $\mathcal{A}^{can}$ to maximize the survival rate. A Go / No Go decision of whether or

not to conduct a phase 3 trial then is made, based on the predictive probability (PP) that the hazard ratio of $X(\hat{d}^{opt})$ versus C, computed using simulated future phase 3 data, will be below a fixed threshold. The PP quantifies how promising $X(\hat{d}^{opt})$ is compared to C. If the decision is "Go", stage 3 consists of a phase 3 trial of $X(\hat{d}^{opt})$ versus C. If it is "No Go", stage 3 is not conducted and it is concluded that $X(\hat{d}^{opt})$ does not provide an improvement over C.

The Gen 1-2-3 design includes two key screening parameters. A *proximity parameter*, $\rho \in [0,\ 1]$, determines whether a dose is close enough to being optimal so that it is included in $\mathcal{A}^{can}$, and a *Go - No Go parameter*, $p_U(Go) \in [0,\ 1]$, determines how large the estimated PP that $X(\hat{d}^{opt})$ is promising compared to C must be in order to conduct phase 3. A CT design may be obtained as a Gen 1-2-3 design by setting $\rho = 1$ to ensure that $\mathcal{A}^{can}$ includes exactly one dose, setting $p_U(Go) = 0$ to ensure that phase 3 always is begun, and later re-setting $\rho = 0$ at the first GS decision of phase 3 to allow dose re-optimization.

Section 2 presents details of the Gen 1-2-3 design's stages and decision criteria. Section 3 illustrates the design using the Bayesian utility-based phase 1-2 optimal interval design, BOIN12 (Lin et al., 2020), assuming that $p(Y_S \mid Y_E, Y_T, d)$ is a Weibull distribution. A simulation study is presented in Section 4, including particular versions of the Gen 1-2-3 and CT designs that may be considered comparable, and two conventional designs that only use (Y_E, Y_T) for dose selection. A brief discussion is given in section 5.

2 A Generalized Phase 1-2-3 Design Paradigm

2.1 Preliminaries

A Gen 1-2-3 design requires assumed regression models $p(Y_E, Y_T \mid d, \boldsymbol{\theta}_{ET})$ and $p(Y_S \mid Y_E, Y_T, d, \boldsymbol{\theta}_S)$, where $\boldsymbol{\theta}_{ET}$ and $\boldsymbol{\theta}_S$ are parameter vectors, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{ET}, \boldsymbol{\theta}_S)$. Like the CT proposal, Gen 1-2-3 is a paradigm for constructing designs, since its component phase 1-2

and phase 3 designs, regression models, and decision criteria may be tailored to accommodate the application at hand. For example, if the goal is to explore combinations of dose and administration schedule for X, then the stages and decision rules are as described here, but with the regression models suitably modified to include (dose, schedule) effects rather than only doses. We first present a general form for the Gen 1-2-3 design paradigm, and in section **3.2** we illustrate it with a particular design that will be used in the simulations.

Denote the time to treatment failure or independent right censoring by Y_S^o, with $\delta = I(Y_S^o = Y_S)$. For each treatment $\tau = X(d)$ or C, denote $m = ET$ for (Y_T, Y_E, τ) data and $m = S$ for (Y_S^o, δ, τ) data. Index the design's stages by $k = 1, 2, 3$, and let n_k denote the maximum sample size at stage k, with overall maximum sample size $N = n_1 + n_2 + n_3$. To keep track of datasets by outcome type and stage, let $\mathcal{D}_{m,k}$ denote the data for outcomes m = ET or S at the end of stage k=1, 2, or 3. To keep track of data in terms of patients enrolled during stages 1 and 2, let $\mathcal{D}_{ET}^n \subseteq \mathcal{D}_{ET,1} \cup \mathcal{D}_{ET,2}$ denote the ET data from the first n patients for $n = 1, \cdots, n_{12} = n_1 + n_2$. Denote the doses of X to be studied by $\mathbf{d} = \{d_1, \cdots, d_J\}$, and denote $d = d_0$ for C. As noted above, $\mathbf{d}$ may instead correspond to a set of (dose, schedule) combinations of X to be evaluated, and the paradigm easily accommodates such settings, with appropriate modifications of the regression models.

Following the Gen 1-2 paradigm, the main requirements for the phase 1-2 design used by a Gen 1-2-3 design are that (i) Y_E and Y_T are discrete ordinal variables observed relatively soon after initiation of treatment; (ii) decisions rely on a dose optimality criterion, $\phi(d, \boldsymbol{\theta})$, defined in terms of (Y_E, Y_T); and (iii) the following two dose acceptability conditions are imposed throughout. For fixed lower limit $\underline{\pi}_E$ on $\pi_E(d, \boldsymbol{\theta})$ and fixed upper limit $\overline{\pi}_T$ on $\pi_T(d, \boldsymbol{\theta})$, given current data $\mathcal{D}_{ET}^n$, *a dose d is acceptable* if

$$\Pr\{\pi_E(d, \boldsymbol{\theta}_{ET}) > \underline{\pi}_E \mid \mathcal{D}_{ET}^n\} > .10 \quad \text{and} \quad \Pr\{\pi_T(d, \boldsymbol{\theta}_{ET}) < \overline{\pi}_T \mid \mathcal{D}_{ET}^n\} > .10. \tag{1}$$

Cutoffs other than .10 may be used in (1), with values in the range .05 to .20 generally working well. For each sample size $n \leq n_{12}$, let $\mathcal{A}_n$ denote the set of acceptable doses determined by (1). For each $d \in \mathcal{A}_n$, the posterior mean dose optimality criterion is denoted by $\hat{\phi}_n(d) = E\{\phi(d, \boldsymbol{\theta}_{ET}) \mid \mathcal{D}_{ET}^n\}$.

The regression model for the distribution of Y_S is formulated to borrow strength from regression of Y_S on Y_E, Y_T, and d. Denote the joint probability $\pi_{a,b}(d, \boldsymbol{\theta}) = \Pr(Y_{\mathrm{E}} = a, Y_{\mathrm{T}} = b \mid d, \boldsymbol{\theta}_{ET})$ for all possible values (a, b) of (Y_E, Y_T). Recalling that $d = d_0$ represents C, the pdf of Y_S as a function of d is the mixture

$$f_S(y_S \mid d, \boldsymbol{\theta}) = \sum_{a,b} f_{S|E,T}(y_S \mid Y_E = a, Y_T = b, d, \boldsymbol{\theta}_S)\ \pi_{a,b}(d, \boldsymbol{\theta}_{ET}). \tag{2}$$

We define the survival function at t as $\overline{F}_{\mathrm{S}}(t \mid X(d), \boldsymbol{\theta}) = \Pr\{Y_S > t \mid X(d), \boldsymbol{\theta}\}$.

2.2 Stages of a Generalized Phase 1-2-3 Design

If the early outcomes $(Y_{\mathrm{E}}, Y_{\mathrm{T}})$ are discrete ordinal variables then, for example, the phase 1-2 optimality criterion may be an expected utility $\phi(d, \boldsymbol{\theta}) = \sum_{a,b} U(y_E, y_T)\pi_{a,b}(d, \boldsymbol{\theta})$. If either Y_E or Y_T has three or more possible values, then it is necessary to define binary versions of these variables to that $\pi_E(d, \boldsymbol{\theta}_{ET})$ and $\pi_T(d, \boldsymbol{\theta}_{ET})$ may be defined in order to specify the dose admissibility criteria (1).

To construct a Go / No Go rule, it will be necessary to consider the future failure time data that would be available upon completion of a phase 3 trial, if it were conducted. This is $\mathcal{D}_{S,3}^{future} = \{(Y_{S,i}^o, \delta_i, \tau_i) :\ i = n_{12} + 1, \cdots, N\}$, where τ_i denotes the i^{th} phase 3 patient's treatment, which is either $X(\hat{d}^{opt})$ or C. At the end of stage 2, since $\mathcal{D}_{S,3}^{future}$ consists of potential outcomes that may or may not be observed depending on whether or not phase 3 is conducted, a predictive probability is used as the criterion for making the Go / No Go decision. Let t_S^* be the maximum follow up time for observing (Y_S^o, δ). As depicted in Figure

1, the design's stages are as follows.

Stage 1. Use a phase 1-2 design based on $(Y_{\rm E}, Y_{\rm T})$ and the criterion function $\phi(d, \boldsymbol{\theta})$ to do sequentially adaptive dose-finding, subject to the dose acceptability rules (1). When n_1 patients have been treated and evaluated, identify the RP2S, $\mathcal{A}_{n_1}^{can} \subseteq \mathbf{d}$, computed from the data $\mathcal{D}_{ET,1}$, as follows. Denoting the estimated criterion function of the empirically best acceptable dose by $\hat{\phi}_{n_1}^{max} = max\{\hat{\phi}_{n_1}(d) : \ d \in \mathcal{A}_{n_1}\}$, the RP2S is

$$\mathcal{A}_{n_1}^{can} = \{d \in \mathcal{A}_{n_1} : \hat{\phi}_{n_1}(d) \geq \rho \ \hat{\phi}_{n_1}^{max}\}.$$

For example, if $\rho = .70$, then any acceptable dose with estimated optimality criterion at least 70% of the maximum value is a candidate.

Stage 2. For each $n = n_1 + 1, \cdots, n_{12}$, let $\mathcal{A}_n^{can}$ denote the current RP2S. When n_2 patients have been randomized and treated, they are followed for an additional time $L \leq t_S^*$, to harvest survival time data for use in later treatment evaluation. The following two decisions are made at the end of stage 2.

Selection of a Best Candidate Dose. Under the mixture model (2), the survival function $\overline{F}_{\rm S}(t \mid X(d), \boldsymbol{\theta})$ evaluated at follow up time t_S^* is used to select an optimal dose from $\mathcal{A}_{n_{12}}^{can}$. Given the data at the end of stage 2, the *optimal dose* $\hat{d}^{opt}$ is defined as the dose in $\mathcal{A}_{n_{12}}^{can}$ with the largest posterior probability of having the largest $\overline{F}_{\rm S}(t_S^* \mid X(d), \boldsymbol{\theta})$, formally,

$$\hat{d}^{opt} = \underset{d \in \mathcal{A}_{n_{12}}^{can}}{\rm argmax} \ \Pr\left\{ \overline{F}_{\rm S}(t_S^* \mid d, \boldsymbol{\theta}) = \underset{d' \in \mathcal{A}_{n_{12}}^{can}}{max} \overline{F}_{\rm S}(t^* \mid d', \boldsymbol{\theta}) \mid \mathcal{D}_{ET,1} \cup \mathcal{D}_{ET,2} \cup \mathcal{D}_{S,2} \right\}. \qquad (3)$$

If desired, $\hat{d}^{opt}$ may be selected using other criteria, such as the posterior mean of $\overline{F}_{\rm S}(t^* \mid d, \boldsymbol{\theta})$.

Go / No Go Decision Based on a Predictive Probability. After selecting d^{opt}, the Go / No Go decision of whether proceed to stage 3 (i.e., phase 3) to further evaluate its

long-term efficacy Y_S will be made based on the PP of the hazard ratio (HR) of $X(d)$ to C, denoted by $\lambda(d)$. Denote the future phase 3 data that would become available at the end of phase 3, if it were conducted, by $\mathcal{D}_{S,3}^{future}$. A parameter $\underline{\lambda} \in (0,1)$ must be specified to determine whether the HR of $X(\hat{d}^{opt})$ versus C is small enough to infer that $X(\hat{d}^{opt})$ provides a meaningful improvement in survival. At the end of stage 2, a Bayesian criterion that quantifyies how superior $X(\hat{d}^{opt})$ will be compared to C at the end of phase 3 is

$$\xi(\mathcal{D}_{S,3}^{future}) =_{def} \Pr\{\lambda(\hat{d}^{opt}) \leq \underline{\lambda} \mid \mathcal{D}_{S,2}, \mathcal{D}_{S,3}^{future}\}. \tag{4}$$

Expression (4) is the posterior probability, given all future randomized data available at the end of phase 3, that the HR of $X(\hat{d}^{opt})$ versus C is smaller than $\underline{\lambda}$. In practice, the HR cutoff $\underline{\lambda}$ might be chosen from the range .50 to .90. Define the phase 3 success indicator $\Xi(\mathcal{D}_{S,3}^{future})$ $= 1$ if $\xi(\mathcal{D}_{S,3}^{future}) > p_U$ and 0 otherwise, where p_U may be chosen from the range .50 to .99. If $\Xi(\mathcal{D}_{S,3}^{future}) = 1$, this says that, based on all randomized data from stage 2 and a future phase 3 trial, it is likely that the HR of $X(\hat{d}^{opt})$ compared to C is desirably small.

Since the future data $\mathcal{D}_{S,3}^{future}$ are not available at the end of stage 2, the Go / No Go decision is based on the predictive distribution of $\mathcal{D}_{S,3}^{future}$ given the observed stage 2 randomized data, which we define as

$$p(\mathcal{D}_{S,3}^{future} \mid \mathcal{D}_{ET,2}, \mathcal{D}_{S,2}) = \int_{\boldsymbol{\theta}} p\{\mathcal{D}_{S,3}^{future} \mid \boldsymbol{\theta}\} p(\boldsymbol{\theta} \mid \mathcal{D}_{ET,2}, \mathcal{D}_{S,2}) d\boldsymbol{\theta}. \tag{5}$$

This PP may be computed in the following steps. First, simulate a large sample of parameters $\{\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(B)}\}$ from the posterior $p(\boldsymbol{\theta} \mid \mathcal{D}_{ET,2}, \mathcal{D}_{S,2})$. For each $\boldsymbol{\theta}^{(b)}$, simulate a future phase 3 dataset, $\mathcal{D}_{S,3}^{future,(b)}$ using the mixture model (2). Evaluate the phase 3 success indicator $\Xi(\mathcal{D}_{S,3}^{future,(b)})$ using a proportional hazard model $f_{S,PH}(y_S \mid \tau, \lambda(\hat{d}^{opt}))$ for $\tau = X(\hat{d}^{opt})$ or C. In our simulation, we assumed an independent exponential-gramma model for $X(\hat{d}^{opt})$ and C to obtain the posterior of $\lambda(\hat{d}^{opt})$. This approach is highly efficient in

terms of computation, and previous research (Thall et al., 2005; Zhou et al., 2020), as well as our simulation described later, has demonstrated its remarkable robustness in facilitating Go/No-Go decisions. Nevertheless, other more sophisticated models, such as the piecewise exponential model (Cai et al., 2014), can be entertained when desirable.

Repeating this for $b = 1, \cdots, B$, the estimated PP of phase 3 success is the mean

$$\widehat{PP} = \frac{1}{B} \sum_{b=1}^{B} \Xi(\mathcal{D}_{S,3}^{future,(b)}). \tag{6}$$

The Go / No Go decision is to conduct phase 3 if $\widehat{PP} > p_U(Go)$. If the decision is No Go, then do not conduct phase 3, and stop the treatment development process, with the conclusion that $X(\hat{d}^{opt})$ is not promising compared to C. If the decision is Go, then continue to stage 3.

Stage 3. Using the GS design, conduct phase 3 to test the difference between the survival distributions of $X(\hat{d}^{opt})$ and C. Each patient is followed for up to t_S^* months to harvest the (Y_S^o, δ, τ) data, for $\tau = X(\hat{d}^{opt})$ or C. The toxicity rate for $X(\hat{d}^{opt})$ is stilled monitored in stage 3 with the toxicity acceptable rule in (1), using the toxicity data accumulated from stages 1 to 3.

3 Illustration of the Gen 1-2-3 Design

3.1 Dose-outcome models

In this section, we illustrate how a Gen 1-2-3 design may be constructed by describing a particular case. We assume binary Y_T and Y_E and denote $\pi_{a,b}(d) = \Pr(Y_{\rm E} = a, Y_{\rm T} = b \mid d)$ for $a, b \in \{0, 1\}$ and dose d, recalling that $d = d_0$ represents C. Rather than formulating a dose-response model, for each d we define the phase 1-2 parameter vector as the probabilities

of the elementary events, $\boldsymbol{\theta}_{ET}(d) = (\pi_{0,0}(d), \pi_{0,1}(d), \pi_{1,0}(d))$, with $\pi_{1,1}(d) = 1 - \{\pi_{0,0}(d) + \pi_{0,1}(d) + \pi_{1,0}(d)\}$. This model does not borrow strength between doses, but is very tractable, as well as robust as it does not make any parametric assumption on dose-toxicity and -efficacy curves. Given sample size n and dose d, denote the elementary event counts

$$V_{a,b,n}(d) = \sum_{i=1}^{n} I(Y_{E,i} = a, Y_{T,i} = b, \tau_i = d)$$

for $a, b \in \{0, 1\}$, and denote the vector of counts by $\boldsymbol{V}_n(d)$. For each d, denoting $n(d) = \sum_{i=1}^{n} I(\tau_i = d)$, $\boldsymbol{V}_n(d) \sim \text{multinomial}(n(d), \boldsymbol{\theta}_{ET}(d))$. We denote the concatenated vectors as $\boldsymbol{V}_n = (\boldsymbol{V}_n(0), \boldsymbol{V}_n(1), \cdots, \boldsymbol{V}_n(J))$ and $\boldsymbol{\theta}_{ET} = (\boldsymbol{\theta}_{ET}(0), \boldsymbol{\theta}_{ET}(1), \cdots, \boldsymbol{\theta}_{ET}(J))$.

We assume that the failure time is Weibull with conditional hazard function

$$h_S(t \mid Y_{\mathrm{E}}, Y_{\mathrm{T}}, d, \theta_{\mathrm{S}}) = \Big(\frac{\gamma}{\psi}\Big)\Big(\frac{t}{\psi}\Big)^{\gamma-1} \exp\Big\{\beta_1 Y_{\mathrm{E}} + \beta_2 Y_{\mathrm{T}} + \sum_{j=0}^{J} \beta_{3,j} \mathrm{I}(d = d_j)\Big\}, \quad t > 0, \tag{7}$$

setting $\beta_{3,0} = 0$ for C, so $\boldsymbol{\theta}_S = (\gamma, \psi, \beta_1, \beta_2, \beta_{3,1}, \cdots, \beta_{3,J})$. For sample size $n \leq n_{12}$ in stage 1 or 2, denote the early outcome data by $\mathcal{D}_{ET}^{n} = \big\{(Y_{\mathrm{E},i}, Y_{\mathrm{T},i}, d_{[i]}); i = 1, \cdots, n\big\}$, and denote the randomized time-to-event data from stage 2 by $\mathcal{D}_{\mathrm{S},2} = \big\{(Y_{\mathrm{S},i}^{o}, \delta_i, d_{[i]}); i = n_1 + 1, \cdots, n_{12}\big\}$. Denoting the Weibull pdf by f_{S} and survivor function by $\overline{F}_{\mathrm{S}}$, the likelihood for the data at the end of stage 2 is

$$\begin{aligned} \mathcal{L}(\mathcal{D}_{ET}^{n_{12}}, \mathcal{D}_{\mathrm{S},2} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n_{12}} \prod_{a=0}^{1} \prod_{b=0}^{1} \{\pi_{a,b}(d_{[i]})\}^{I[Y_{\mathrm{E},i}=a, Y_{\mathrm{T},i}=b]} \\ &\times \prod_{i=n_1+1}^{n_{12}} \Big\{f_{\mathrm{S}}(Y_{\mathrm{S},i}^{o} \mid Y_{\mathrm{E},i}, Y_{\mathrm{T},i}, d_{[i]}, \boldsymbol{\theta}_{\mathrm{S}})\Big\}^{\delta_i} \Big\{\overline{F}_{\mathrm{S}}(Y_{\mathrm{S},i}^{o} \mid Y_{\mathrm{E},i}, Y_{\mathrm{T},i}, d_{[i]}, \boldsymbol{\theta}_{\mathrm{S}})\Big\}^{1-\delta_i} \end{aligned} \tag{8}$$

To complete the Bayesian model, we assume the non-informative priors

$$\Big(\pi_{0,0}(d), \pi_{0,1}(d), \pi_{1,0}(d)\Big) \sim \text{Dirichlet}(0.25, 0.25, 0.25, 0.25),$$

with γ and $\psi \sim \text{Gamma}(0.01, 0.01)$, and β_1, β_2, $\beta_{3,1}, \cdots, \beta_{3,J} \sim iid\ \text{N}(0, 10^2)$. We use MCMC to compute the posterior for $\boldsymbol{\theta}$ given the data $\mathcal{D}_{12} = \mathcal{D}_{ET}^{n_{12}} \cup \mathcal{D}_{\text{S},2}$.

The marginal survivor function at the end of follow-up is the probability weighted average

$$\overline{F}_{\text{S}}(t_S^* \mid X(d), \boldsymbol{\theta}) = \sum_{a=0}^{1} \sum_{b=0}^{1} \overline{F}_{\text{S}}(t_S^* \mid Y_{\text{E}} = a, Y_{\text{T}} = b, d, \boldsymbol{\theta}_S) \pi_{a,b}(d). \tag{9}$$

3.2 A utility-based Gen 1-2-3 design

Under this dose-outcome model, the following steps may be used to conduct a utility-based Gen-1-2-3 design. Dose-finding in stage 1 is done using the BOIN12 design (Lin et al., 2020), which relies on a numerical utility score for each of the four possible outcomes of (Y_E, Y_T). To establish a utility, we first assign the score $U_{1,0} = 100$ to the most desirable outcome $(Y_{\text{E}} = 1, Y_{\text{T}} = 0)$, and $U_{0,1} = 0$ to the least desirable outcome $(Y_{\text{E}} = 0, Y_{\text{T}} = 1)$. Using these two utilities as boundaries, we then ask the clinicians to provide their subjective utility scores $U_{1,1}$ for $(Y_{\text{E}} = 1, Y_{\text{T}} = 1)$, and $U_{0,0}$ for $(Y_{\text{E}} = 0, Y_{\text{T}} = 0)$. Table 1 provides an illustrative example. The mean utility of dose d based on (Y_E, Y_T), normed to take values between 0 and 1, is

$$\overline{U}(d, \boldsymbol{\theta}_{ET}) = \frac{1}{100} \sum_{a=0}^{1} \sum_{b=0}^{1} U_{a,b} \pi_{a,b}(d),$$

which is used as a phase 1-2 dose optimality criterion, $\phi(d, \boldsymbol{\theta}_{ET})$. The utility approach has several advantages (Zhou et al., 2019; Lin et al., 2020). It is scalable and readily accommodates ordinal Y_E and Y_T, as well as more than two endpoints (Liu et al., 2018). In addition, it is general and contains the marginal-probabilities-based risk-benefit tradeoff method as a special case. Lastly, as $U_{a,b}$ directly links to the clinical outcomes of patients, it can be easily interpreted and understood by clinicians.

The BOIN12 design treats $\overline{U}(d, \boldsymbol{\theta}_{TE})$ as a "probability" and uses the quasi-binomial method to estimate its posterior starting with a non-informative $Beta(0.5, 0.5)$ prior. The

design uses the posterior of $\overline{U}(d, \boldsymbol{\theta}_{ET})$ to make dose escalation and de-escalation decisions, and adaptively allocates patients to the dose that optimizes the posterior mean of $\phi(d, \boldsymbol{\theta}_{ET})$. The dose-finding algorithm of the BOIN12 design is given by (Lin et al., 2020). A Gen 1-2-3 trial with the above utility-based phase 1-2 component may be conducted as follows:

Stage 1. Enroll the first cohort of patients at a pre-specified starting dose. For each subsequent cohort, use the BOIN12 design to choose doses, up to n_1 patients.

Stage 2. For sample sizes $n = n_1 + 1, \cdots, n_1 + n_2$, if $\mathcal{A}_n^{can}$ is empty, terminate the trial. Otherwise, randomize each cohort of patients fairly among C and the doses in $\mathcal{A}_n^{can}$. When desirable, $\mathcal{A}_n^{can}$ can be updated after treating each cohort. At the end of stage 2, continue to follow patients for an additional of L months to collect PFS time data (Y_S^o, δ). Select the optimal dose $\widehat{d}^{\text{opt}}$ based on the survival probability $\overline{F}_{\text{S}}(t_S^* \mid X(d), \boldsymbol{\theta})$, and make the Go / No Go decision based on the PP. Details for calculating PP are provided in the online supporting materials. If the decision is No Go, the trial is stopped with the conclusion that $X(d^{opt})$ is not promising compared to C.

Stage 3. If the Go decision is made, then a GS phase 3 trial is conducted to compare the survival time distributions of $X(\widehat{d}^{\text{opt}})$ and C. In addition, the toxicity profile of $X(\widehat{d}^{\text{opt}})$ is still monitored in stage 3 during each interim analysis and the final test to ensure the safety of the selected optimal dose. For each interim decision $k = 1, \cdots, K-1$, after $n_{3,k}$ patients have been enrolled and followed for PFS, do two-sided tests for superiority or futility using a logrank test based on the combined data $\mathcal{D}_{S,2} \cup \mathcal{D}_{S,3}$. Denoting the approximately normal test statistic computed from the logrank statistic by Z, for the futility bound $\underline{b}_k$ and superiority bound $\bar{b}_k$, stop the trial if $|Z| < \underline{b}_k$ for futility, $|Z| > \bar{b}_k$ for superiority, or $\Pr\{\pi_T(d, \boldsymbol{\theta}_{ET}) < \overline{\pi}_T \mid \mathcal{D}_{ET}^n\} \leq .10$ for toxicity. Otherwise, do the final test when a total of n_3 patients have been enrolled and followed for PFS at $X(\widehat{d}^{\text{opt}})$ and C in stage 3.

As a result of the dose selection process at the end of stage 2, the use of the standard

hypothesis testing method based on the combined data $\mathcal{D}_{S,2} \cup \mathcal{D}_{S,3}$ may result in an inflated type I error rate. There are several ways to address this issue. One such approach involves using a combination test (Bauer and Köhne, 1994), which combines the p-value based on $\mathcal{D}_{S,2}$ (after applying the closing test procedure (Markus et al., 1976)) with the p-value based on $\mathcal{D}_{S,3}$. In this case, the standard GS boundaries, such as the O'Brien-Fleming boundary (O'Brien and Fleming, 1979), can be used as $\underline{b}_k$ and $\bar{b}_k$. This approach effectively controls the family-wise error rate (FWER), but may be conservative in some cases. For the Gen 1-2-3 design, the FWER is defined as the probability that any dose from the new treatment is selected at the end of stage 3, assuming that there is no existing dose with both acceptable toxicity and superior survival in comparison to the control.

Another simpler and often more powerful approach involves leveraging the built-in Go/No Go decision, which deflates the type I error rate, to cancel out the inflation caused by the dose selection process. Thus, the standard GS boundaries can still be used. Specifically, the cutoff $p_U(Go)$ is calibrated under the null hypothesis that none of the doses are effective with respect to Y_S using simulation, such that the FWER is controlled at the nominal value. This is the approach we employed in our simulation, and we found that a cutoff of $p_U(Go) = 0.5$ or higher is often sufficient to cancel out the type I error inflation and effectively control FWER.

The maximum sample size n_3 in stage 3 depends on the accumulated numbers of patients treated with $X(\hat{d}^{opt})$ and C, because we fix the maximum sample size for the GS design to be n_{GSD} rather than fixing n_3. For example, suppose that a GS design is planned in stage 3 with an interim analysis in the middle of the trial using a maximum sample size of $n_{\text{GSD}} = 500$, and 15 and 15 patients have been treated at $X(\widehat{d}^{\text{opt}})$ and C in stage 2. In stage 3, we first enroll 250-30=220 patients before the interim analysis and then, if stage 3 is continued, enroll the additional 250 patients after the interim analysis, which gives $n_3 = 220 + 250 = 470$.

4 Simulation Studies

4.1 Simulation Settings

We conducted simulations to evaluate the operating characteristics (OCs) of the utility-based Gen 1-2-3 design presented in the last section. Let $\pi_T^{true}(d_j)$ be the true toxicity probability, $\pi_E^{true}(d_j)$ the true short-term response probability, and $\overline{F}^{true}(t_S^*, d_j)$ the true survival probability at follow up time t_S^*. We evaluated the design's performance under eight scenarios having a variety of different patterns for $\pi_T^{true}(d_j)$, $\pi_E^{true}(d_j)$ and $\overline{F}^{true}(t_S^*, d_j)$, shown in Figure 2.

As comparators, we used two conventional utility-based phase 1-2 designs followed by a phase 3 design, referred to as Conv 1 and Conv 2. The Conv 1 design consists of stages 1 and 2 of the Gen 1-2-3 design, but does not use any Y_S data to make decisions in these stages. Rather, an optimal dose $\widehat{d}_{ET}^{opt}$ is selected by maximizing the posterior mean of $\overline{U}(d, \boldsymbol{\theta}_{ET})$, and $X(\widehat{d}_{ET}^{opt})$ is used in phase 3 if

$$\Pr\left\{\pi_E(X(\widehat{d}_{ET}^{opt}), \boldsymbol{\theta}_{ET}) > \pi_E(C, \boldsymbol{\theta}_{ET}) \mid \mathcal{D}_{ET,1} \cup \mathcal{D}_{ET,2}\right\} > 0.8.$$

This is a Go / No Go rule based on the ET data but no survival time data, and the decision criterion is a posterior probability rather than a predictive probability involving simulated future data. If phase 3 is conducted, the same GS design used by Gen 1-2-3 is used in the phase 3 portion of the Conv 1 design. The Conv 2 design is similar to the Conv 1 design, with the two differences that (1) C is excluded from the phase 1-2; and (2) there is no Go / No Go decision between phase 1-2 and phase 3, so phase 3 always is conducted. We also considered a modified version of the CT design as a comparator. The original CT design used the EffTox phase 1-2 design, which relies on a regression model with parametric dose-response and dose-toxicity curves to borrow strength between doses, and an efficacy-toxicity

trade-off contour as a criterion for dose selection (Thall and Cook, 2004; Thall et al. 2014). To obtain a more fair comparison, we modified the CT design by using the same BOIN12 design and mean utility objective function in stages 1 and 2 of the Gen 1-2-3 design in the application. We refer to this as the CT-B12 design.

For the admissibility criteria (1), we set $\overline{\pi}_T = 0.35$ and $\underline{\pi}_E = 0.20$. The fixed follow up window for Y_S^o was $t_S^* = 6$ months. We also set the additional follow-up time to harvest survival time data from all patients in stage 2 who have been randomized and treated, prior to selection of optimal dose, to $L = 1$ month. Patients were treated in cohorts of size 3 in stage 1 and size 5 in stage 2, assuming a mean accrual rate of 1 cohort per month in stages 1 and 2 and 10 patients per month in stage 3. The maximum sample sizes were $n_1 = 30$, $n_2 = 50$, and $n_{\text{GSD}} = 500$ for the phase 3 GS design. The remaining design parameters were $\rho = 0.5$, $\underline{\lambda} = 0.85$, $p_U = 0.8$ and $p_U(Go) = 0.5$, determined from preliminary simulation studies.

As data generation models for the simulation studies, to obtain (Y_E, Y_T) we first simulated latent variables $\mathbf{W} = (W_E, W_T)$ following a bivariate normal distribution with mean (0,0), variances 1, and correlation .10. We then defined

$$Y_E = \begin{cases} 0 & \text{if } W_E < \kappa_E(d_j) \\ 1 & \text{if } W_E \geq \kappa_E(d_j) \end{cases}, \quad Y_T = \begin{cases} 0 & \text{if } W_T < \kappa_T(d_j) \\ 1 & \text{if } W_T > \kappa_T(d_j) \end{cases}$$

with the cut-offs $\kappa_E(d_j)$ and $\kappa_T(d_j)$ chosen to obtain the maginal probabilities $\pi_E^{true}(d_j)$ and $\pi_T^{true}(d_j)$ for each d_j specified under each scenario. To generate survival times, we assumed that $t_S^* = 6$, and that $[Y_S \mid Y_E, Y_T, d]$ followed a piecewise exponential (PE) distribution with survival function

$$\overline{F}_{Y_S}(t \mid Y_E, Y_T, d) = \exp\{-t/\tilde{\lambda}(t, Y_E, Y_T, d)\}, \quad 0 < t \leq t_S^*,$$

where the log hazard function takes the piecewise form

$$\log\{\tilde{\lambda}(t, Y_E, Y_T, d)\} = \begin{cases} \tilde{\beta}_{01} + \tilde{\beta}_{E1} Y_E + \tilde{\beta}_{T1} Y_T + \sum_{j=0}^{J} \tilde{\gamma}_{j1} I[d = d_j] & \text{if } 0 < t \leq 3 \\ \tilde{\beta}_{02} + \tilde{\beta}_{E2} Y_E + \tilde{\beta}_{T2} Y_T + \sum_{j=0}^{J} \tilde{\gamma}_{j2} I[d = d_j] & \text{if } 3 < t \leq 6, \end{cases}$$

setting $\tilde{\gamma}_{01} = \tilde{\gamma}_{0,2} = 0$ and $d = d_0$ for C. Since $t_S^* = 6$, the parameters $\kappa_E(d_j)$, $\kappa_T(d_j)$, $\tilde{\beta}_{01}$, $\tilde{\beta}_{02}$, $\tilde{\beta}_{E1}$, $\tilde{\beta}_{E2}$, $\tilde{\beta}_{T1}$, $\tilde{\beta}_{T2}$, and $\tilde{\gamma}_{j1}, \tilde{\gamma}_{j2}$ were derived under each scenario to match the pre-determined values of $\pi_T^{true}(d_j)$, $\pi_E^{true}(d_j)$ and $\overline{F}^{true}(t_S^*, d_j)$ for $j = 0, \cdots, 5$. For theoretical consistency, if desired, the final interval for defining the PE hazard may be extended to be infinite to ensure that the distribution of all possible Y_S is well-defined.

4.2 Simulation Results

Table 2 summarizes OCs of the Gen 1-2-3, Conv 1, Conv 2 and CT-B12 designs, including optimal dose selection percentages at the end of stage 2, final treatment selection percentages at the end of stage 3, mean number of patients treated at each dose and C, mean trial duration by month, mean overall sample size, the mean percentage of Go decisions, and the mean percentage of the Go decision with the true optimal dose given that a Go decision is made (in parentheses). All results are based on 5,000 simulated replicates of the trial using each design.

Scenario 1 is a null case where no dose d gives $X(d)$ with better survival than C. Because a Go / No Go decision is included in the Gen 1-2-3 and the Conv 1 designs, both may stop the trial early, at the end of stage 2. The Gen 1-2-3 design correctly selects C 94.7% of the time, with a similar value 94.8% for the Conv 1 design. This percentage drops to 77.4% for the Conv 2 design and 45.2% for the CT-B12 design. Gen 1-2-3 and Conv 1 have much shorter trial durations of 33.7 and 43.3 months, and much smaller sample sizes of 198.7 and 283.8, compared to respective durations 65.7 and 62.5 and sample sizes 479.4 and 447.2 for

the Conv 2 and CT-B12 designs. In addition, the Gen 1-2-3 has only a 32.5% chance of conducting phase 3. These results quantify the substantial advantage obtained by including a Go / No Go rule.

In scenario 2, d_4 is truly optimal with the highest $\overline{F}^{true}(t_S^*, X(d_4)) = 0.60$, whereas d_5 has the highest mean utility $\overline{U}^{true}(X(d_5)) = 62$. The Gen 1-2-3 design has the highest percentage, 56.4%, of correctly selecting d_4 as optimal, and a 80.9% chance of making a Go decision, and 69.7% of the Go decision eventually goes to d_4, whereas the optimal dose selection percentages for d_4 are 12.2%, 30.9% and 21.2% under the Conv 1, Conv 2 and CT-B12 designs. For patient allocation, the Gen 1-2-3 design assigns an average of 81.8 patients to d_4, compared to 31.8, 57.2, and 42.3 for the Conv 1, Conv 2 and CT-B12 designs. The Conv 1 and Conv 2 designs assign most patients, 50.7 and 108.1, to d_5, which has the highest mean utility. The CT-B12 design assigns patients evenly among the doses, with the numbers ranging from 34.1 to 58.3. For trial duration and mean sample size, the designs with a Go / No Go decision, Gen 1-2-3 and Conv 1, have better results than Conv 2 and CT-B12. In scenario 3, d_2 is truly optimal dose and has the highest mean utility. The Gen 1-2-3 design again outperforms the other designs, with at least a 30% higher correct treatment selection percentage, and it also assigns at least 26 more patients to d_2.

In scenario 4, doses d_2, d_3, d_4, and d_5 all have the same PFS probability 0.60 at six months, but d_4 and d_5 have unacceptably high toxicity rates of 0.40 and 0.50. Thus, there are two true optimal treatments, $X(d_2)$ and $X(d_3)$, in this scenario. However, $X(d_2)$ is preferable due to its much smaller toxicity probability 0.05, compared with 0.30 for $X(d_3)$. The Gen 1-2-3 design performs well compared to the Conv 1 and CT-B12 designs, with a 76.6% chance of selecting a true optimal treatment. Between the two true optimal treatments, the Gen 1-2-3 design slightly favors $X(d_2)$, with a higher treatment selection percentage of 46.8% and a larger number of patients, 77.8. This is because the Gen 1-2-3 design uses the posterior mean utility for dose screening and randomization in stage 2, and the utility function accounts for

toxicity. The Conv 2 design has a slightly better performance than the Gen 1-2-3 design in this scenario. That is because the Conv 2 selects the optimal dose based on the mean utility, and doses d_2 and d_3, which are the true optimal doses in terms of survival, happen to have better mean utility values than other doses in scenarios 4. Besides, there is no Go/ No Go decision equipped for the Conv 2 design, which further improves the power of the design. However, as noticed in scenario 1, the Conv 2 design fails to control the FWER if there is no existing optimal dose. Additional simulation results, for scenarios 5, 6, 7, and 8, are given in Table S1 of the online supplementary materials. The results are qualitatively similar to those obtained for scenarios 1, 2, 3, and 4.

4.3 Sensitivity Analyses

We performed additional simulations to explore the sensitivity of the Gen 1-2-3 design to several of is parameters. The results are summarized in the online supporting information. The simulations in Table 2 were based on $p_U(Go) = 0.5$ for the Go / No Go decision. Table S2 summarizes the OCs of the Gen 1-2-3 design using each of the values $p_U(Go) = 0$, 0.5 and 0.9, where $p_U(Go) = 0$ implies that phase 3 always is conducted. As expected, larger $p_U(Go)$ is more favorable under the null scenarios 1 and 5, while smaller $p_U(Go)$ is more favorable when a truly optimal experimental treatment $X(d_j)$ exists for $j = 1, \cdots$, or 5, in scenarios 2 – 4 and 6–8. Considering both null and alternative scenarios, it appears that $p_U(Go) = 0.5$ is the best selection to obtain generally good performance. A possible reason that $p_U(Go) = 0.5$ works well is that the Go/No Go decision serves as a screening rule, and when making a screening decision, it may be more desirable to use a relatively low cut-off value that primarily eliminates very unpromising doses. If a dose d for which $X(d)$ in fact does not improve survival compared to C passes the Go/No Go decision, the chance of it passing the subsequent GS test and being selected as an effective treatment still is low.

Table S3 shows the effects of using different randomization methods in stage 2. We denote

equal randomization in Table 1 by ER, and consider two alternative methods. Method 2, referred to as Half, randomizes half of the patients to C and uses AR for the doses in $\mathcal{A}_n^{can}$. Method 3 uses AR to randomize patients among C and the doses in $\mathcal{A}_n^{can}$. The results show that ER and AR give very similar overall performances, and both outperform Half. It thus appears that stage 2 of the Gen 1-2-3 design does not require the use of AR.

Table S4 studies the effect of different additional follow up times L on the OCs of the Gen 1-2-3 designs. The results confirm that larger L improves the Go / No Go decision, and therefore can generally give better performances for optimal treatment selection and patient allocation. The changes in mean sample size are minimal, but the total trial duration is prolonged. In practice, L should be chosen, based on preliminary simulations, in terms of the tradeoff among correct treatment selection, patient allocation, and trial duration.

Table S5 considers different types of candidate dose sets. In Table S5, an alternative $\mathcal{A}_{n_1}^{can}$ was defined that uses only the n_1 patients in stage 1 and keeps this set unchanged thereafter. The results show that these two ways to define a candidate dose set yield similar performances. In particular, $\mathcal{A}_n^{can}$ performs slightly better than $\mathcal{A}_{n_1}^{can}$ under the null scenarios and $\mathcal{A}_{n_1}^{can}$ is slightly better than $\mathcal{A}_n^{can}$ under the alternative scenarios. Also, using $\mathcal{A}_{n_1}^{can}$ can increase the mean sample size slightly.

Table S6 investigates different parametric models to generate the survival outcomes for the simulation studies. In addition to the piece-wise exponential distribution used in Table 1, the Weibull and log-logistic distributions are also considered in this Table. The results show that different distributions to generate survival outcomes give very similar OCs, which confirms the robustness of the proposed exponential-gamma model.

Lastly, the FWER using the Gen 1-2-3 design is also evaluated. Table S7 lists various representative scenarios for FWER evaluation. In particular, scenarios S1 and S2 represent cases where all the doses are safe but none of them has better survival than the control. In scenarios S3 and S4, doses d_4 and d_5 are overly toxic, and there is either no existing dose

with better survival (scenario S3), or the dose with higher survival rate is unsafe (scenario S4). In scenarios S5 and S6, all the doses are overly toxic. Figure S1 depicts the empirical FWER of the Gen 1-2-3 design under these scenarios. The results show that the Gen 1-2-3 can well control the overall FWER around a nominal level of 5%.

5 Discussion

Stages 1 and 2 of a Gen 1-2-3 design may be thought of as a refined dose screening and selection process based on both early and late outcomes. Stage 1 uses efficacy and toxicity with conventional phase 1-2 design machinery to assign patients sequentially to doses, screen out unsafe or ineffective doses, and reduce the original dose set to the RP2S $\mathcal{A}_{n_1}^{can}$. This step in stage 1 is important because identifying a candidate dose set rather than selecting one optimal dose allows the dose space to be explored more fully based on survival time in subsequent steps. Stage 2, which may be regarded as a link between sequentially adaptive dose finding and phase 3, has four key elements. (1) Including C ensures that each $X(d_j)$ evaluation is based on a comparison to C, and (2) randomization ensures that these comparisons are unbiased. (3) In contrast with conventional phase 1-2 designs, the dose optimization criterion is based on survival time Y_S, rather than on the early (Y_E, Y_T) outcomes. Finally, (4) stage 2 includes a Go / No Do decision based on a PP of $X(d^{opt})$ superiority over C in terms of survival. This greatly reduces the risk of wasting phase 3 resources on an agent that is unlikely to provide a survival benefit over C.

In terms of trial conduct, the main additional practical requirement of a Gen 1-2-3 design is the inclusion of stage 2 between a phase 1-2 design and phase 3, including computations for the decisions, dose screening, and randomizations made by the computer program. Due to its additional structure and complexity, the necessary computer simulations required to construct a design may be time-consuming, and certainly will require careful planning and

interactions between statisticians and physician investigators. The simulations indicate that this additional effort is warranted, since they show that the Gen 1-2-3 design has very desirable properties compared to more conventional approaches. R code for implementing the Gen 1-2-3 design is available from https://github.com/yongzang2020.

Acknowledgment

The authors thank the associate editor and two referees for valuable comments. The research of Yong Zang is partial supported by NIH grants R01 GM150808, R21 CA264257, P30 CA082709, and the Ralph W. and Grace M. Showalter Research Trust award. Peter and Ying: please add your funding support.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this paper. The R code for the Gen 1-2-3 design is available from https://github.com/yongzang2020.

References

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.

Bashir, Q., Thall, P. F., Qazilbash, M. H., et al. (2019) Conditioning with busulfan plus melphalan versus melphalan alone before autologous haemopoietic cell transplantation for multiple myeloma: an open-label, randomised, phase 3 trial. *The Lancet Hematology* **6**: 266-275.

Bauer, P. and Köhne, K. (1994) Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**, 1029–1041.

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.

Brock K, Homer V, Soul G, Potter C, Chiuzan C, and Lee S. (2021) Is more better? An analysis of toxicity and response outcomes from dose-finding clinical trials in cancer. *BMC Cancer* **21**: 777.

Cai C, Liu S, Yuan Y. (2014) A Bayesian design for phase II clinical trials with delayed responses based on multiple imputation. *Statistics in Medicine* **33(23)**:4017-4028.

Chapple, A. G. and Thall, P. F. (2019) A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III (with discussion). *Biometrics* **75**: 371-381.

Chapple, A.G. and Thall, P.F. (2020) Comparison of phase I-II designs with parametric or semi-parametric models using two different risk-benefit trade-off criteria. *Contemporary Clinical Trials* **97**: 106099.

Guo B, Yuan Y. (2017) Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *Journal of American Statistical Association* **112**: 508-520.

Guo B, Yuan Y. (2023) DROID: dose-ranging approach to optimizing dose in oncology drug development. *Biometrics* In press.

Korn, E. L., Freidlin, B., Abrams, J. S., Halabi, S. (2012). Design issues in randomized phase II/III trials. *Journal of Clinical Oncology* **30**: 667-671.

Inoue LYT., Thall P.F., and Berry D.A. (2002) Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* **58**: 823-831.

Lee J, Thall PF, 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.

Lin, R., Zhou, Y., Yan, F., Li, D. and Yuan, Y. (2020) BOIN12: Bayesian optimal interval phase I/II trial design for utility-based dose finding in immunotherapy and targeted therapies. *Journal of Clinical Oncology Precision Oncology* **4**: 1393-1402.

Liu S, Yuan Y. (2015) Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C***64**: 507-523.

Liu S, Guo B, Yuan Y. (2018) A Bayesian phase I/II design for immunotherapy trials. *Journal of American Statistical Association* **113**: 1016-1027.

Markus, R., Pertiz, E. and Gabriel, K.R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

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

O'Brien PC, Fleming TR. (1979) A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Ratain, M. (2014) Redefining the primary objective of phase Ioncology trials. *Nature Reviews: Clinical Oncology*, **11**: 50-504.

Ratain, M., Tannock, I. and Lichter, A. (2021) Dose optimization of Sotorasib: is the US Food and Drug Administration sending a message? *Journal of Clinical Oncology*, **39**: 3423-3426.

Schaid, D.J., Wieand, S., and Therneau, T. (1990) Optimal two-stage screening designs for survival comparisons. *Biometrika*, **77**: 507-513.

Shah, M., Rahman, A., Theoret, M. and Pazdur, R. (2021) The drug dosing conundrum in oncology-when less is more. *The New England Journal of Medicine*, **385**: 1445-1447.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*, **22**: 286-703.

Storer BE. (1989) Design and analysis of phase I clinical trials. *Biometrics* **45**: 925-937.

Storer BE (2001) An evaluation of phase I clinical trials in the continuous dose-response setting. *Statistics in Medicine* **20**: 2399-2408.

Thall, P.F., Simon, R., and Ellenberg, S.S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**: 303-310.

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

Thall PF, Cook JD. (2004) Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* **60**: 684-693.

Thall, P.F. (2008). A review of phase 2-3 clinical trial designs. *Life time Data Analysis* **14**: 37-53.

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

Thall, P. F. (2020) *Statistical Remedies for Medical Researchers*. Springer Nature Switzerland.

Thall PF, Wooten LH, Tannir NM. (2005) Monitoring event times in early phase clinical trials: some practical issues. *Clinical Trials* **2(6)**: 467-478.

Thall PF., Zang Y. and Yuan Y. (2023) Generalized phase I-II designs to increase long term therapeutic success rate. *Pharmaceutical Statistics* in press.

Yan, F., Thall, P. and Yuan, Y., (2018) Phase I–II clinical trial design: a state-of-the-art paradigm for dose finding. *Annals of Oncology* **29**: 694-699.

Yin, G., Chen, N. and Lee, J.J. (2018) Bayesian adaptive randomization and trial monitoring with predictive probability for time-to-event endpoint. *Statistics in Biosciences* **10**: 420-438.

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

Zang, Y., Lee, J. and Yuan, Y. (2014) Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials* **11**: 319-327.

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

Zhang Y, Cao S, Zhang C, Jin IH, Zang Y. (2021) A Bayesian adaptive phase I/II clinical trial design with late-onset competing risk outcomes.*Biometrics* **77**: 796-808.

Zhou H, Chen C, Sun L, Yuan Y. (2020) Bayesian optimal phase II clinical trial design with time-to-event endpoint. Pharmaceutical Statistics **19(6)**:776-786.

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 fortargeted and immune therapies. *Statistics in Medicine* **38**: 5299-5316.

Figure 1: Schematic for the Gen 1-2-3 design.

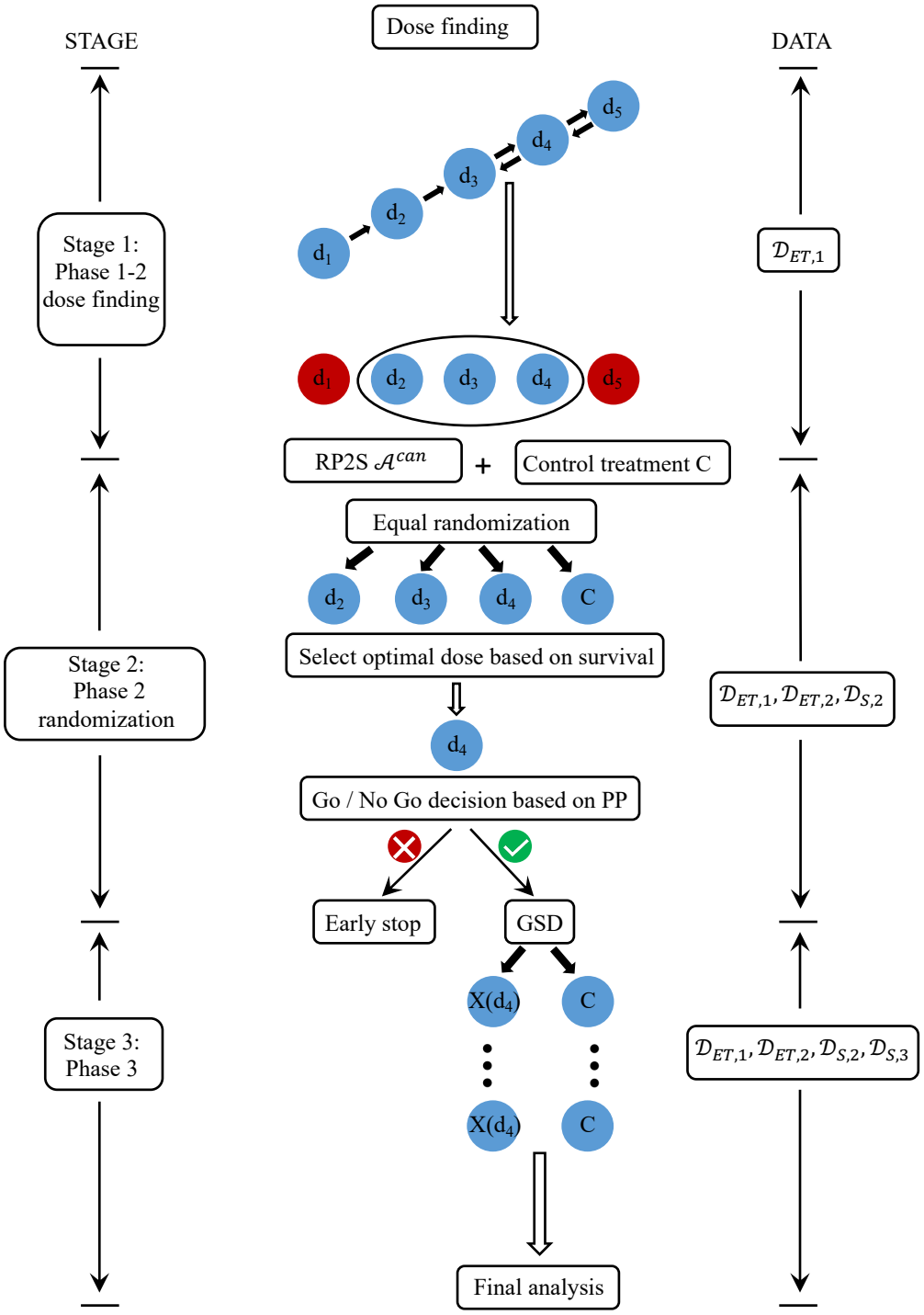

Figure 2: Dose-outcome curves for the scenarios in the simulation study. The red, green, and blue curves are $\pi_T^{true}(d_j)$, $\pi_E^{true}(d_j)$ and $\overline{F}^{true}(t_S^*, d_j)$. The dashed line shows $\overline{F}^{true}(t_S^*, C)$ for the control. The doses with unacceptable toxicity probabilities and short-term response probabilities are given in **red**.

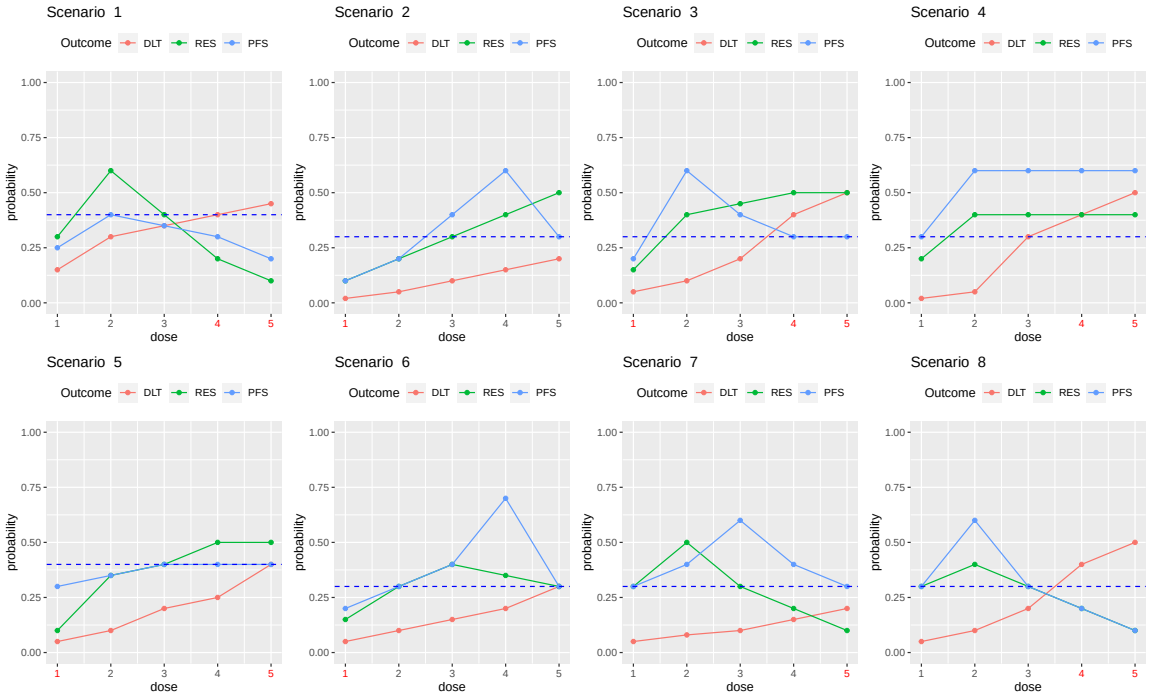

Table 1: An example of a utility table for two binary outcomes.

	$Y_{\mathrm{T}} = 0$	$Y_{\mathrm{T}} = 1$
$Y_{\mathrm{E}} = 0$	$u_{0,0} = 40$	$u_{0,1} = 0$
$Y_{\mathrm{E}} = 1$	$u_{1,0} = 100$	$u_{1,1} = 60$

Table 2: Optimal dose selection %, final treatment (Trt) selection %, mean number of patients treated with each dose level and the control (0*), mean trial duration, and mean sample size. **Boldface** indicates results for the true optimal decision. True toxicity probabilities $> .35$ and short-term response probabilities $< .2$ are given in **red**.

Design		Dose levels 0*	1	2	3	4	5	Trial Dur	Sample size	Go %
					Scenario 1					
	$\pi_T^{true}(d_j)$	.20	.15	.30	.35	**.40**	**.45**			
	$\pi_E^{true}(d_j)$	.30	.30	.60	.40	.20	**.10**			
	$\overline{U}^{true}(d_j)$	50.0	52.0	64.0	50.0	36.0	28.0			
	$\overline{F}^{true}(t_S^*, d_j)$	**.40**	.25	.40	.35	.30	.20			
Gen 1-2-3	Dose %	1.2	26.1	50.9	18.9	2.9	0	33.7	198.7	32.5
	Trt %	**94.7**	2.8	1.1	1.1	0.3	0			
	# Pats	75.6	30.7	62.1	25.1	4.5	0.8			
Conv 1	Dose %	1.4	17.8	70.2	10.4	0.2	0.0	43.3	283.8	52.3
	Trt %	**94.8**	2.5	1.3	1.4	0.0	0.0			
	# Pats	117.5	26.5	114.9	21.6	2.5	0.8			
Conv 2	Dose %	1.4	16.9	73.0	8.5	0.2	0.0	65.7	479.4	100
	Trt %	**77.4**	16.3	3.8	2.3	0.2	0.0			
	# Pats	200.1	64.1	180.4	30.9	3.0	0.8			
CT-B12	Dose %	1.6	43.1	24.6	22.3	6.7	1.7	62.5	447.2	100
	Trt %	**45.2**	41.8	1.4	5.0	4.9	1.7			
	# Pats	200.1	99.0	75.3	52.8	16.7	3.3			
					Scenario 2					
	$\pi_T^{true}(d_j)$	.10	.02	.05	.10	.15	.20			
	$\pi_E^{true}(d_j)$	.30	**.10**	.20	.30	.40	.50			
	$\overline{U}^{true}(d_j)$	54.0	45.2	50.0	54.0	58.0	62.0			
	$\overline{F}^{true}(t_S^*, d_j)$	.30	.10	.20	.40	**.60**	.30			
Gen 1-2-3	Dose %	0.1	0.9	5.5	20.4	64.0	9.1	48.1	312.6	80.9
	Trt %	29.2	0.3	2.0	11.7	**56.4**	0.4			(69.7)
	# Pats	126.1	11.7	18.2	48.1	81.8	26.7			
Conv 1	Dose %	0.1	2.0	6.1	16.8	32.4	42.6	34.9	207.2	36.3
	Trt %	83.0	0.2	0.3	3.4	**12.2**	0.9			(33.6)
	# Pats	73.3	11.7	14.5	25.3	31.8	50.7			
Conv 2	Dose %	0.0	1.4	5.1	15.1	30.9	47.5	61.5	435.3	100
	Trt %	51.2	1.4	4.4	10.1	**30.9**	2.0			
	# Pats	177.6	14.6	26.6	51.1	57.2	108.1			
CT-B12	Dose %	0.0	22.9	21.8	21.9	21.2	12.2	58.9	409.2	100
	Trt %	23.3	22.9	16.6	14.9	**21.2**	1.1			
	# Pats	177.5	34.1	55.9	58.3	42.3	41.1			

Table 2 (continued)

Designs		Dose levels 0*	1	2	3	4	5	Trial duration	Sample size	Go %
				Scenario 3						
	$\pi_T^{true}(d_j)$	.10	.05	.10	.20	**.40**	**.50**			
	$\pi_E^{true}(d_j)$	.30	**.15**	.40	.45	.50	.50			
	$\overline{U}^{true}(d_j)$	54.0	47.0	60.0	59.0	54.0	50.0			
	$\overline{F}^{true}(t_S^*, d_j)$	.30	.20	**.60**	.40	.30	.30			
Gen 1-2-3	Dose %	0.1	2.5	73.2	18.4	4.8	1.0	46.6	297.5	81.7
	Trt %	26.3	0.7	**63.1**	9.9	0.0	0.0			(77.2)
	# Pats	120.7	15.8	93.9	47.9	15.8	3.5			
Conv 1	Dose %	0.1	1.6	27.8	39.3	26.1	5.1	40.1	253.4	45.3
	Trt %	77.2	0.3	**9.2**	12.0	1.1	0.2			(20.3)
	# Pats	98.1	14.5	32.4	58.6	39.7	10.2			
Conv 2	Dose %	0.1	1.0	32.5	38.7	24.4	3.3	62.7	446.7	100
	Trt %	37.4	0.7	**32.5**	27.8	1.3	0.3			
	# Pats	183.4	18.2	67.8	109.5	58.4	9.4			
CT-B12	Dose %	0.2	27.8	31.4	19.7	16.5	4.4	60.5	425.2	100
	Trt %	33.1	22.2	**31.4**	12.3	0.7	0.3			
	# Pats	186.8	67.9	58.8	60.2	40.9	10.7			
				Scenario 4						
	$\pi_T^{true}(d_j)$	.10	.02	.05	.30	**.40**	**.50**			
	$\pi_E^{true}(d_j)$	.30	.20	.40	.40	.40	.40			
	$\overline{U}^{true}(d_j)$	54.0	51.2	62.0	52.0	48.0	44.0			
	$\overline{F}^{true}(t_S^*, d_j)$	.30	.30	**.60**	**.60**	.60	.60			
Gen 1-2-3	Dose %	0.0	4.1	50.5	32.3	10.5	2.6	47.0	294.3	92.2
	Trt %	17.0	0.4	**46.8**	**29.8**	5.9	0.1			(83.1)
	# Pats	119.7	23.5	77.8	50.2	18.2	5.0			
Conv 1	Dose %	0.0	4.1	48.0	33.2	11.8	2.9	30.4	164.0	33.1
	Trt %	67.2	0.0	**14.2**	**12.5**	5.0	1.1			(80.7)
	# Pats	54.3	17.0	43.3	32.0	13.8	3.7			
Conv 2	Dose %	0.0	3.2	50.2	32.0	11.4	3.2	51.6	336.3	100
	Trt %	2.9	0.3	**50.2**	**32.0**	11.4	3.2			
	# Pats	128.1	25.4	93.8	59.1	22.8	7.0			
CT-B12	Dose %	0.0	27.2	22.3	31.3	14.2	5.0	52.6	345.8	100
	Trt %	26.3	0.9	**22.3**	**31.3**	14.2	5.0			
	# Pats	147.4	63.3	54.7	49.9	22.8	7.9			