Received: 6 February 2018 | Accepted: 17 May 2018
DOI: 10.1111/biom.12994

ORIGINAL ARTICLE

WILEY Biometrics

# A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III

**Andrew G. Chapple**[1] | **Peter F. Thall**[2]

[1]Department of Statistics, Rice University, Houston, Texas

[2]Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas

**Correspondence**
Andrew G. Chapple, Department of Statistics, Rice University, Houston, TX
Email: achapp@lsuhsc.edu
Peter F. Thall, Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX
Email: rex@mdanderson.org

**Abstract**

Conventionally, evaluation of a new drug, $A$, is done in three phases. Phase I is based on toxicity to determine a "maximum tolerable dose" (MTD) of $A$, phase II is conducted to decide whether $A$ at the MTD is promising in terms of response probability, and if so a large randomized phase III trial is conducted to compare $A$ to a control treatment, $C$, usually based on survival time or progression free survival time. It is widely recognized that this paradigm has many flaws. A recent approach combines the first two phases by conducting a phase I-II trial, which chooses an optimal dose based on both efficacy and toxicity, and evaluation of $A$ at the selected optimal phase I-II dose then is done in a phase III trial. This paper proposes a new design paradigm, motivated by the possibility that the optimal phase I-II dose may not maximize mean survival time with $A$. We propose a hybridized design, which we call phase I-II/III, that combines phase I-II and phase III by allowing the chosen optimal phase I-II dose of $A$ to be re-optimized based on survival time data from phase I-II patients and the first portion of phase III. The phase I-II/III design uses adaptive randomization in phase I-II, and relies on a mixture model for the survival time distribution as a function of efficacy, toxicity, and dose. A simulation study is presented to evaluate the phase I-II/III design and compare it to the usual approach that does not re-optimize the dose of $A$ in phase III.

**KEYWORDS**

Bayesian design, clinical trial, dose finding, phase I-II clinical trial, phase III clinical trial

## 1 | INTRODUCTION

After a new treatment agent, $A$, is identified in pre-clinical studies, conventional clinical drug development and evaluation is carried out in three phases (Cancer.org, 2018). In phase I, the aim is to identify a dose, called the "maximum tolerable dose" (MTD), having acceptable toxicity probability. Phase I trials typically are small, with a wide variety of designs, including the 3+3 algorithm (Storer, 1989), continual reassessment method (O'Quigley et al., 1990), and escalation with overdose control (Babb et al., 1998). Efficacy of $A$ at the MTD then is evaluated in phase II using the estimated probability $\pi_E$ of a short-term event ("response"), such as 50% solid tumor shrinkage or complete remission of leukemia. Most phase II designs compare $\pi_E(A)$ with $A$ at the MTD to an assumed $\pi_E(C)$ of a conventional therapy, $C$. Phase II trials often are small, and may include an early stopping rule if $\pi_E(A)$ is poor compared to $\pi_E(C)$. If $A$ is found to be promising in phase II, this may motivate a randomized phase III trial of $A$ versus $C$ based on a long-term outcome, such as survival time.

Many phase II designs have been published. Simon et al. (1985) proposed a randomized selection design for two or more experimental treatments. For single-arm phase II trials, two-stage designs were proposed by Simon (1989) based on response, and by Bryant and Day (1995) based on response

  

and toxicity. Bayesian sequential designs were proposed by Thall and Simon (1994) for a binary response, and by Thall et al. (1995) for monitoring multiple outcomes. Lee and Liu (2005) used predictive probabilities for futility rules, and Yin et al., (2012) used adaptive randomization to favor empirically better treatment arms.

It now is recognized widely that the conventional phase I → phase II → phase III paradigm has many flaws, and has led to many negative phase III trials. Two studies (Arrowsmith, 2011; Bio, 2016) showed that only about 50% of phase III trials yield an improvement over standard therapy. Seruga, et al. (2015) discussed causes of failure in phase III, including insufficient evidence of anti-disease activity in early phase trials, disagreements about how phase II trials should be designed, and reliance on phase II efficacy events or other surrogates not associated with longer survival. Yuan et al. (2016, Chapter 1) discuss problems with the conventional phase I → phase II paradigm, mainly due to limited sample sizes and ignoring efficacy when determining an MTD in phase I.

Many alternatives have been proposed that create hybrid designs by combining conventional phases, most commonly phase I-II or phase II-III. Thall (2008) reviewed phase II-III designs and discussed problems with the conventional phase II → phase III paradigm. "Select-and-test" phase II-III designs, where two or more experimental agents are chosen in phase II and randomized against $C$ in phase III while maintaining desired overall type I and type II error rates, are given by Thall et al. (1988), Schaid et al. (1990), Stallard and Todd (2003), and many others. A phase II-III design proposed by Inoue et al. (2004) uses both an early efficacy (response) indicator, $Y_E$, and survival time, $Y_S$. Denote $\pi_E = \Pr(Y_E = 1)$, the probability density function (pdf) of $Y_S$ by $f_S(t)$, and the conditional pdf of $[Y_S \mid Y_E]$ by $f_S(t \mid Y_E = y)$ for $y = 0, 1$. Their approach relies on a mixture model of the general form

$$f_S(t) = f_S(t \mid Y_E = 1)\pi_E \; + \; f_S(t \mid Y_E = 0)(1 - \pi_E). \quad (1)$$

Denote the indicator of early toxicity by $Y_T$ and $\pi_T = \Pr(Y_T = 1)$. Because phase I designs use $Y_T$ but ignore $Y_E$ when choosing a MTD, they are likely to choose a dose having reasonable $\pi_T$ but ineffectively low $\pi_E$. For example, consider a dose-finding scenario with five doses, true toxicity probabilities (0.05, 0.10, 0.20, 0.30, 0.35), and true efficacy probabilities (0.05, 0.10, 0.20, 0.30, 0.60). If the CRM is used with target toxicity probability 0.30, this most likely will select dose 4 as optimal. By ignoring $Y_E$, however, dose 5 is chosen less frequently, despite the fact that it has only a 0.05 higher $\pi_T$ than dose 4 but doubles $\pi_E$ from 0.30 to 0.60. Phase I-II designs are motivated, in part, by the desire to overcome this sort of problem. Examples include the two-stage design of Hoering et al. (2011), studying combination therapies (Huang et al., 2006), using the odds ratio between $\pi_E$ and $\pi_T$ (Yin et al., 2006), and basing decisions on elicited numerical utilities of the possible elementary events determined by efficacy and toxicity (Thall and Nguyen, 2012). Thall and Cook (2004) proposed, and Thall et al. (2014) refined, the so-called "Eff-Tox" phase I-II design based on maximizing an estimate of an efficacy-toxicity trade-off function, $\phi(\pi_E, \pi_T)$. The function $\phi(\pi_E, \pi_T)$ increases in $\pi_E$, decreases in $\pi_T$, and quantifies the desirability of each probability pair $(\pi_E, \pi_T)$.

This article presents a new Bayesian hybrid design that combines a phase I-II design followed by a modified phase III design, based on both early and late outcomes. We will call this a phase I-II/III trial design. For simplicity, we will use survival time, $Y_S$, as the long-term outcome, although progression-free survival (PFS) time will work in precisely the same way. Our approach relies on a mixture model for the distribution of $Y_S$ that generalizes the model (1) by including both efficacy and toxicity indicators, $(Y_E, Y_T)$, to characterize early outcome. After phase I-II and an initial stage of phase III have been completed, the phase I-II/III design may re-optimize the dose of the experimental agent $A$ based on mean survival time, $\mu_S$. This approach hybridizes the phase I-II → phase III paradigm, in which dose-finding for $A$ is done using $(Y_E, Y_T)$ in phase I-II, rather than using only $Y_T$ for dose-finding as in the more conventional phase I → phase II → phase III paradigm.

Our proposed phase I-II/III design has $K \geq 3$ stages. In stage 1, a phase I-II trial is conducted based on the short-term binary indicators $(Y_E, Y_T)$, including adaptive randomization (AR) among doses of $A$ based on the dose desirability criterion $\phi$. The use of AR reduces the risk of getting stuck at a suboptimal dose in phase I-II. It addresses the "exploration versus exploitation" or "stickiness" problem, which is well known in sequential analysis (Sutton and Barto, 1998; Azriel, et al., 2010). AR improves the reliability of our proposed phase I-II/III design because it obtains more data on doses that may be sub-optimal in terms of the phase I-II criterion $\phi$ based on $(Y_E, Y_T)$ but optimal in terms of $\mu_S$.

Denote $A$ given at $x$ by $A(x)$. At the end of phase I-II (stage 1), an optimal dose $\hat{x}_{ET}^{\text{opt}}$ of $A$ based on $\phi$ is determined. In stage 2, phase III begins with patients randomized fairly to $C$ and $A(\hat{x}_{ET}^{\text{opt}})$. Phase I-II patients are followed to observe their times of death or follow up. After a pre-specified number of deaths, $n_2^*$, have been observed from patients receiving $C$ or $A(\hat{x}_{ET}^{\text{opt}})$ in stage 2, all $(Y_E, Y_T)$ and survival data of patients treated with $A$ in stages 1 and 2 are used to determine an optimal dose $\hat{x}_S^{\text{opt}}$ such that $A(\hat{x}_S^{\text{opt}})$ maximizes $\mu_S$, for use in the rest of phase III. The re-optimized dose $\hat{x}_S^{\text{opt}}$ may or may not be the same as $\hat{x}_{ET}^{\text{opt}}$. Stages 3,…,K are a randomized group sequential trial with up to $K - 2$ tests comparing the mean survival times of $A(\hat{x}_S^{\text{opt}})$ versus $C$. To provide a concrete illustration, for stage 1 we use the Eff-Tox phase I-II design of Thall et al. (2014), extended to include AR.

The rest of the paper is organized as follows. In Section 2, the data structure, models, and decision criteria are presented. Section 3 presents details of trial conduct. Section 4 describes possible decisions, outcomes, and potential consequences of re-optimizing dose versus the conventional approach of using $\hat{x}_{ET}^{\text{opt}}$ in phase III. Section 5 presents results of simulation study to compare the phase I-II/III design to the phase I-II → phase III paradigm. Section 6 concludes with a discussion. A computer program to implement the phase I-II/III design is available on CRAN in the package *Phase123*.

## 2 | DATA STRUCTURE, MODELS, AND DECISION CRITERIA

Given raw doses $\boldsymbol{d} = (d_1, \cdots, d_J)$ of the experimental agent $A$, denote the standardized doses by $x_j = (d_j - \overline{\boldsymbol{d}})/sd(\boldsymbol{d})$ for $j = 1, \ldots, J$. Let $Y_S^o$ denote the observed time to death or administrative censoring and $\delta = I(Y_S = Y_S^o)$. Denote the parameters for the distribution of $[Y_E, Y_T|x]$ by $\boldsymbol{\theta}_{ET}$, and the parameters for the distribution of $[Y_S|Y_E, Y_T, x]$ by $\boldsymbol{\theta}_S$.

The Eff-Tox design is reviewed in Web Appendix Section A. Briefly, for each $m = E, T$, and $x_j$, it is assumed that $\pi_m(x_j, \boldsymbol{\theta}_{ET}) = P(Y_m = 1 \mid x_j, \boldsymbol{\theta}_{ET}) = logit^{-1}\{\eta_m(x_j, \boldsymbol{\theta}_{ET})\}$, with $\eta_T(x_j, \boldsymbol{\theta}_{ET}) = \tau_{T,1} + \tau_{T,2}x_j$ and $\eta_E(x_j, \boldsymbol{\theta}_{ET}) = \tau_{E,1} + \tau_{E,2}x_j + \tau_{E,3}x_j^2$, with $\tau_{T,2} > 0$, so that $\pi_T(x_j, \boldsymbol{\theta}_{ET})$ increases with $x_j$, but $\pi_E(x_j, \boldsymbol{\theta}_{ET})$ may be non-monotone. An association parameter $\psi$ determines the joint distribution of $(Y_E, Y_T)$ from their marginals using a copula, so $\boldsymbol{\theta}_{ET} = (\tau_{T,1}, \tau_{T,2}, \tau_{E,1}, \tau_{E,2}, \tau_{E,3}, \psi)$. These parameters are assumed to be independent with priors $\psi \sim N(0, 1)$, $\tau_{E,2} \sim N(0, .20)$, and $\tau_{m,r} \sim N(\widetilde{\mu}_{m,r}, \widetilde{\sigma}_{m,r}^2)$ for $m = E, T$ and $r = 1, 2$. Numerical values of $(\widetilde{\mu}_{m,r}, \widetilde{\sigma}_{m,r})$ for $m = E, T, r = 1, 2$ are determined from elicited means of $\pi_m(x_j, \boldsymbol{\theta}_{ET})$, for $j = 1, \cdots, J$, $m = E, T$, and a desired prior effective sample size. Adaptive dose-finding decisions are based on a trade-off function $\phi(\pi_E, \pi_T)$ for $\boldsymbol{\pi} \in [0, 1]^2$.

Denote $\phi_j = \phi[E\{\pi_E(x_j, \boldsymbol{\theta}_{ET}), \pi_T(x_j, \boldsymbol{\theta}_{ET})\}|data]$ for each dose $x_j$ at any point during phase I-II based on the current data. The estimated optimal dose in an Eff-Tox trial is $\hat{x}_{ET}^{\text{opt}} = \text{argmax}_{x_j}\{\phi_j\}$. To extend this design to include AR, rather than choosing $\hat{x}_{ET}^{\text{opt}}$ for each cohort during phase I-II, we adaptively randomize the next cohort to dose $x_j$ with probability

$$\frac{\exp\{(\phi_j - \overline{Q})/sd(\boldsymbol{Q})\}}{\sum_{r:\,\phi_r \in \boldsymbol{Q}} \exp\{(\phi_r - \overline{Q})/sd(\boldsymbol{Q})\}},$$

where $\boldsymbol{Q}$ is the current set of posterior mean desirabilities, $\phi_j$, of doses that are acceptably safe and efficacious. This shrinks the selection probability of less desirable doses toward 0 while allowing selection of doses that are suboptimal in terms of $\phi$. After the $(Y_E, Y_T)$ outcomes of all $N_{ET}$ patients in phase I-II have been evaluated, $A(\hat{x}_{ET}^{\text{opt}})$ based on the final phase I-II data is moved forward to stage 2, which is the first portion of phase III.

In the phase I-II/III design, we define two different types of truly optimal doses of $A$. Let $\boldsymbol{\theta}_m^{true}$ denote an assumed true value of $\boldsymbol{\theta}_m$ for $m = ET$ or $S$. The truly optimal dose that maximizes $\phi\{\pi_E(x, \boldsymbol{\theta}_{ET}^{true}), \pi_T(x, \boldsymbol{\theta}_{ET}^{true})\}$ is $x_{ET}^{\text{opt}}$. The truly optimal dose that maximizes the mean survival time $\mu_S(x, \boldsymbol{\theta}_S^{true})$ is $x_S^{\text{opt}}$. Let $k = 1, \ldots, K$ index the stages of the phase I-II/III trial. Thus, $k = 1$ indexes the phase I-II trial, $k = 2$ indexes the first portion of phase III at the end of which the dose of $A$ may be re-optimized based on $\mu_S$, and $k = 3, \ldots, K$ index the subsequent group sequential stages in phase III for comparing $A(\hat{x}_S^{\text{opt}})$ to $C$. Thus, there are up to $K - 2$ group sequential comparisons in phase III. Let $\mathcal{D}_{I-II,k}$ and $\mathcal{D}_{III,k}$ denote the data for patients at the end of stage $k$, from the phase I-II and phase III portions of the trial, respectively. Therefore, $\mathcal{D}_{I-II,1}$ consists only of the $(x, Y_E, Y_T)$ data from phase I-II patients, while $\mathcal{D}_{I-II,2}$ also includes these patients' survival time data $(Y_S^o, \delta)$, up to the time at which the decision of whether to switch the dose based on mean survival time is made. $\mathcal{D}_{III,1}$ does not exist because phase III has not begun in stage 1. The re-optimized dose $\hat{x}_S^{opt}$ is chosen based on $\mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}$, which includes all $(x, Y_E, Y_T)$ and $(Y_S^0, \delta)$ data at the end of stage 2.

Since the phase I-II → phase III paradigm uses $\hat{x}_{ET}^{\text{opt}}$ throughout phase III, the primary motivation for our design is the possibility that $x_{ET}^{\text{opt}} \neq x_S^{\text{opt}}$, and that re-optimizing the dose of $A$ may produce larger $\mu_S$ by comparing $C$ to $A(\hat{x}_S^{\text{opt}})$ rather than $A(\hat{x}_{ET}^{\text{opt}})$ in the group sequential trial. To evaluate the effects of re-optimizing the dose of $A$ based on $\mu_S$ during the first part of phase III, we require models for $[Y_E, Y_T|x]$ and $[Y_S|Y_E, Y_T, x]$, in order to formulate a mixture model for $[Y_S|x]$. This will include the effects of $x$ on the indicators $(Y_E, Y_T)$, and the effects of $(Y_E, Y_T)$ and $x$ on the hazard function of $Y_S$. Let $\pi(y_E, y_T \mid x, \boldsymbol{\theta}_{ET})$ denote the probability distribution of $(Y_E, Y_T)$ at dose $x$, where $(y_E, y_T) \in \{0, 1\}$. Let $f_{S|E,T}(y_S \mid y_E, y_T, x, \boldsymbol{\theta}_S)$ denote the conditional pdf of $Y_S$ given the early binary outcomes and dose $x$ of $A$. The mixture pdf of $Y_S$ for patients treated with $A(x)$ is

$$f_S(y_S \mid x, \boldsymbol{\theta}_S, \boldsymbol{\theta}_{ET}) = \sum_{y_E=0}^{1}\sum_{y_T=0}^{1} f_{S|E,T}(y_S \mid y_E, y_T, x, \boldsymbol{\theta}_S) \times \pi(y_E, y_T \mid x, \boldsymbol{\theta}_{ET}). \quad (2)$$

The conditional mean survival time given $(y_E, y_T)$ of a patient treated with $A(x)$ is

$$\mu_{S,A(x)}(y_E, y_T, \boldsymbol{\theta}_S) = \int_0^\infty y_S f_{S|E,T}(y_S \,|\, y_E, y_T, x, \boldsymbol{\theta}_S) dy_S. \quad (3)$$

At the end of stage 2 of phase I-II/III, we choose $\hat{x}_S^{\text{opt}}$ based on all observed data, where $\hat{x}_S^{\text{opt}}$ maximizes the posterior mean of the parametric mean survival time

$$\mu_{S,A(x)}(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{ET}) = \sum_{y_E=0}^{1}\sum_{y_T=0}^{1} \mu_{S,A(x)}(y_E, y_T, \boldsymbol{\theta}_S) \times \pi(y_E, y_T \mid x, \boldsymbol{\theta}_{ET}) \quad (4)$$

at $A(x)$. Conventionally $(Y_E, Y_T)$ are used as surrogates for $Y_S$ in choosing a dose $\hat{x}_{ET}^{\text{opt}}$ in phase I-II, but $(Y_E, Y_T)$ are ignored when modeling survival in phase III.

We assume that the distribution of $[Y_S \mid Y_E, Y_T, x]$ has the Cox type hazard function

$$h(t|Y_E, Y_T, x, \boldsymbol{\theta}_S) = h_0(t)\exp\{\beta_1 x + \beta_2 x^2 - e^{\beta_E} Y_E + e^{\beta_T} Y_T\}, \; t > 0. \quad (5)$$

We assume that $\beta_1$, $\beta_2$, $\beta_E$, and $\beta_T$ are independent with identical non-informative $N(0, 100)$ priors. For robustness, we assume that the baseline hazard is piecewise exponential with $h_0(t) = \exp(\lambda_l)$ for $t \in (t_l, t_{l+1}]$ under the partition $t_0 = 0 < t_1 < \ldots < t_{L+1} = \max(\boldsymbol{Y_S}^o)$. We allow the dimension $L$ of the baseline hazard to vary, with prior $L \sim Poi(\zeta_S)$ and assume that the locations of the split points $\boldsymbol{t}$ vary according to the even order statistics with a uniform distribution of size $2L$, as in Lee et al. (2015) and Chapple et al. (2017). This prevents obtaining intervals in $h_0(t)$ having few events for estimating $\lambda_l$. We suggest values of $\zeta_S \in \{3, 4, 5, 6, 7\}$, since most hazard shapes can be approximated very accurately with 1 to 5 pieces. The resulting posterior distribution is not sensitive to the choice of $\zeta_S$ in this range for sample sizes greater than 50. We assume a normal prior with mean 0 and variance 25 for $\lambda_1$, denoted $\lambda_1 \sim N(0, 25)$, and borrow strength when $L > 1$ for adjacent intervals via the prior $\lambda_l \sim N(\lambda_{l-1}, \sigma_\lambda^2)$, with the prior of $\sigma_\lambda$ proportional to $1/\sigma_\lambda$. The variance of $\lambda_1$ ensures posterior hazard values seen in practice, while maintaining prior non-informativeness.

Denote $\hat{\mu}_{S,A(x)} = E\left(Y_S | x, \mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}\right)$, the posterior mean survival time for $A(x)$ given the data from phases I-II and III at the end of stage 2. We compute this quantity under the mixture model (2) by estimating the posterior mean survival time

$$\hat{\mu}_{S,A(x)}(y_E, y_T | \mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}) = E\{\mu_{S,A(x)}(y_E, y_T, \boldsymbol{\theta}_S) \mid \mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}\}$$

for each pair $(y_E, y_T) \in \{0, 1\}$, under the formula (3), and computing the posterior mean

$$\hat{\pi}(y_E, y_T \mid x, \mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}) = E\{\pi(y_E, y_T \mid x, \boldsymbol{\theta}_{ET}) \mid \mathcal{D}_{I-II,2} \cup \mathcal{D}_{III,2}\}$$

of each bivariate probability under the Eff-Tox model given in Web Appendix A.

Since there will be limited survival time follow up information after $n_2^*$ events, the design only evaluates the means until the maximum observed patient follow up time. The trial is continued after $n_2^*$ patient events using the dose $\hat{x}_S^{\text{opt}}$ of $A$ having the highest posterior mean $\hat{\mu}_S^{\text{opt}} = \max_{x_j}\{\hat{\mu}_{S,A(x_j)}\}$. After making this decision, the design does not use data from patients who were treated at doses $x_j \neq \hat{x}_S^{\text{opt}}$. After obtaining values of $n_3^*, \ldots, n_K^*$ from East 6 statistical software (2016), $n_2^*$ is chosen such that the design can switch doses with high accuracy, but can still yield high power for phase III trials, given the truly optimal dose $x_S^{\text{opt}}$ has been selected. Suitable values of $n_2^*$ can be determined using the function SimPhase123 in the package *Phase123*. This approach will result in a larger sample size of patients in the $C$ arm being compared to $A(\hat{x}_S^{\text{opt}})$ if $\hat{x}_{ET}^{\text{opt}} \neq \hat{x}_S^{\text{opt}}$. We use Markov chain Monte Carlo to obtain posterior distributions for $\boldsymbol{\theta}_{ET}$ and $\boldsymbol{\theta}_S$, using 2000 iterations and 1000 discarded as burnin. This gives good convergence of the parameters, shown by the posterior of $L$ settling on one or two values as well as traceplots for the parameters $\boldsymbol{\lambda}|L, \boldsymbol{s}|L$, and the coefficients in the linear terms of the Eff-Tox and survival hazard models. A detailed account of computational algorithms used to simulate posterior samples is given in Web Appendix B.

## 3 | TRIAL CONDUCT

In this section, we give specific rules for conducting a phase I-II/III clinical trial. Each of the computer functions described below is contained in the R package *Phase*123, available on CRAN, including documentation of inputs and examples. Additional information on the trial parameters is given in Web Appendix C and a tutorial on several of the functions is given in Web Appendix D. When designing a phase I-II/III trial, the statistician should consult with the physician to establish design parameters, such as $\phi$, maximum sample sizes $N_{ET}$ and $N_S$, and the number of comparative tests $K - 2$ following dose re-optimization. The group sequential boundaries for stopping the trial due to futility $\underline{u}_k$ or superiority $\overline{u}_k$ may be obtained using East 6 statistical software (2016), specifying a null value of $\mu_C$, desired improvement $\Delta$, type I error, power under the alternative, maximum sample size $N_S$, and information proportions for determining $n_k^*$ for $k = 3, \ldots, K$. If no futility decision is desired at look $k$ then $\underline{u}_k = 0$. The information proportions used to determine $n_k^*$ should be large enough ($> 30\%$) to avoid making unreliable decisions based on a small amount of patient data if a dose is re-optimized for $A$.

The phase I-II/III design parameters must be calibrated to obtain good operating characteristics (OCs) under a reasonable array of possible scenarios. A smaller value of $n_2^*$ ($\leq 20\%$

of the total information proportion) may be obtained by simulating the phase I-II/III trial under sets of different (a) Eff-Tox scenarios quantifying effects of $x$ on $(Y_E, Y_T)$, (b) effects of $(x, Y_E, Y_T)$ on survival, and (c) survival distributions. The stage 2 sample size $n_2^*$ should be set by examining the design's OCs for several different values, to find $n_2^*$ (1) large enough to give a high probability of selecting the optimal dose, but (2) small enough so, given that the design switches to a true optimal dose in stage 2, it has good generalized power figures. This can be done using the function $SimPhase123$. Specific rules for conducting a phase I-II/III trial are as follows:

**(1)** Enroll the first cohort of patients in the phase I-II portion at the lowest dose. For each subsequent cohort until $N^F$ patients have been treated, use the function $AssignEffTox$ to obtain the next dose to give.
**(2)** Once $N^F$ patients have been enrolled in phase I-II, use the function $RandomEffTox$ to adaptively randomize the next cohort of patients among acceptable doses, which allows doses that are empirically suboptimal in terms of $\phi(\pi_E, \pi_T)$ to be chosen.
**(3)** After $N_{ET}$ patients have been enrolled in phase I-II and their efficacy and toxicity outcomes have been evaluated, use the function $AssignEffTox$ to obtain the dose $\hat{x}_{ET}^{\text{opt}}$ to continue to phase III.
**(4)** Start phase III, randomizing patients equally between $C$ and $A(\hat{x}_{ET}^{\text{opt}})$.
**(5)** After $n_2^*$ deaths have been observed, use the function $Reoptimize$ to determine the dose $\hat{x}_S^{\text{opt}}$ to continue with for the remainder of the trial.
**(6)** Remove any patients treated with $\hat{x}_{ET}^{\text{opt}}$ from consideration if the dose was switched and begin randomizing patients between $C$ and $A(\hat{x}_S^{\text{opt}})$.
**(7)** For each stage $k = 3, \ldots, K$, after $n_k^*$ deaths occur, do two-sided tests for superiority or futility using the logrank test in R. Denoting the Z-score corresponding to the logrank statistic by $Z$, for futility bound $\underline{u}_k$ and superiority bound $\overline{u}_k$, stop the trial if

$$|Z| > \overline{u}_k \text{ for superiority} \quad \text{or} \quad |Z| < \underline{u}_k \text{ for futility.}$$

**(8)** Stop accrual after $N_S$ patients have been enrolled in the phase III portion, including patients treated with a dose that is no longer considered optimal.

## 4 | POSSIBLE TRIAL OUTCOMES

Before presenting our simulation results, we discuss possible design decisions and comment on each under different true states of nature. Because the phase I-II/III design may change the phase I-II selected dose of $A$ in phase III before comparing $A$ to $C$, the sequence of decisions that it makes may be correct and optimal, correct but suboptimal, wrong, or disastrously wrong, depends on $x_S^{\text{opt}}$ and $x_{ET}^{\text{opt}}$, their estimates, and whether $\mu_{A(x_S^{\text{opt}})} = \mu_C$ or $\mu_{A(x_S^{\text{opt}})} \geq \mu_C + \Delta$. Since more than one dose of $A$ may provide the desired improvement in $\mu_S$ of at least $\Delta$ over $C$, we denote the set of all such doses by $X^{\text{opt}} = \{x_j : \mu_{A(x_j)} \geq \mu_C + \Delta\}$. We define the generalized power (GP) to be the probability of (1) selecting a dose $x_j \in X^{\text{opt}}$ in stage 2 and (2) declaring $A(x_j)$ superior to $C$ in one of stages $3, \cdots, K$. The GP is the sum over $x_j \in X^{\text{opt}}$ of the probability of selecting $x_j$ and declaring $A(x_j)$ superior to $C$. If $X^{\text{opt}}$ contains more than one dose, then the GP is larger than the probability of the best possible decision, which is to select the optimal dose $x_S^{\text{opt}} \in X^{\text{opt}}$ that maximizes $\mu_S$ with $A$ and declare $A(x_S^{\text{opt}})$ superior to $C$. We denote the probability of making this best decision by $\gamma_1$ and the GP by $\gamma_2$. Thus, $\gamma_1 \leq \gamma_2$, with $\gamma_1 = \gamma_2$ if $X^{\text{opt}}$ contains exactly one dose, which in this case must be $x_S^{\text{opt}}$.

To help sort this out, Table 1 provides explanatory comments on scenarios in stages $k = 1$ (phase I-II) and $k = 2$ (the first portion of phase III) regarding the true relationship between the optimal doses $x_S^{\text{opt}}$ and $x_{ET}^{\text{opt}}$ and their posterior estimates $\hat{x}_S^{\text{opt}}$ and $\hat{x}_{ET}^{\text{opt}}$. If $\hat{x}_{ET}^{\text{opt}} = \hat{x}_S^{\text{opt}}$, then the phase I-II/III and phase I-II → phase III designs make equivalent decisions. However, if $\hat{x}_{ET}^{\text{opt}} \neq \hat{x}_S^{\text{opt}}$, then switching provides a potential advantage. In this case survival data from patients who were treated with $\hat{x}_{ET}^{\text{opt}}$ during phase III are no longer relevant. Depending on the accrual rate, maximum sample size $N_S$, and number of patient events $n_2^*$ needed to re-optimize dose, this may result in 30 to 100 patients being treated at doses no longer considered a part of the trial as phase III proceeds.

After choosing $\hat{x}_S^{\text{opt}}$, the phase I-II/III design makes group sequential decisions comparing $A(\hat{x}_S^{\text{opt}})$ to $C$, so the decisions in phase III depend on the selected $\hat{x}_S^{\text{opt}}$. But it may not be the case that $\hat{x}_S^{\text{opt}} = x_S^{\text{opt}}$. That is, the design may not choose the truly optimal dose in terms of mean survival time in stage 2. Table 2 lists possible decisions of a phase I-II/III design in stages $k = 2, \ldots, K$ and how each decision may be viewed in terms of $x_S^{\text{opt}}$. Table 2 is ordered with the best outcomes listed first and the worst listed last, with outcomes 1, 2, and 3 being good and outcomes 4 and 5 being bad. In outcome 1, the design declares the dose that increases $\mu_S$ the most to be superior to $C$. In outcome 2, a dose of $A$ is selected that provides a clinically meaningful improvement $\geq \Delta$ in $\mu_S$ compared to $C$, but the best dose of $A$ is not chosen, so the decision is correct but the dose has not been truly optimized. Outcome 3 represents a correct decision, but it does not improve $\mu_S$ since it declares $C$ superior to or equivalent to $A(\hat{x}_S^{\text{opt}})$. Outcome 4 gives a false positive result, including cases where the design wrongly chooses an inferior dose for which $\mu_{A(x_j)} < \mu_C$, which is worse than a conventional type I error. Outcome 5 represents the worst possible case, since

**TABLE 1** Possible relationships between $\hat{x}_{ET}^{\text{opt}}$, $x_{ET}^{\text{opt}}$, $\hat{x}_{S}^{\text{opt}}$, and $x_{S}^{\text{opt}}$, including comments related to the phase I-II → phase III and phase I-II/III designs

| $\hat{x}_{ET}^{\text{opt}} \stackrel{?}{=} x_{ET}^{\text{opt}}$ | $x_{ET}^{\text{opt}} \stackrel{?}{=} x_{S}^{\text{opt}}$ | Comments |
|---|---|---|
| $\hat{x}_{ET}^{\text{opt}} = x_{ET}^{\text{opt}}$ | $x_{ET}^{\text{opt}} = x_{S}^{\text{opt}}$ | The optimal dose in terms of $\mu_S$ was selected in phase I-II, so it is not desirable to switch doses at stage 2. In this scenario, the phase I-II/III design cannot provide an improvement over phase I-II → III. |
| $\hat{x}_{ET}^{\text{opt}} = x_{ET}^{\text{opt}}$ | $x_{ET}^{\text{opt}} \neq x_{S}^{\text{opt}}$ | The dose selected in phase I-II is optimal in terms of $\phi$ but is not optimal in terms of $\mu_S$. This illustrates the advantage of the phase I-II/III design over phase I-II → III design. |
| $\hat{x}_{ET}^{\text{opt}} \neq x_{ET}^{\text{opt}}$ | $x_{ET}^{\text{opt}} = x_{S}^{\text{opt}}$ | The dose selected in phase I-II is suboptimal based on $\phi$, but the optimal doses in terms of $\phi$ and $\mu_S$ are identical. This scenario illustrates the advantage of the phase I-II/III design over phase I-II → III design. |
| $\hat{x}_{ET}^{\text{opt}} \neq x_{ET}^{\text{opt}}$ | $x_{ET}^{\text{opt}} \neq x_{S}^{\text{opt}}$ | The dose selected in phase I-II is suboptimal based on $\phi$, but the optimal doses in terms of $\phi$ and $\mu_S$ are not identical. This scenario illustrates the advantage of the phase I-II/III design over phase I-II → III design. |

Note: $\hat{x}_{ET}^{\text{opt}} \stackrel{?}{=} x_{ET}^{\text{opt}}$ refers to whether or not the optimal dose is selected at the end of phase I-II based on $\phi$, and $x_{ET}^{\text{opt}} \stackrel{?}{=} x_{S}^{\text{opt}}$ refers to whether the optimal dose based on $\mu_S$ is the same as that based on $\phi$.

not only does the design wrongly conclude that the chosen dose gives $A(\hat{x}_S^{\text{opt}})$ superior to $C$, but it might have obtained a successful trial result if it had correctly selected $x_S^{\text{opt}}$ in stage 2.

These same decisions and interpretations are made in the conventional phase I-II → phase III paradigm, with the difference that $\hat{x}_S^{\text{opt}}$ is replaced with $\hat{x}_{ET}^{\text{opt}}$. Compared to this conventional design, allowing the optimal dose to be switched in the phase I-II/III design makes selecting $\hat{x}_S^{\text{opt}} = x_S^{\text{opt}}$ more likely, which increases the probabilities of outcomes 1 and 2 and decreases the probabilities of the disastrous outcome 5. Under outcome 3, the phase I-II/III design is likely to treat more patients because it is more likely to correctly pick the dose $x_S^{\text{opt}}$ having the largest $\mu_S$, thus making stopping the trial early for superiority of $C$ or futility less likely. It will be more likely to switch to the dose having the longest mean survival time for outcome 4, however, which makes a false positive event more likely.

## 5 | SIMULATION STUDY

To perform a simulation study comparing the phase I-II/III design to the phase I-II → phase III paradigm, we first specify three different Eff-Tox scenarios, consisting of true efficacy and toxicity dose-probability vectors. We will use these to specify different relationships between $(Y_E, Y_T)$ and $Y_S$. We evaluate the design with $J = 5$ doses using raw dose values $(d_1, \cdots, d_5) = (1, 2, 3, 3.5, 5)$. For this study, each patient's $(Y_E, Y_T)$ are evaluated in one month, and we assume for simplicity that no patients die before this month long window. For a dose $x_j$ chosen in phase I-II, we test the null hypothesis $H_0 : \mu_C = \mu_{A(x_j)} = 24$ months versus $H_0 : \mu_C \neq \mu_{A(x_j)}$ with target $\mu_{A(x_S^{opt})} = 36$ months, a $\Delta = 12$ month improvement.

To implement phase I-II using the Eff-Tox design, the three equivalent $(\pi_E, \pi_T)$ pairs used to establish the desirability function $\phi$ were $(0.35, 0)$, $(0.70, 0.40)$ and $(1, 0.75)$. The contour created by these three pairs is seen in Web Figure 1.

**TABLE 2** Possible phase I-II/III trial outcomes, $O$

| $O$ | Decision | Truth | Comments |
|---|---|---|---|
| 1 | $\hat{x}_S^{\text{opt}} = x_S^{\text{opt}}$<br>$A(\hat{x}_S^{\text{opt}}) \succ C$ | $\mu_{A(x_S^{\text{opt}})} > \mu_C + \Delta$ | This is the generalized power event at the optimal dose $x_S^{\text{opt}}$. The design correctly selects $x_S^{\text{opt}}$ as optimal and declares $A(x_S^{\text{opt}})$ superior to $C$. |
| 2 | $\hat{x}_S^{\text{opt}} \neq x_S^{\text{opt}}$<br>$A(\hat{x}_S^{\text{opt}}) \succ C$ | $\mu_{A(\hat{x}_S^{\text{opt}})} > \mu_C + \Delta$<br>$\mu_{A(\hat{x}_S^{\text{opt}})} > \mu_{A(x_S^{\text{opt}})}$ | This is a generalized power event in a case where the design correctly concludes $A(\hat{x}_S^{\text{opt}})$ is superior to $C$ but $\hat{x}_S^{\text{opt}}$ is suboptimal, so it could have improved survival more had if it chosen the truly optimal dose $x_S^{\text{opt}}$. |
| 3 | $\hat{x}_S^{\text{opt}} = \text{any } x_j$<br>$C \succeq A(\hat{x}_S^{\text{opt}})$ | $\mu_{A(x_S^{\text{opt}})} \leq \mu_C$ | This is a correct conclusion, but the phase I-II/III design will require an increased sample size compared to the phase I-II → III design due to correctly switching to $x_S^{\text{opt}}$. |
| 4 | $\hat{x}_S^{\text{opt}} = \text{any } x_j$<br>$A(\hat{x}_S^{\text{opt}}) \succ C$ | $\mu_{A(x_S^{\text{opt}})} \leq \mu_C$ | This is a false positive conclusion. While the design may pick the best dose of $A$, it incorrectly concludes that $A$ at that dose is superior to $C$. |
| 5 | $\hat{x}_S^{\text{opt}} \neq x_S^{\text{opt}}$<br>$A(\hat{x}_S^{\text{opt}}) \succ C$ | $\mu_{A(\hat{x}_S^{\text{opt}})} \leq \mu_C$<br>$\mu_{A(x_S^{\text{opt}})} \geq \mu_C + \Delta$ | This is a disastrous false negative conclusion. The design chooses a suboptimal dose based on $\mu_S$ and incorrectly concludes $A(\hat{x}_S^{\text{opt}})$ is inferior to $C$, instead of correctly selecting $x_S^{\text{opt}}$ and declaring $A(x_S^{\text{opt}})$ superior to $C$. |

Note: $x_S^{\text{opt}}$ is the truly optimal dose in terms of $\mu_{A(x)}$ and $\Delta$ is the desired improvement over $\mu_C$. Column 2 gives the two trial decisions, the first row for selecting $\hat{x}_s^{\text{opt}}$ and the second row for determining superiority, inferiority, or futility, with $A(\hat{x}_S^{\text{opt}}) \succ C$ indicating that $A(\hat{x}_S^{\text{opt}})$ is declared superior to $C$, and $C \succeq A(\hat{x}_S^{\text{opt}})$ indicating that the trial is stopped due to either superiority of $C$ or futility.

The upper limit on $\pi_T$ was $\overline{\pi}_T = 0.40$ and the lower limit on $\pi_E$ was $\underline{\pi}_E = 0.30$. The threshold on the posterior probability that $\pi_E > 0.30$ and $\pi_T < 0.40$ was set to be $p_E = p_T = 0.10$ for both acceptability rules. Patients were treated in cohorts of size 3, with up to $N_{ET} = 60$ patients enrolled in phase I-II (stage 1). We calibrated the phase I-II hyperparameters to have prior effective sample size .90 as suggested by Yuan et al. (2016). We used prior mean toxicity probabilities of $(0.05, 0.10, 0.15, 0.20, 0.30)$ and mean efficacy probabilities (.20, .40, .60, .65, .70) for the five doses, to produce the hyperparameter means $(-4.23, 3.1, .02, 3.45, 0, 0)$ and standard deviations $(3.13, 3.12, 2.68, 2.69, 0.2, 1)$ for the prior of $\boldsymbol{\theta}_{ET}$. The *EffTox* program is freely available on the MDAnderson biostatistics software page.

For the phase I-II portion of the simulated trials, patients were treated in cohorts of size three and assigned doses after the previous cohort was fully evaluated, assuming an accrual rate of five patients per month with adaptive randomization begun after $N^F = 15$ patients. The three simulation scenarios' assumed true $\pi_E(x_j)$ and $\pi_T(x_j)$ are given in Table 3, with their selection percentages, true $\phi$ values, and numbers of patients treated, based on 5000 simulated trials using the *EffTox* program.

In the three Eff-Tox scenarios in Table 3, the respective optimal doses in terms of the tradeoff contour are doses = 3, 5, and 2. We only consider simulated phase I-II trials that advance to phase III, ignoring simulation replications where the trial stopped early. In scenario 1, doses 3 and 4 have nearly equivalent desirability, so we expect most patients in the phase I-II portion of the phase I-II/III trial to be treated at these two doses. In scenario 2, the highest dose 5 is considered optimal, most patients are treated at this dose, and it is selected in 49% of the simulations. In scenario 3, the dose 2 is optimal and doses 4 and 5 have unacceptably high toxicity probabilities, so we expect to treat fewer patients at these doses. The design treats the most patients at dose 2, which is selected with probability 0.51, but substantial numbers of patients are treated at doses 1 and 3. The use of AR assigns more patients to doses 1 and 3, which allows the phase I-II /III design to better assess the functional relationship between dose and mean survival time. For the control group, we assume that the effects of toxicity and efficacy on overall survival are the same as those for the experimental group, and set the probabilities of toxicity and efficacy to be $(0.15, 0.40)$, $(0.10, 0.30)$, and $(0.20, 0.35)$ for the three Eff-Tox scenarios, respectively. For each of these simulation scenarios, we assume two different forms for the linear terms of the log hazard of $Y_S$. For $A(x)$, we assume $\eta_S(x, Y_E, Y_T) = \beta_0 + \beta_1 x + \beta_2 x^2 - \exp(\beta_E)Y_E + \exp(\beta_T)Y_T$. For the simulated data from the control group, we assume $\eta_S(C, Y_E, Y_T) = \beta_C - \exp(\beta_E)Y_E + \exp(\beta_T)Y_T$ and calibrate the additional parameter $\beta_C$ so that we obtain the desired null value of 24 months for mean survival time. We first consider an exponential distribution with pdf $f(t|\rho) = (1/\rho)\exp(-t/\rho)$ where $\rho = \exp\{\eta_S(x, Y_E, Y_T)\}$, since the O'Brien Fleming group sequential bounds (O'Brien and Fleming, 1979) for the logrank test are based on this assumption. Later, we will consider several other distributions to evaluate the robustness of the methodology. Table 4 displays the six scenarios considered, which correspond to the different Eff-Tox scenarios listed in Table 3, as well as differing effects of dose, efficacy, and toxicity on survival time.

These scenarios encompass several qualitatively and quantitatively different possible cases in connecting phase I-II to phase III. In scenario 1, the optimal dose in terms of $\mu_S$ is dose 3, which is selected with probability 0.29. In scenario 2, there is a large efficacy effect, leading to dose 5 being optimal in terms of $\mu_S$, but this dose is only selected with probability 0.49 in phase I-II. Thus, we expect to see a large improvement in this scenario by using a phase I-II/III design. Similarly, in scenario 3, dose 3 is optimal in terms of $\mu_S$, but is only selected with probability 0.18 in phase I-II. Scenario 4 represents a case with a large toxicity effect and small efficacy effect, making dose 1 optimal in terms of $\mu_S$, but dose

**TABLE 3** Eff-Tox scenarios

| Scenario | Value | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | $(\pi_E, \pi_T)^{TR}$ | (0.20, 0.10) | (0.40, 0.15) | (0.60, 0.25) | (0.65, 0.35) | (0.70, 0.50) |
| | $\phi\{(\pi_E, \pi_T)^{TR}\}$ | −0.37 | −0.13 | 0.05 | −0.01 | −0.13 |
| | % Selected | 3 | 26 | 29 | 27 | 13 |
| | # Treated | 6.4 | 16.0 | 16.1 | 12.1 | 9.0 |
| 2 | $(\pi_E, \pi_T)^{TR}$ | (0.2, 0.05) | (0.25, 0.08) | (0.35, 0.10) | (0.40, 0.15) | (0.55, 0.20) |
| | $\phi\{(\pi_E, \pi_T)^{TR}\}$ | −0.30 | −0.26 | −0.14 | −0.13 | 0.04 |
| | % Selected | 5 | 11 | 18 | 16 | 49 |
| | # Treated | 8.4 | 9.1 | 10.1 | 9.8 | 22.4 |
| 3 | $(\pi_E, \pi_T)^{TR}$ | (0.40, 0.10) | (0.50, 0.15) | (0.60, 0.35) | (0.65, 0.60) | (0.70, 0.70) |
| | $\phi\{(\pi_E, \pi_T)^{TR}\}$ | −0.06 | 0.03 | −0.09 | −0.35 | −0.40 |
| | % Selected | 26 | 51 | 20 | 2 | 0 |
| | # Treated | 16.7 | 27.3 | 11.6 | 3.3 | 0.8 |

Note: True outcome probabilities $(\pi_E, \pi_T)^{TR}$, desirabilities $\phi\{(\pi_E, \pi_T)^{TR}\}$, and operating characteristics for the usual (non-adaptively randomized) EffTox phase I-II trial design.

**TABLE 4** Simulation parameters

| Scenario | Eff-Tox Scen | Hyp | $(\beta_1, \beta_2, e^{\beta_E}, e^{\beta_T}, \beta_0)^{TR}$ | $(\mu_{A(x_1)}, \mu_{A(x_2)}, \mu_{A(x_3)}, \mu_{A(x_4)}, \mu_{A(x_5)})^{TR}$ |
|---|---|---|---|---|
| 1 | 1 | Null | (0.1, −0.5, 0.5, 0.5, 2.9) | (8.3, 17.9, 24, 22.5, 9.8) |
| | | Alt | (0.25, −2, 0.5, 0.5, 3.4) | (1, 14.5, 36.2, 28.3, 1) |
| 2 | 2 | Null | (0.1, −0.1, 1, 0.5, 2.6) | (14.0, 17.8, 21.9, 23, 24) |
| | | Alt | (0.5, 0, 1, 0.5, 2.3) | (7.1, 10.3, 16.0, 19.5, 36) |
| 3 | 2 | Null | (0.1, −0.5, 0.3, 1, 3.1) | (9.5, 18.5, 24, 22.5, 10.4) |
| | | Alt | (0.1, −1, 0.3, 1, 3.6) | (6.9, 24.7, 38, 33.1, 6.3) |
| 4 | 3 | Null | (−0.3, 0.3, 0.3, 1, 2.3) | (24,13.6, 8.9, 6.8,7.8) |
| | | Alt | (−0.1, 0.3, 0.3, 1, 3.0) | (38, 24.6, 18.4, 15.0, 21.1) |
| 5 | 3 | Null | (0.1, −0.5, 0.3, 0.1, 3.0) | (9.3, 18.7, 24, 22.7, 10.4) |
| | | Alt | (0.1, −1, 0.3, 0.1, 3.6) | (7.8, 28.8, 44, 38.6, 7.4) |
| 6 | 1 | Null | (0.75, −0.5, 0.3, 0.25, 2.8) | (3.2, 10.1, 20.4, 24, 20.4) |
| | | Alt | (1, −0.6, 0.3, 0.25, 3.3) | (3.0, 12.9, 31.8, 40, 36.6) |

Note: True survival parameters $\boldsymbol{\beta}^{TR}$ corresponding to the phase I-II scenarios in Table 3, and true mean survival time of each $A(x_1), \cdots, A(x_5)$ for each phase I-II/III scenario's null and alternative hypotheses.

1 is only chosen in 26% of the usual phase I-II trials. In scenario 5, dose 3 is the third best dose in terms of $\phi$, but is best in terms of $\mu_S$, and it is only selected with probability 0.20. In this scenario, dose 4 also gives a significant improvement in $\mu_S$ compared to $C$, with $\mu_{A(x_4)} = 38.6$ months. In scenario 6, there is a large efficacy effect on overall survival, making dose 4 optimal in terms of both overall survival and $\phi$, but dose 5 also has significantly improved survival compared to $C$. These two scenarios provide a basis for evaluating improvements in both $\gamma_1$ and the GP, $\gamma_2$. To control the possibility of incorrectly switching due to chance outcomes, we do not allow the design to continue with a dose that had less than 6 patients treated. These scenarios also have varying effects of toxicity and efficacy on $h_S$, quantified by the coefficients $\beta_E$ and $\beta_T$. This will evaluate the sensitivity of the method to these effects. Since the parameters $(\beta_1, \beta_2)$ must be changed substantially to obtain similar $\mu_S$ values for different values of $(\beta_E, \beta_T)$, we do not perform a sensitivity analysis to these parameters within each scenario.

We assume that 10 patients, on average, are accrued each month during phase III, and that the phase III trial will begin 1 month after the phase I-II trial concludes. This waiting time could be increased to obtain longer survival follow up and thus improve the design's ability to re-optimize doses during stage $k = 2$. We enroll a maximum of $N_S = 500$ patients in phase III, which has up to three interim looks after $n_3^* = 200$, $n_4^* = 300$ and $n_5^* = 400$ deaths, with superiority decisions possible at each. We calibrated the stopping boundaries with East 6 statistical software (2016) using O'Brien-Fleming bounds (O'Brien and Fleming, 1979) with power 0.80 and type I error probability 0.05. We included a rule to determine if the trial should be stopped for futility, that is, neither $C$ nor $A(\hat{x}_S^{\text{opt}})$ is superior, after $n_4^* = 300$ deaths. The boundaries for declaring superiority of $A(\hat{x}_S^{opt})$ or $C$ based on the standardized logrank statistics are $(\overline{u}_3, \overline{u}_4, \overline{u}_5) = (2.96, 2.53, 1.99)$, and the futility bound at the second look is $\underline{u}_4 = 1.001$. At the start of the phase III portion of the trial, we begin randomizing patients equally to $A(\hat{x}_{ET}^{\text{opt}})$ and $C$. After $n_2^* = 50$ deaths in the trial have occurred, we determine the dose $\hat{x}_S^{\text{opt}}$ that patients receiving $A$ should receive for the remainder of the trial. This is the re-optimization step. Survival times for patients in phase I-II and phase III are generated after their toxicity and efficacy are scored, which does not allow the possibility that a patient may die before their short-term indicators are seen.

For each scenario and design, trial replications were simulated. The simulation results are summarized in Table 5. In each of scenarios 1-4, $\gamma_1 = \gamma_2$, since there is one dose for which $A$ is superior to $C$. Mean improvement in patient survival time with each design is denoted by $\overline{W}$, computed by averaging the differences between the true mean survival time with the selected dose of $A$ and $\mu_C$, if $A$ is declared superior to $C$. In the simulations, $\overline{W}$ is computed as the mean over $\{W^b, b = 1, \cdots, 5000\}$, where

$$W^b = (\mu_{A(\hat{x}_S^{\text{opt}})} - \mu_C) \; I \times \left[A(\hat{x}_S^{\text{opt}}) \text{ is declared superior to } C \text{ in simulated trial b}\right].$$

Table 5 shows that, in general, the phase I-II/III design maintained type I error probability $\leq 0.05$ under $H_0$ and had a uniformly higher $\gamma_1$ and GP, $\gamma_2$, compared to the conventional phase I-II → phase III approach without dose re-optimization. The values of $\gamma_1$, $\gamma_2$, and $\overline{W}$ are uniformly larger for the phase I-II/III design than for the conventional phase I-II → phase III paradigm. The differences are extremely large in scenario 3, with an improvement of 0.73 in $\gamma_1$, and a 9.68 month improvement in $\overline{W}$. The smallest advantage of the phase I-II/III design is seen in scenario 2, with an improvement of 0.09 for $\gamma_1$ and 0.90 for $\overline{W}$. In scenarios 5 and 6, where two doses of $A$ give mean survival time larger than $\mu_C + \Delta = 36$ months, the phase I-II/III design provides respective improvements in $\gamma_2$ of 0.61 and 0.23, and improvements in $\gamma_1$ of 0.57 and 0.25. These scenarios illustrate the potential advantage of the phase I-II/III

**TABLE 5** Simulation results

| | | Alternative Hypothesis | | | | | | Null Hypothesis | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Scenario** | **Design** | $\overline{W}$ | $\gamma_1$ | $\gamma_2$ | $\overline{Dur}$ | $\overline{N}$ | | $\alpha$ | $\overline{Dur}$ | $\overline{N}$ |
| 1 | Phase I-II → phase III | 4.04 | 0.29 | 0.29 | 4.05 | 431.1 | – | 0.02 | 4.05 | 461.5 |
| | Phase I-II/III | 10.15 | 0.83 | 0.83 | 4.73 | 479.2 | – | 0.03 | 4.32 | 492.0 |
| 2 | Phase I-II → phase III | 7.87 | 0.66 | 0.66 | 4.29 | 459.2 | – | 0.06 | 4.18 | 489.4 |
| | Phase I-II/III | 8.97 | 0.75 | 0.75 | 4.45 | 470.7 | – | 0.02 | 4.18 | 489.9 |
| 3 | Phase I-II → phase III | 1.83 | 0.06 | 0.06 | 3.10 | 355.0 | – | < 0.01 | 3.28 | 385.7 |
| | Phase I-II/III | 11.51 | 0.79 | 0.79 | 4.56 | 476.9 | – | 0.04 | 4.22 | 485.6 |
| 4 | Phase I-II → phase III | 3.52 | 0.25 | 0.25 | 4.28 | 475.8 | – | 0.05 | 3.48 | 407.8 |
| | Phase I-II/III | 5.86 | 0.42 | 0.42 | 4.30 | 472.0 | – | 0.05 | 3.81 | 442.0 |
| 5 | Phase I-II → phase III | 5.61 | 0.21 | 0.25 | 3.98 | 428.5 | – | 0.01 | 3.83 | 440.4 |
| | Phase I-II/III | 16.71 | 0.68 | 0.88 | 4.24 | 464.4 | – | 0.03 | 4.37 | 493.9 |
| 6 | Phase I-II → phase III | 9.46 | 0.34 | 0.52 | 4.16 | 447.9 | – | 0.02 | 4.16 | 466.5 |
| | Phase I-II/III | 12.67 | 0.59 | 0.75 | 4.53 | 472.7 | – | 0.04 | 4.39 | 494.0 |

Note: $\alpha$ is the probability of a type I error or concluding an inferior version of $A$ is better than $C$ under the null. $\gamma_1$ is the generalized power at $A(x_S^{opt})$ (probability of selecting the best dose $x_S^{opt}$ and declaring it to be superior to $C$) under the alternative hypothesis. $\gamma_2$ is the generalized power (probability of selecting any truly superior dose of $A$ and declaring it superior to $C$). $\overline{W}$ is the mean improvement in patient survival under the alternative hypothesis, $\overline{Dur}$ is the mean trial duration, and $\overline{N}$ is the mean sample size.

design compared to the conventional phase I-II → phase III approach.

The phase I-II/III design does have the drawback that it requires treating more patients and longer trial durations, on average, than the conventional paradigm. Part of this required increase is due to the design correctly switching to the best dose of $A$ in terms of overall survival, which decreases the likelihood that a trial will stop early by declaring $C$ to be superior or due to futility. This increase in required sample size and trial duration are the price paid for the much larger probability of a successful phase III trial in cases where dose switching increases mean survival time with $A$.

Since phase I-II trials may have sample sizes ranging from 24 to 90 in practice, we chose the Eff-Tox sample size $N_{ET} = 60$ in the simulations as a practical compromise that obtains a reasonable amount of information in stage 1. Web Tables 1 and 2, seen in Web Appendix E, summarize additional simulations with $N_{ET} = 90$. Values of $\gamma_1$, $\gamma_2$ and $\overline{W}$ for the phase I-II/III design all increased substantially with $N_{ET}$ for all six scenarios. This is because more information at different doses in phase I-II makes switching to the best dose in stage 2 more likely.

To assess robustness of the phase I-II/III design to different event time distributions, we evaluated its performance for two lognormal distributions, with variances 0.25 and 1, a Weibull distribution with increasing or decreasing hazard, with shape parameters 4 or 0.5, and a gamma distribution with scale parameter 2. The true coefficients of $Y_E$ and $Y_T$ in the hazard function's linear term were kept constant for each distribution, and the remaining constant parameters $\beta_0$, $\beta_C$, $\beta_1$, $\beta_2$ were adjusted to obtain similar true means as in the exponential distribution simulation study. We exponentiated the linear term for the gamma and Weibull distribution rate parameters, but did not do this for the lognormal distribution. The means under the null and alternative hypotheses for each distribution are given in Web Table 3. Table 6 summarizes the robustness study, showing that under the alternative hypothesis, for each distribution, the phase I-II/III design has uniformly higher values of $\gamma_1$, $\gamma_2$, and $\overline{W}$, with substantially higher values for scenarios 1, 3, and 5. For the Weibull distribution with decreasing hazards in each scenario, the decrease in $\gamma_1$ and $\gamma_2$ for both designs is due to the assumptions of the logrank test being grossly violated by a high early failure rate. Because so many patients have early failures, patients are not followed as long before the final group sequential test. For this distribution, however, the phase I-II/III design still improved the probability of selecting the optimal dose of $A$ compared to the conventional paradigm by 0.49, 0.21, 0.15, 0.59, 0.46, and 0.14 in the six scenarios, respectively. Similar improvements are seen under the other distributions. This shows that the logrank test is not robust to the Weibull distribution with decreasing hazard. An extension of the phase I-II/III design might incorporate a robust group sequential test in place of the logrank test, to reduce the loss in power under a Weibull with decreasing hazard. The type I error constraints are nearly met for each distribution. Some slight inflation in $\alpha$ above 0.05 may be attributed to the proportional hazards assumption being violated. For each distribution, the phase I-II/III design treats more patients, on average, under both $H_0$ and $H_1$, and has slightly longer trial duration, but makes the correct decision much more often.

## 6 | DISCUSSION

We have proposed a new drug development strategy, which we call a phase I-II/III design, that re-optimizes the dose

**TABLE 6** Robustness simulation results

| | | Phase I-II → phase III | | | | Phase I-II/III design | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Scenario** | **Distribution** | $\overline{W}$ | $\gamma_1$ | $\gamma_2$ | $\alpha$ | $\overline{W}$ | $\gamma_1$ | $\gamma_2$ | $\alpha$ |
| 1 | Lognormal, $\sigma = 0.5$ | 7.72 | 0.30 | 0.30 | 0.02 | 14.56 | 0.80 | 0.80 | 0.03 |
| | Lognormal, $\sigma = 1$ | 6.18 | 0.30 | 0.30 | 0.02 | 14.88 | 0.85 | 0.85 | 0.03 |
| | Weibull increasing | 6.37 | 0.30 | 0.30 | 0.02 | 12.00 | 0.77 | 0.77 | 0.03 |
| | Weibull decreasing | 1.96 | 0.13 | 0.13 | 0.02 | 4.14 | 0.32 | 0.32 | 0.02 |
| | Gamma | 5.02 | 0.30 | 0.30 | 0.02 | 11.72 | 0.89 | 0.89 | 0.04 |
| 2 | Lognormal, $\sigma = 0.5$ | 12.18 | 0.73 | 0.73 | 0.05 | 15.14 | 0.91 | 0.91 | 0.05 |
| | Lognormal, $\sigma = 1$ | 11.70 | 0.70 | 0.70 | 0.06 | 13.78 | 0.82 | 0.82 | 0.03 |
| | Weibull increasing | 9.08 | 0.73 | 0.73 | 0.06 | 11.20 | 0.90 | 0.90 | 0.06 |
| | Weibull decreasing | 3.98 | 0.32 | 0.32 | 0.05 | 4.75 | 0.38 | 0.38 | 0.03 |
| | Gamma | 9.11 | 0.73 | 0.73 | 0.06 | 11.30 | 0.90 | 0.90 | 0.03 |
| 3 | Lognormal, $\sigma = 0.5$ | 2.17 | 0.06 | 0.06 | $< 0.01$ | 13.91 | 0.85 | 0.85 | 0.04 |
| | Lognormal, $\sigma = 1$ | 2.02 | 0.06 | 0.06 | $< 0.01$ | 12.91 | 0.81 | 0.81 | 0.03 |
| | Weibull increasing | 2.14 | 0.06 | 0.06 | $< 0.01$ | 13.63 | 0.84 | 0.84 | 0.03 |
| | Weibull decreasing | 0.90 | 0.03 | 0.03 | $< 0.01$ | 4.94 | 0.32 | 0.32 | 0.03 |
| | Gamma | 2.07 | 0.06 | 0.06 | $< 0.01$ | 13.23 | 0.86 | 0.86 | 0.03 |
| 4 | Lognormal, $\sigma = 0.5$ | 4.50 | 0.32 | 0.32 | 0.05 | 7.23 | 0.52 | 0.52 | 0.05 |
| | Lognormal, $\sigma = 1$ | 3.86 | 0.28 | 0.28 | 0.05 | 5.61 | 0.40 | 0.40 | 0.06 |
| | Weibull increasing | 4.67 | 0.33 | 0.33 | 0.04 | 6.61 | 0.47 | 0.47 | 0.04 |
| | Weibull decreasing | 4.14 | 0.28 | 0.28 | 0.06 | 6.21 | 0.42 | 0.42 | 0.07 |
| | Gamma | 4.62 | 0.31 | 0.31 | 0.04 | 6.93 | 0.47 | 0.47 | 0.04 |
| 5 | Lognormal, $\sigma = 0.5$ | 6.63 | 0.21 | 0.25 | 0.01 | 18.19 | 0.76 | 0.96 | 0.03 |
| | Lognormal, $\sigma = 1$ | 5.74 | 0.21 | 0.25 | 0.01 | 16.78 | 0.69 | 0.87 | 0.03 |
| | Weibull increasing | 7.41 | 0.21 | 0.25 | 0.01 | 15.88 | 0.64 | 0.77 | 0.03 |
| | Weibull decreasing | 4.00 | 0.16 | 0.18 | 0.01 | 12.27 | 0.48 | 0.63 | 0.03 |
| | Gamma | 6.73 | 0.21 | 0.25 | 0.01 | 18.73 | 0.80 | 0.92 | 0.03 |
| 6 | Lognormal, $\sigma = 0.5$ | 11.92 | 0.34 | 0.52 | 0.04 | 15.56 | 0.58 | 0.83 | 0.06 |
| | Lognormal, $\sigma = 1$ | 11.50 | 0.34 | 0.52 | 0.02 | 15.66 | 0.60 | 0.80 | 0.04 |
| | Weibull increasing | 10.29 | 0.34 | 0.52 | 0.04 | 12.76 | 0.54 | 0.77 | 0.05 |
| | Weibull decreasing | 5.46 | 0.21 | 0.31 | 0.02 | 6.63 | 0.32 | 0.38 | 0.03 |
| | Gamma | 10.90 | 0.34 | 0.52 | 0.02 | 14.80 | 0.59 | 0.83 | 0.04 |

Note: $\alpha$ = probability of a type I error or concluding that an inferior version of $A$ is better than $C$ under the null. $\gamma_1$ = generalized power at $x_S^{\text{opt}}$ under the alternative hypothesis (probability of selecting the best dose $x_S^{\text{opt}}$ and declaring $A(x_S^{\text{opt}})$ superior to $C$). $\gamma_2$ = generalized power (probability of selecting any superior dose and declaring $A(\hat{x}_S^{\text{opt}})$ superior to $C$). $\overline{W}$ is the mean improvement in survival time under the alternative hypothesis. $\overline{Dur}$ and $\overline{N}$ are the mean trial duration and sample size, respectively.

of an experimental agent $A$ chosen in phase I-II during phase III based on mean survival time. We use information from all patients treated with $A$, including their short-term efficacy and toxicity indicators, dose assigned, and survival time information, in order to more accurately select the dose of $A$ that provides the highest posterior mean survival time. The design is based on an assumed a mixture model for the survival time distribution that averages over the possible short-term phase I-II outcomes. While we have used the Eff-Tox trade-off based phase I-II design for stage 1 of the phase I-II/III design, one could replace the Eff-Tox design with any phase I-II design based on $(Y_E, Y_T)$ that uses some dose optimality criterion $\phi$ and includes AR. However, the necessary modifications of the design parameters and computer software to accommodate such a change would be non-trivial. Similarly, a complicated but straightforward extension of the methodology may address the problem of possible deaths before evaluation of $(Y_E, Y_T)$.

The simulations shows that, under a range of alternative cases, the generalized power $\gamma_2$, and probability $\gamma_1$ of the best possible decision, both are greatly increased by the phase I-II/III design compared to the phase I-II → phase III paradigm. The phase I-II/III design also has a much lower probability of making the least desirable decision, where a suboptimal dose is chosen and a true treatment advance is missed. A drawback of the phase I-II/III design is that it requires more patients and a slightly longer trial duration, on average, compared to the phase I-II → phase III paradigm. This seems like a very reasonable price to pay for the much larger values of $\gamma_1$, $\gamma_2$, and $\overline{W}$, in cases where re-optimizing the dose of the experimental agent increases its associated mean survival time.

## ACKNOWLEDGMENTS

Peter Thall's research was supported by NCI grants R01 CA 83932 and P30 CA 016672. Andrew Chapple's research was partially supported by the NIH grant 5T32-CA096520-07.

## ORCID

*Andrew G. Chapple* http://orcid.org/0000-0001-5332-2730

## REFERENCES

Arrowsmith, J. (2011). Trial watch: Phase III and submission failures: 2007–2010. *Nat Rev Drug Discovery* 10, 87–87.

Azriel, D., Mandel, M., and Rinott, Y. (2011). The treatment versus experimentation dilemma in dose-finding studies. *J Stat Plan Inference* 141, 2759–2768.

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

BIO, Biomedtracker, Amplion. Clinical Development Success Rates 2006-2015. https://www.bio.org/bio-industry-analysis-published-reports. (Accessed December 20, 2017).

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 51, 1372–1382.

Cancer.org. (2018). What are the phases of clinical trials? https://www.cancer.org/ treatment/ treatments-and-side-effects/ clinical-trials/ what-you-need-to-know/phases-of-clinical-trials.html (Accessed 17 Jan. 2018).

Chapple, A.G., Vannucci, M., Thall, P.F., and Lin, S.H. (2017). Bayesian variable selection for a semi-competing risks model with multiple components. *J Comput Stat Data Anal* 112, 170–185.

Chu, Y., Pan, H., and Yuan, Y. (2016). Adaptive dose modification for phase I clinical trials. *Stat Med* 35, 3497–3508.

East 6 (2016). *Statistical Software for the Design, Simulation and Monitoring of Clinical Trials*. Cambridge, MA: Cytel Inc.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 4, 711–732.

Hoering, A., LeBlanc, M., and Crowley, J. (2011). Seamless phase I-II trial design for assessing toxicity and efficacy for targeted agents. *Clin Cancer Res* 17,4, 640–646.

Huang, X., Biswas, X., Oki, Y., Issa, J.P., and Berry, D. (2006). A parallel phase I/II clinical trial design for combination therapies. *Biometrics* 63,2, 429–436.

Inoue L.Y.T., Thall P.F., and Berry D.A. (2002). Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 58, 823–831.

Jin, I-H, Liu, S., Thall, P.F., and Yuan, Y. (2014). Using data augmentation to facilitate conduct of phase I/II clinical trials with delayed outcomes. *J Am Stat Assoc* 109, 525-536.

Lee, J. and Liu, D. (2008). A predictive probability design for phase II cancer clinical trials. *Clin Trials* 5, 93–106.

Lee, K., Haneuse, S., Schrag, D., and Dominici, F. (2015). Bayesian semiparametric analysis of semicompeting risks data: Investigating hospital readmission after a pancreatic cancer diagnosis. *J R Stat Soc Ser C (Appl Stat)* 64, 253–273.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 549–556.

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

Schaid, D.J., Wieand, S., and Therneau, T. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* 3, 1, 507–513.

Seruga, B., Ocana, A., Amir, E., and Tannock, I.F. (2015). Failures in phase III: Causes and consequences. *Clin Cancer Res* 21, 4551–4560.

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

Simon, R., Wittes R.E., and Ellenberg, S.S. (1985). Randomized phase II clinical trials. *Cancer Treat Rep* 69, 1375–81.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med* 22, 286–703.

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

Sutton, R. S. and Barto, A. G. (1998), *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thall, P.F. (2008). A review of phase 2-3 clinical trial designs. *Lifetime Data Anal* 14, 37–53.

Thall, P.F., Herrick, R.C., Nguyen, H.Q., Venier, J.J., and Norris, J.C. (2014). Effective sample size for computing prior hyperparameters in Bayesian phase I-II dose-finding. *Clin Trials* 11, 657–666.

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

Thall, P.F. and Nguyen H.Q. (2012). Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *J Biopharm Stat* 22, 785–801.

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

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

Yin, G., Chen, N., and Lee, J. (2012). Phase II trial design with Bayesian adaptive randomization and predictive probability. *J R Stat Soc Ser C (Appl Stat).* 61, 219-235.

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

Yuan, Y., Nguyen, H., and Thall, P. F. (2016). *Bayesian Designs for Phase I-II Clinical Trials*. Boca Raton, Florida: Chapman and Hall/CRC Biostatistics Series.

## SUPPORTING INFORMATION

Web Appendices and tables, referenced in Sections 2, 3, and 5, are available at the Biometrics website on Wiley Online Library.

**How to cite this article:** Chapple AG, Thall PF. A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III. *Biometrics.* 2019;75:371–381. https://doi.org/10.1111/biom.12994

DOI: 10.1111/biom.12993

WILEY Biometrics

# Discussion of "A Hybrid Phase I-II/III Clinical Trial Design Allowing Dose Re-Optimization in Phase III" by Andrew G. Chapple and Peter F. Thall

**Eric S. Leifer** | **Nancy L. Geller**

Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland

**Correspondence**
Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, MD
*Email: Eric.Leifer@nih.gov
**Email: Nancy.Geller@nih.gov

With the rapid pace of medical innovation, there is a growing interest in streamlining the drug development process. In particular, there is a growing literature of trial designs which combine the traditional phases II and III into a single trial. Such designs have the potential to rapidly abandon ineffective treatments and more rapidly bring effective treatments to the relevant patients. Korn, Freidlin, et al. (2012) discuss several different designs which combine phase II and III into a single trial. One example they provide is CALGB-30610 (Cancer and Leukemia Group B 30610) which is a combined phase II/III trial which initially randomized small-cell lung cancer patients to one of three treatment arms. The control arm received standard radiotherapy while the two experimental arms received increased radiation doses with two different radiotherapy regimens. The phase II portion assessed toxicity scores and dropped the experimental arm with the higher average toxicity score. The phase III portion continues to randomize patients between the control arm and the remaining experimental arm with an overall survival (OS) primary endpoint. There are a myriad of possibilities for combining different trial phases including choice of short vs. long term endpoint, endpoint comparisons to make, and type I and II error probabilities or Bayesian quantities to use for the various comparisons.

We congratulate Chapple and Thall on their innnovative addition to this literature which shows theoretical promise. They are interested in the situation in which several doses of a new treatment are under consideration as comparators to a control treatment. Their goal is to choose the optimal dose, with respect to mean survival, among the possible treatment doses and to compare that dose to a control treatment in a randomized fashion. Their method, which they call the *phase I-II/III* design, has $K$ stages. Their paper switches between the *phase* and *stage* nomenclatures. We clarify that the phase I-II part of the design corresponds to stage 1 while the phase III part corresponds to stages 2 through $K$. For the remainder of this discussion, we primarily use the *stage* nomenclature.

Stage 1 uses a Bayesian outcome adaptive dose selection algorithm to assign successive small cohorts of patients, typically three patients each, to the various doses. For stage 1, the outcomes are short term efficacy and toxicity. Thall, Cook, and Estey (2006) provide an example of such outcomes in an acute myelogenous leukemia trial in which several doses of a biologic agent are tested. Efficacy is defined as complete disease remission by day 35 of treatment while toxicity is defined as death or development of life-threatening (grade 4) symptomatic toxicity by day 35. By the end of stage 1, an optimal dose $\hat{x}_{ET}^{opt}$ is identified with respect to a "trade-off" function of the Bayesian posterior probabilities of efficacy and toxicity. To identify $\hat{x}_{ET}^{opt}$, Chapple and Thall use the *Eff-Tox* method proposed by Thall and Cook (2004) and refined by Thall et al. (2014) in conjunction with their freely available Eff-Tox software. However, Chapple and Thall's general method could use other dose selection algorithms.

In stage 2, patients are randomized to receive either $\hat{x}_{ET}^{opt}$ or the control treatment $C$, until a prespecified $n_2^*$ deaths have occurred. At the conclusion of stage 2, the optimal dose $\hat{x}_{S}^{opt}$ is determined as the dose with the largest posterior mean survival time. The posterior is computed with respect to an exponential survival model with log hazard that is quadratic in dose and has linear indicator terms for short term efficacy and toxicity. It is important to note that $\hat{x}_{S}^{opt}$ may not be the

  

same dose as $\hat{x}_{ET}^{opt}$. Indeed, while stage 2 randomizes patients between $\hat{x}_{ET}^{opt}$ and $C$, the survival of all stage 1 patients continues to be followed in stage 2. Thus, the stage 2 randomization is not for the purpose of comparing $\hat{x}_{ET}^{opt}$ to $C$ at the end of stage 2, but rather to get an "early start" on the ultimate survival comparison between $\hat{x}_{S}^{opt}$ and $C$, assuming that $\hat{x}_{S}^{opt}$ turns out to be $\hat{x}_{ET}^{opt}$. If $\hat{x}_{S}^{opt} \neq \hat{x}_{ET}^{opt}$, then the phase III randomized survival comparison between $\hat{x}_{S}^{opt}$ and $C$ begins in stage 3 and continues through stage $K$, unless the trial is stopped early for efficacy, harm, or futility.

Chapple and Thall do not apply their method to a real data example, so we must rely on their simulation studies to evaluate their method. Chapple and Thall compare their phase I-II/III design to a design which they call the *phase I-II → phase III* design. The two designs are exactly the same except that the phase I-II/III design allows the optimal dose to be changed at the end of stage 2 while the phase I-II → phase III design does not allow such a change. That is, the phase I-II/III design allows for the possibility that $\hat{x}_{S}^{opt} \neq \hat{x}_{ET}^{opt}$ while the phase I-II → phase III design requires $\hat{x}_{S}^{opt} = \hat{x}_{ET}^{opt}$. It is appealing to obtain survival data on $\hat{x}_{ET}^{opt}$ in stage 2 before deciding whether to carry $\hat{x}_{ET}^{opt}$ forward in stages 3-$K$ in a head-to-head survival comparison with the control $C$.

Evidently, the situation in which the phase I-II/III design would be preferable to the phase I-II → phase III design when a particular dose is distinctly optimal with respect to efficacy and toxicity and a different dose is distinctly optimal with respect to survival. This situation is simulated in the alternative hypothesis case for scenario 3 of Table 5; note that average control group survival is 24 months for all of the Table 5 scenarios. In scenario 3, dose 5 is truly optimal with respect to the efficacy-toxicity tradeoff according to the efficacy-toxicity probabilities which are given in Table 3, scenario 2. However, as seen in Table 4, dose 5 has much lower mean survival as compared to control (6.3 months vs. 24 months) while doses 3 and 4 have substantially better mean survival as compared to control (33.1-38 months vs. 24 months). This is a rather extreme disconnect between short term efficacy-toxicity and survival.

Scenario 2 in Table 5 is more realistic. Here dose 5 is truly optimal with respect to efficacy-toxicity as well as survival. Nevertheless, in the alternative hypothesis case, the phase I-II/III design still has noticeably higher power than the phase I-II → phase III design (0.75 vs. 0.66). There are two reasons for the power advantage. First, as seen in Table 3, there is only a 49% chance that dose 5 is selected as optimal with respect to efficacy-toxicity. Second, as seen in Table 4, doses 1-4 have substantially worse average survival as compared to the 24 month control group survival. Thus, there is a 51% chance that the phase I-II → phase III design will compare one of the poorer survival doses 1-4 against control in stages 3-$K$. On the other hand, the phase I-II/III design can use dose 5's promising survival outcomes in stage 2 to increase the probability above 49% that dose 5 will be tested against control in stages 3-$K$. Parenthetically, Table 4 could be improved by including the $\hat{x}_{S}^{opt}$ selection percentages.

In addition to the power studies, Table 5 shows that the phase I-II/III design controls the type I error. At first glance, this is a bit of a surprise since $\hat{x}_{S}^{opt}$ is kept equal to $\hat{x}_{ET}^{opt}$ only if $\hat{x}_{ET}^{opt}$ has the highest posterior mean survival among the doses at the end of stage 2. Without any multiple comparison adjustment, this seems to give $\hat{x}_{S}^{opt} = \hat{x}_{ET}^{opt}$ a "good start" in stages 3-$K$ in its survival comparison against the control treatment. It seems as though that could inflate the type I error for the survival comparison. The type I error control is aided by the choice of the true mean survival rates under the null hypothesis as given in Table 4. Many of those rates are substantially lower than the control group's true mean survival of 24 months. For example, we see in Table 5, scenario 4 that the phase I-II/III design type I error is 0.05. However, only dose 1 has true mean survival equal to the control group's true 24 month mean survival; the other doses have mean survival 6.8 to 13.6 months. To more fully explore their phase I-II/III design's type 1 error control, Table 5 would include scenarios in which several (or all) of the doses had mean survival equal to the mean survival of the control treatment.

A potential shortcoming of Chapple and Thall's method is that the stage 1 patients may be different from the stage 2 patients. Often very ill patients are entered into stage 1 and possibly patients with improved prognosis are entered as the trial progresses. Such a situation could make it difficult to discard $\hat{x}_{ET}^{opt}$ at the end of stage 2 since $\hat{x}_{ET}^{opt}$ is the only dose studied in stage 2, in addition to the control. More research is needed to understand the potential effect of temporal changes in patient prognosis. Another question is how the trial should proceed if in stage 2, $\hat{x}_{ET}^{opt}$ has substantially worse survival than control. While Chapple and Thall's method would likely choose a different dose to proceed to stages 3-$K$, we might be concerned that all of the doses might be inferior to control. However, we could not be certain since $\hat{x}_{ET}^{opt}$ is the only dose which had comparable patients to control.

Chapple and Thall mention a few future avenues for research. One possibility which they did not mention was randomizing to multiple doses, in addition to control, in stage 2. Such a design would be the first step towards comparison of the Phase I-II/III to a multi-arm multi-stage (MAMS) design (Ghosh, Liu, et al. 2017). In particular, one could imagine in stage 1 eliminating doses that were unacceptably toxic. Stages 2-$K$ would then be a MAMS design which would compare the remaining doses to the control treatment. That comparison could be with respect to overall survival for all stages, or, an intermediary endpoint at earlier stages. Royston, Barthel, et al. (2012) discuss the STAMPEDE (Systemic

Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy) trial. The trial assesses three alternative classes of treatment in men starting androgen suppression. In the first three stages of the trial, failure-free survival is used for dropping ineffective treatments. Treatment failure is defined as radiologic, clinical, or PSA progression or death from prostate cancer. The primary endpoint is overall survival in the fourth and final stage. Thus, the STAMPEDE trial uses a time-to-event intermediary endpoint as opposed to a short-term success/failure short-term endpoint. Also, the STAMPEDE trial does not have a dose ordering in the manner of Chapple and Thall, so parametric modeling of mean survival as a function of dose is not done. However, with the growing interest in MAMS designs, a comparison to Chapple and Thall's design would be quite worthwhile.

Chapple and Thall are to be commended for introducing a design which has the potential for more rapidly developing new treatments for patients in need. We look forward to further study of the theoretical properties of their design and its application.

## ACKNOWLEDGEMENTS

The authors thank Neal Jeffries and James Troendle for their careful review of the manuscript. The authors are employees of the National Heart, Lung, and Blood Institute. The views expressed in this article are the authors' and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; National Institutes of Health; or the United States Department of Health and Human Services.

## REFERENCES

Ghosh, P., Liu, L., Senchaudhuri, P., Gao, P., Mehta, C. (2017). Design and monitoring of multi-arm multi-stage clinical trials. *Biometrics* 73, 1289–1299.

Korn, E. L., Freidlin, B., Abrams, J. S., Halabi, S. (2012). Design issues in randomized phase II/III trials. *J Clin Oncol* 30, 667–671.

Royston, P., Barthel, F. M. S., Parmar, M. K. B., Choodari-Oskooei, B., Isham, V. (2011). Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials* 12, 81.

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

Thall, P. F., Cook, J. D., Estey, E. H. (2006). Adaptive dose selection using efficacy-toxicity trade-offs: Illustrations and practical considerations. *J Biopharm Stat* 16, 623-638.

Thall, P. F., Herrick, R. C., Nguyen, H. Q., Venier, J. J., Norris, J. C. (2014). Effective sample size for computing prior hyperparameters in Bayesian phase I-II dose-finding. *Clin Trials* 11, 657–666.

DOI: 10.1111/biom.12992

WILEY Biometrics

# Discussion of "A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III" by Andrew G. Chapple and Peter F. Thall

**Tianjian Zhou**[1] | **Yuan Ji**[1,2,*]

[1]Research Institute, NorthShore University HealthSystem, Evanston, Illinois

[3]Department of Public Health Sciences, The University of Chicago, Chicago, Illinois

**Correspondence**
*Research Institute, NorthShore University HealthSystem, Evanston, IL; Department of Public Health Sciences, The University of Chicago, Chicago, IL
Email: yji@health.bsd.uchicago.edu

## 1 | REVIEW OF STANDARD ONCOLOGY DRUG DEVELOPMENT AND THE HYBRID DESIGN

We would like to congratulate Chapple and Thall (CT) for their nice work on a hybrid phase I-II/III clinical trial design for new drug development. CT propose a new strategy that allows a new investigational drug to be tested for safety, efficacy, and life-prolonging effects (patient survival) in a seamless fashion. Although not explicitly stated, the new strategy is suitable for cancer drugs. The proposed hybrid design combines a dose-finding scheme using the joint efficacy and toxicity outcome with a phase III confirmatory trial in which the optimal dose from the dose-finding part can be re-optimized based on survival outcome. As a discussion, we first briefly review the standard drug development regime and provide a few comments on CT's new hybrid design later.

### 1.1 | Standard Oncology Drug Development

In a sequential manner, a new cancer drug entering the clinical development stage usually goes through three phases of clinical trials before it can be approved by regulatory agencies as a commercial treatment. Phase I trials establish the safety profile of the drug by testing various doses of the drug. The primary endpoint is the dose-limiting toxicity (DLT) as a binary variable $Y_T$, using the notation in CT. The definition of DLT is based on clinical guidance such as the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) classification, and usually encompasses all grade 3 or higher toxicities with the exception of grade 3 nonfebrile neutropenia and alopecia. If the drug shows reliable safety profile and potential disease-fighting activity, one or few doses are selected for testing in phase II trials which aim at establishing efficacy $Y_E$. Here efficacy refers to a binary endpoint based on tumor shrinkage or pharmacodynamic biomarkers in a relatively short amount of time after treatment, and is measured using medical imaging or laboratory tests (eg, those in El-Maraghi and Eisenhauer 2008). Oftentimes, the drug is compared to a control arm in a randomized fashion. Importantly, in many cases, phase I trials can enroll patients with different types of cancers (ie, all-comers), but phase II trials narrow the investigation to one or few specific cancer types. This means change of enrollment criteria from phase I's all-comers to disease-specific patients in phase II, which could be a challenge for seamless designs. We will discuss this point later for CT's design. If a drug shows superior efficacy, it is further tested in a larger phase III trial in which the primary endpoint is the survival time $Y_S$. A drug can show great efficacy in terms of tumor shrinkage, but the ultimate benefit to cancer patients is evaluated in terms of patient survival.

### 1.2 | The Proposed Hybrid Approach

In the conventional three phases of sequential clinical development, the outcome of interest for each phase is different and is modeled separately, leading to possibly suboptimal dose selection and reduced power. CT therefore propose a hybrid phase I-II/III clinical trial design based on joint models for $[Y_E, Y_T \mid x]$ and $[Y_S \mid Y_E, Y_T, x]$, where $x$ is the dose level. We summarize the proposed design scheme in Figure 1.

  

According to Figure 1, the hybrid approach consists of two seamless phases of clinical development in which a new drug goes through dosage optimization in a phase I-II dose-finding study based on binary efficacy and toxicity response, and subsequently and seamlessly a dosage re-optimization in a phase III randomized trial based on survival outcome. Compared with the standard oncology drug development, the hybrid design can be highlighted with a few features. First, it condenses the two early phases, phases I and II trials into a single phase I-II dose-finding study. Second, it allows direct and seamless transition from dose finding to randomized comparison to a control arm in a confirmatory phase III study assuming an optimal dose can be estimated from the dose-finding stage. Third, the hybrid design proposes to re-optimize the dose in the phase III stage by modeling survival as a function of efficacy and toxicity outcomes observed from the phase I-II stage. Apparently, this is a new paradigm packing several adaptations into one single design that spans the entire clinical drug development spectrum. We provide some discussion next on this new paradigm.

## 2 | COMMENTS ON TRIAL DESIGN

### 2.1 | Skipping Phase II

In the hybrid design, once the optimal dose $A(\hat{x}_{ET}^{\text{opt}})$ is identified, it will enter a phase III trial for testing. This is a major

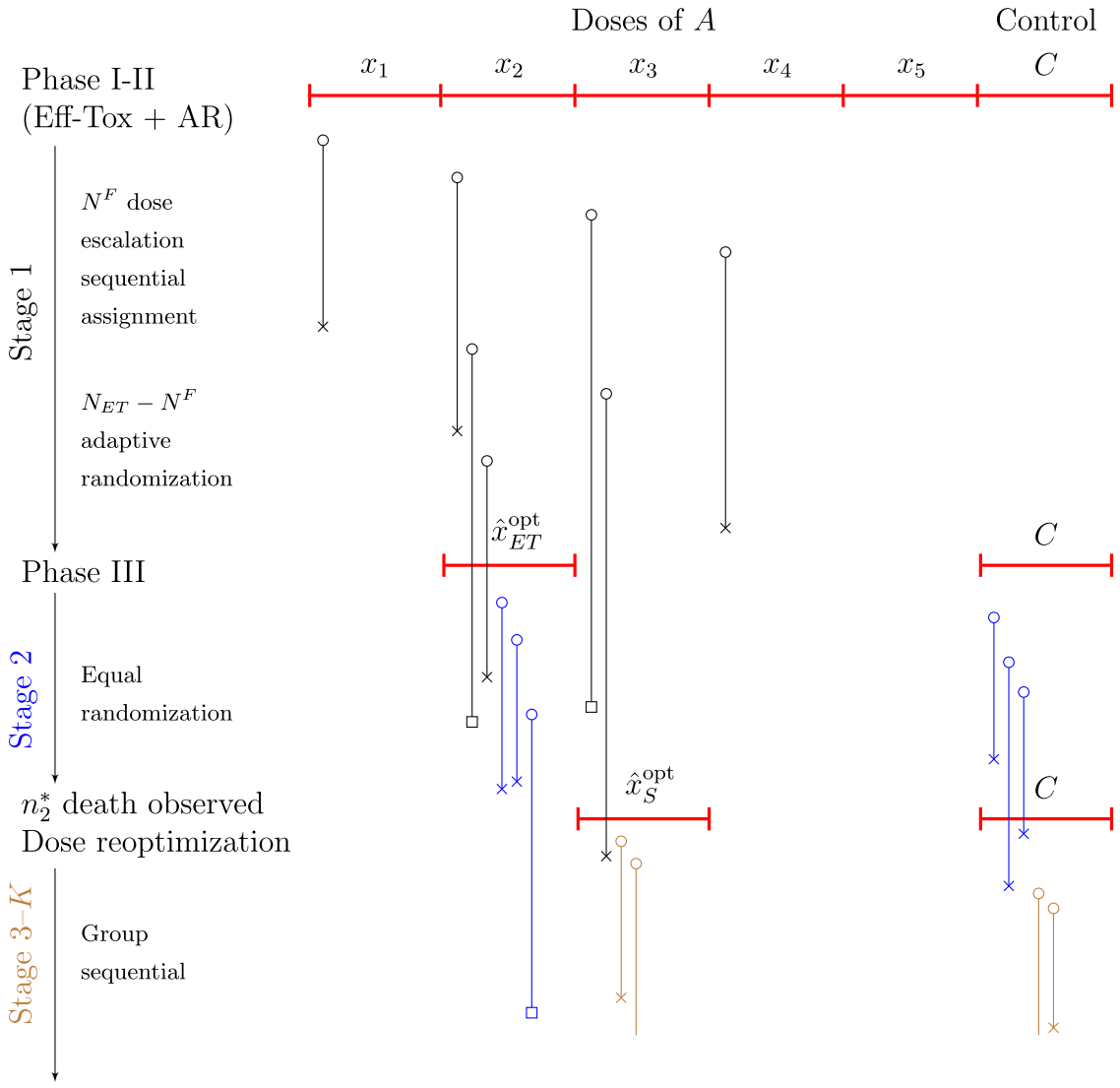


**FIGURE 1** Schema of the hybrid phase I-II/III design. The vertical axis represents time, and the horizontal axis represents dose levels of the experimental agent $x_i$ and the control $C$. A circle, cross or square indicates the accrual, death or censoring of a patient. In this example, the optimal dose based on efficacy and toxicity is $\hat{x}_{ET}^{\text{opt}} = x_2$ and is used in the first stage of phase III. After observing $n_2^*$ deaths, the optimal dose based on survival time is determined as $\hat{x}_S^{\text{opt}} = x_3$ and is used in the remaining stages of phase III. This figure appears in color in the electronic version of this article.

saving since the phase II testing in the standard drug development is skipped. Phase II is traditionally used to test tumor shrinkage ability or early efficacy of the drug on patients with specific indications, such as a cancer type based on tissues and organs. In the hybrid design, efficacy defined as a binary outcome is directly incorporated into the dose-finding investigation and a dose is optimized based on joint efficacy and toxicity outcomes. We note two practical considerations regarding the skipping of phase II. First, in order to carry out a dose-finding study based on joint efficacy and toxicity outcome from patients, both responses need to be observed quickly compared to enrollment speed in order to conduct dose finding in a timely manner. However, in reality, efficacy, if defined as tumor shrinkage, can take months to assess. This might pose challenge in the dose-finding stage since it means that one has to wait until the efficacy outcome of each patient (or each cohort of patients) is assessed before making a dose decision on the next patients. Second, the hybrid design is useful when the study population in phase I-II and in phase III is the same. That is, starting from the very beginning of the phase I-II dose finding, one has to decide the specific indication that will be targeted for potential drug approval. In contrast, phase I trials in standard oncology drug development usually enroll all cancer patients with say, solid tumors, regardless of tissue types, and only determine the specific diseases later in phase II. Since there is no phase II in the hybrid design, it assumes that the study population of the dose finding stage and the phase III stage is the same. We believe this is not a major problem in recent precision medicine and immune oncology, as many drugs like CAR-T therapies target specific diseases from the very beginning of the clinical development.

## 2.2 | Adaptive Randomization within Dose Finding

In dose finding, the hybrid design uses an adaptive randomization (AR) to assign patients across different doses after initial $N^F$ patients are sequentially assigned based on the Eff-Tox design (Thall and Cook, 2004). This is clever for various reasons. As the authors noted, using AR allows suboptimal doses to be assigned where the suboptimality is assessed based on binary efficacy and toxicity outcome, but not survival. By accumulate information on these doses, later on the hybrid design will have increased power to select the dose that might be suboptimal based on eff/tox outcomes but optimal based on survival time. As an added benefit, AR allows faster accrual of patients when compared to the cohort-based sequential enrollment in the Eff-Tox design. One caution is that some doses might never be assigned to patients when AR starts (eg, if $N^F$ is relatively small), and the AR probabilities on these doses are purely driven by data from other doses which have been assigned to patients. In addition, dose exclusion rules may be needed to force the AR probability to be zero if a dose is deemed too toxic or less efficacious than standard therapy.

## 2.3 | Dose Re-optimization in Phase III

The phase III trial is launched after phase I-II completes. The experimental arm of the phase III trial is an estimated optimal dose $\hat{x}_{ET}^{\text{opt}}$ based on the Eff-Tox design (Figure 1). A key feature of the hybrid design is that patients from phase I-II and phase III are continuously followed to accumulate their survival information. When $n_2^*$ survival events are observed, the optimal dose is re-optimized based on a joint model for survival time $Y_S$, efficacy $Y_E$ and toxicity $Y_T$ outcomes across all the doses $x$, using data from all the patients including those from phase I-II and those assigned to $\hat{x}_{ET}^{\text{opt}}$ from the phase III trial. The main argument for this strategy is 1) to speed up the drug development by launching a phase III after phase I-II completes and 2) to allow an opportunity to re-estimate the optimal dose that maximizes survival, the ultimate benefit for patients.

An alternative strategy would be to follow phase I-II patients for a period of time, collect their survival information, re-optimize the dose based on survival benefits using data from patients in phase I-II, make a go/no-go decision for launching a phase III trial using the new optimal dose, and if go, start phase III using the new optimal dose under a proper design (eg, group sequential with randomization to a control arm). This strategy is expected to take longer time than the hybrid design since phase III will not start until all phase I-II patients are followed for sufficient period of time and a go decision is warranted. However, it may reduce the possibility of treating patients at suboptimal doses (in terms of survival) since a phase III trial will be launched only when sufficient evidence points to potential benefits of survival in a go/no-go analysis. The overall tradeoff between the hybrid design and the alternative strategy depends on many factors, such as patients benefits, cost, and the actual phase II data. Perhaps this could be a future direction for new seamless designs that extend the hybrid design proposed in CT.

# 3 | COMMENTS ON SIMULATION STUDIES

The simulation studies in CT show the superiority of the proposed design compared to the phase I-II → III design, which does not re-optimize dose in phase III, but still follows the phase I-II dose finding using Eff-Tox and a phase III afterwards. We appreciate the authors' effort in providing an R package *Phase123*. We believe such a package will allow the public to assess the performance of the hybrid design in various scenarios, and we provide a few examples next.

The first is a null scenario, in which no doses of the experimental agent $A$ have desirable efficacy. For example, all doses of $A$ have efficacy probabilities similar to the control, as well as low toxicity probabilities. The quantity of interest is the study type I error rate, or false positive rate, defined as the probability of concluding any dose level of $A$ is better than $C$ after phase III completes. This is important for complex Bayesian adaptive designs as recently noted in Cunanan et al. (2017). The second is a scenario where the survival time improvement for $A(x_S^{\text{opt}})$ over $C$ is moderate, say $\Delta = 5$ months, as opposed to a 12 months improvement over $C$ (ie, a 50% improvement) in the paper. This allows one to see the benefit of the hybrid design when a new cancer drug improves the survival by a moderate amount, as is the case in many real-world trials (Prasad, 2017).

## 4 | CONCLUSION

The work by CT suggests a novel approach to the seamless combination of trial phases and provides important insight into the advancement of drug development. For general use of the proposed design, some practical considerations might need to be addressed, and several extensions can be made. For example, the Eff-Tox design used in the phase I-II part of the proposed design may be replaced by the SEARS design (Pan et al., 2014), which allows seamless transition from phase I to phase II. This assumes that efficacy and toxicity are observed at different time points and it is more suitable to separate toxicity-based dose finding from efficacy-centric phase II investigation. In general, we find the work by CT to be thought-provoking. We hope to see more work on hybrid trial designs and look forward to implementations of the proposed design in real-world trials in the future.

## REFERENCES

El-Maraghi R. H., Eisenhauer E. A. (2008). Review of phase II trial designs used in studies of molecular targeted agents: Outcomes and predictors of success in phase III. *J Clin Oncol* 26, 1346–1354.

Thall Peter F., Cook J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. Biometrics 60, 684–693.

Cunanan K. M, Iasonos A., Shen R., et al. (2017). Specifying the true-and false-positive rates in basket trials *JCO Precis Oncol* 1, 1–5.

Prasad V. (2017). Do cancer drugs improve survival or quality of life? *BMJ* 359, j4528.

Pan H., Xie F., Liu P., Xia J., Ji Y. (2014). A phase I/II seamless dose escalation/expansion with adaptive randomization scheme (SEARS). *Clin Trials* 11, 49–59.

DOI: 10.1111/biom.12991

REJOINDER

WILEY Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

# Rejoinder to "A hybrid phase I-II/III clinical trial design allowing dose reoptimization in phase III"

**Andrew G. Chapple**[1] | **Peter F. Thall**[2]

[1]Biostatistics Program, School of Public Health, Louisiana State University, New Orleans, Louisiana
[2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas

**Correspondence**
Andrew G. Chapple, Biostatistics Program, School of Public Health, Louisiana State University, New Orleans, LA, 70112.
Email: achapp@lsuhsc.edu

## 1 | SOME RELEVANT HISTORY AND STRONG OPINIONS

It is useful to begin our rejoinder by providing some historical context for our proposed three-phase "all-in-one" design. Three decades ago, two so-called "two-stage select-and-test" phase II-III designs were proposed, by Thall *et al.* (1988) (TSE) for binary outcomes, and by Schaid *et al.* (1990) (SWT) for event time outcomes. We mentioned these papers in the Introduction of our paper but did not explain them in any detail due to space limitations. The TSE design randomizes patients throughout, among experimental treatments $E_1, ..., E_k$ and a standard control treatment $C$ in stage 1, brings at most one best $E_{j^{\text{opt}}}$ forward to stage 2 if $E_{j^{\text{opt}}}$ is good enough compared to $C$ based on the stage 1 data, and if stage 2 is conducted then it does a final comparison of $E_{j^{\text{opt}}}$ to $C$ based on data pooled from both stages. The SWT design allows more than one $E_j$ to be selected and brought forward to stage 2, and it controls the power and type I error probability for each pairwise $E_j$ vs $C$ comparison. The TSE and STW design papers triggered a great deal of subsequent research on numerous extensions, for example, to group sequential designs with more than two stages, and designs using both early discrete and event time outcomes. It is encouraging that, in the intervening years, select and test designs have been applied to conduct real clinical trials, although the number of applications has been limited.

A key historical point, which is closely related to the new phase I-II/III design, is that the TSE design controls the generalized power (GP). The GP is computed under a "least favorable configuration (LFC)" of response probabilities, $\theta_C, \theta_1, ..., \theta_k$, in which exactly one $\theta_{j^{\text{opt}}}$ provides a meaningful improvement $\delta > 0$ over $C$, with $\theta_{j^{\text{opt}}} = \theta_C + \delta$. In the TSE design, the GP is defined as the probability, under the LFC, of correctly (1) selecting $E_{j^{\text{opt}}}$ as best in stage 1, (2) deciding that, compared to $C$, $E_{j^{\text{opt}}}$ is good enough to be brought forward to stage 2, and (3) at the end of stage 2, rejecting the global null hypothesis $\theta_C = \theta_1 = \cdots = \theta_k$ in favor of $\theta_{j^{\text{opt}}} > \theta_C$. This three-component correct decision event (CDE) is a much smaller event than simply rejecting a null hypothesis under a specifically targeted alternative, the probability of which is conventional power. A key property of the TSE design is that the GP accounts for the stage 1 selection process. In contrast, this is ignored by conventional two-arm group sequential designs, which assume implicitly that the experimental treatment somehow magically appears rather than arising from a preliminary screening process. The consequence is that the reported power figure for a two-arm phase III trial that was preceded by a screening process to determine $E$ is misleading. What matters is the GP, which is the probability of the CDE under the alternative, since the CDE accounts for the entire treatment evaluation and decision-making process.

In the phase I-II/III design, which hybridizes all three stages of the conventional phase I → phase II → phase III paradigm, defining the GP is a much harder problem because the data structure and CDE are much more complex. This motivated our preliminary explanation, summarized in Tables 1 and 2, of the many possible states of nature and decisions. Again, we strongly believe that it is the GP of the entire drug development and evaluation process that matters and that the power of the phase III portion of the process, considered by itself while ignoring what preceded phase III, may be very misleading.

  

## 2 | RESPONSES TO LEIFER AND GELLER

We thank Leifer and Geller (LG) for providing historical background, for their careful and detailed review of our proposed hybrid phase I-II/III clinical trial design, and for their thoughtful insights and suggestions. We agree that the Eff-Tox design used for phase I-II in our paper could be replaced by other phase I-II designs. We feel that this certainly should be done for qualitatively different early phase settings, for example where the early outcomes are ordinal variables, as in Thall and Nguyen (2012), or event times, as in Jin *et al.* (2014). In such cases, the mixture model for the survival distribution necessarily would be changed accordingly. As noted by LG, we do not include a real trial as an illustration. Our proposed methodology is new and thus has not yet been applied. We are hopeful that the practicality of the method, and free availability of the necessary software, will lead to applications in the near future.

LG note several limitations of our simulation study and suggest additional cases. We agree that many more cases could, and should, be considered. They also note that some scenarios studied in our simulations reflect "a rather extreme disconnect between short term efficacy-toxicity and survival." We feel that these are very important cases to include in any simulation study of this type of hybrid design since there are numerous examples in the medical literature where this disconnect actually occurs. Indeed, this was, in large part, motivation for our design. In our model, and in general, this is quantified by the strength of the regression of $Y_S$ on $(Y_E, Y_T)$, so it is a matter of degree. A straightforward but tedious sensitivity analysis in the numerical values of the parameters $\beta_E$ and $\beta_T$ appearing in equation (5) could be conducted to explore this issue. This also gets at the more general problem that, even under the conceptually straightforward parametric mixture model given by our equations ((2)-(5)), the number of possible cases that might be studied is immense. Put another way, we have gone far beyond considering only a null versus an alternative hypothesis. Since a computer simulation study is an experiment, this suggests that a more extensive, formally designed computer simulation study would be very useful, although this is a nontrivial problem for complex clinical trial designs. Such a study would be likely to provide additional insights into the properties of our proposed phase I-II/III design, as well as future variants and extensions.

LG have sugested that it would be useful for us to provide the probability that each dose was selected as $\hat{x}_S^{\text{opt}}$. We agree, and thank them for this suggestion. Table 1 provides the reoptimization selection percentage for each dose, along with the true mean survival times, for each scenario.

In table 5, the respective GP figures at $A(x_S^{\text{opt}})$ for the six scenarios, given as percentages, are 100 $\gamma_1 = (83, 75, 79, 42, 68, 59)$. The corresponding selection percentages of the truly optimal doses under the alternative are (90.74, 78.82, 80.20, 58.92, 68.9, 64.42). This shows that the probability of correctly switching to the optimal dose but not declaring $A(x_S^{\text{opt}})$ superior to $C$ under the alternative is small for each scenario. In scenarios 5 and 6, there are much higher probabilities of switching to the dose with the largest mean survival time, compared with the slightly suboptimal dose that has a $> 36$ month improvement.

We agree with LG's observation that an important limitation of our design arises if the characteristics of stage 1 and stage 2 patients are substantively different. In particular, stage 1 patients may have worse prognosis. Our design does not accommodate this setting. This suggests a very important area for future research. We also agree that another useful extension would be to randomize patients among several doses in stage 2, which is in the spirit of the SWT design. We currently are working to develop this generalization.

## 3 | RESPONSES TO ZHOU AND JI

We thank Zhou and Ji (ZJ) for their thoughtful comments, a careful review of our design, and kind remarks. Their initial review of *Standard Oncology Drug Development* is very informative, and it provides a useful context for evaluating the practical utility of our design. We also are very grateful for the figure provided by ZJ that shows how our design may play out over time. We find this to be an extremely informative graphical illustration of a rather complex process.

ZJ point out two important practical limitations. The first is that the phase I-II outcomes may be more complex than simple binary indicators and may take a nontrivial amount of time to evaluate. We strongly agree that this is a major practical issue. It suggests that, in this type of setting, a different approach is needed, such as the "late onset efficacy-toxicity" phase I-II design of Jin *et al.* (2014), which accommodates possibly right-censored time-to-toxicity and time-to-efficacy outcomes observed over pre-specified intervals, and thus is more practical. The second limitation, also noted by LG, is that our design does not account for patient heterogeneity. Again, this suggests several directions in which the design may be extended to accommodate a broader array of clinical settings.

**TABLE 1** Reoptimization selection percentages for each dose $x_j$ under each phase I-II/III scenario's null and alternative hypotheses, corresponding to Table 4 of the main text

| Scenario | Eff-Tox Scenario | Hyp | Percent $\hat{x}_S^{opt} = x_j$ ($x_1$, $x_2$, $x_3$, $x_4$, $x_5$) | True Mean Survival Times ($\mu_{A(x_1)}$, $\mu_{A(x_2)}$, $\mu_{A(x_3)}$, $\mu_{A(x_4)}$, $\mu_{A(x_5)}$) |
|---|---|---|---|---|
| 1 | 1 | Null | (0.7, 12.0, 59.7, 27.1, 0.4) | (8.3, 17.9, 24.0, 22.5, 9.8) |
| | | Alt | (0.0, 1.7, 90.7, 7.5, 0.0) | (1.0, 14.5, 36.2, 28.3, 1.0) |
| 2 | 2 | Null | (4.0, 15.1, 14.4, 33.9, 32.6) | (14.0, 17.8, 21.9, 23.0, 24.0) |
| | | Alt | (0.1, 1.2, 4.6, 15.2, 78.8) | (7.1, 10.3, 16.0, 19.5, 36.0) |
| 3 | 2 | Null | (2.5, 11.3, 61.8, 24.0, 0.4) | (9.5, 18.5, 24.0, 22.5, 10.4) |
| | | Alt | (0.9, 5.1, 80.2, 13.8, 0.0) | (6.9, 24.7, 38.0, 33.1, 6.3) |
| 4 | 3 | Null | (45.9, 16.4, 12.2, 21.5, 4.0) | (24.0, 13.6, 8.9, 6.8, 7.8) |
| | | Alt | (58.9, 16.2, 9.5, 14.1, 1.3) | (38.0, 24.6, 18.4, 15.0, 21.1) |
| 5 | 3 | Null | (0.3, 18.6, 54.0, 26.8, 0.0) | (9.3, 18.7, 24.0, 22.7, 10.4) |
| | | Alt | (0.0, 11.5, 68.9, 19.6, 0.0) | (7.8, 28.8, 44.0, 38.6, 7.4) |
| 6 | 1 | Null | (0.0, 1.6, 15.6, 63.1, 19.7) | (3.2, 10.1, 20.4, 24.0, 20.4) |
| | | Alt | (0.0, 1.6, 18.5, 61.2, 18.7) | (3.0, 12.9, 31.8, 40.0, 36.6) |

ZJ suggest a modification wherein a "go/no go" decision is made based on comparison of $A(\hat{x}_S^{opt})$ to $C$, with this done based on survival data from longer follow-up in phase I-II before proceeding to the randomized trial. This is a potentially useful variant of the design. It seems likely that, if $E_j^{opt}$ is truly no better than $C$, then this go/no go decision can be made more reliably by the phase I-II/III design at the end of its dose reoptimization stage. In any case, it certainly will be worthwhile to examine the comparative properties of design with this suggested alternative early futility stopping rule.

Finally, ZJ suggests that alternative early phase designs may be used in place of the Eff-Tox design. As stated above, we fully agree with doing this, as appropriate for a given clinical setting. In this regard, we view the phase I-II/III design that we have proposed as a modular paradigm, and we anticipate numerous extensions and elaborations. Our main goal is to convince clinicians to use this design, or a modified version tailored to their particular clinical setting, since the potential gain in reliability, savings in resources, and seamless acceleration of the clinical screening and evaluation process promise to be substantial.

## REFERENCES

Jin, I.-H., Liu, S., Thall, P.F., 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.

Schaid, D.J., Wieand, S., and Therneau, T. (1990). Optimal two-stage screening designs for survival comparisons. Biometrika, 77(3), 507–513.

Thall, P.F. and Nguyen, H.Q. (2012). Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. Journal of Biopharmaceutical Statistics, 22, 785–801.

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