Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# A hybrid geometric phase II/III clinical trial design based on treatment failure time and toxicity

Peter F. Thall [a,*], Hoang Q. Nguyen [a], Xuemei Wang [a], Johannes E. Wolff [b]

[a] Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, USA
[b] Department of Pediatrics, Tufts Medical Center, Boston, MA, USA

ARTICLE INFO

*Article history:*
Received 20 January 2011
Received in revised form
28 October 2011
Accepted 31 October 2011
Available online 10 November 2011

*Keywords:*
Bayesian design
Clinical trial
Phase II/III design
Treatment selection

ABSTRACT

The problem of comparing several experimental treatments to a standard arises frequently in medical research. Various multi-stage randomized phase II/III designs have been proposed that select one or more promising experimental treatments and compare them to the standard while controlling overall Type I and Type II error rates. This paper addresses phase II/III settings where the joint goals are to increase the average time to treatment failure and control the probability of toxicity while accounting for patient heterogeneity. We are motivated by the desire to construct a feasible design for a trial of four chemotherapy combinations for treating a family of rare pediatric brain tumors. We present a hybrid two-stage design based on two-dimensional treatment effect parameters. A targeted parameter set is constructed from elicited parameter pairs considered to be equally desirable. Bayesian regression models for failure time and the probability of toxicity as functions of treatment and prognostic covariates are used to define two-dimensional covariate-adjusted treatment effect parameter sets. Decisions at each stage of the trial are based on the ratio of posterior probabilities of the alternative and null covariate-adjusted parameter sets. Design parameters are chosen to minimize expected sample size subject to frequentist error constraints. The design is illustrated by application to the brain tumor trial.

## 1. Introduction

The problem of comparing several experimental treatments, $E_1,\ldots,E_J$, to a standard treatment, $S$, is common in clinical trials. Reviews are given by Simon et al. (1994), Rubinstein et al. (2004), Bretz et al. (2006), Jennison and Turnbull (2006), and Thall (2008). For a one-dimensional outcome, the focus is on $J+1$ parameters, $\vec{\boldsymbol{\theta}}=(\theta_0,\theta_1,\ldots,\theta_J)$, where $\theta_0$ corresponds to $S$ and $\theta_j$ corresponds to $E_j$ for $j=1,\ldots,J$. For example, each entry of $\vec{\boldsymbol{\theta}}$ may be a probability of tumor response, or median survival time.

In this setting, Thall, Simon and Ellenberg (TSE) (1988) proposed a 2-stage phase II/III design for binary outcomes combining ideas from selection (cf. Bechhofer et al., 1995) and group sequential testing (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Wang and Tsiatis, 1978; Jennison and Turnbull, 2006). The TSE design randomizes patients among $E_1,\ldots,E_J$ and $S$ in stage 1, and proceeds to stage 2 only if the empirically best experimental treatment in stage 1, $E_v$, is promising compared to $S$. In stage 2, patients are randomized between $E_v$ and $S$ and a final test is conducted

[*] Corresponding author.
*E-mail address:* rex@mdanderson.org (P.F. Thall).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.10.016

to decide whether $E_v$ provides an improvement over $S$. Design parameters are chosen to minimize expected sample size subject to constraints on overall Type I and Type II error. Schaid et al. (1990) proposed a similar 2-stage design for time-to-event outcomes that allows more than one $E_j$ to move forward to stage 2, allows termination of the trial in stage 1 for either futility or superiority, and does pairwise comparisons with the possibility of concluding that more than one $E_j$ provides an improvement over $S$. Many extensions of these designs have been proposed, including designs with more than two stages (Stallard and Todd, 2003; Stallard and Friede, 2008), a design that continues accrual between stages and uses the stage 1 data to determine the stage 2 sample size adaptively (Liu and Pledger, 2005), and an algorithm for computing decision boundaries (Cheung, 2008).

We address the problem of selecting among $E_1,\ldots,E_J$ and comparing the selected treatment to $S$ based on $(T,Y)$, where $T$ is survival or event-free survival (EFS) time and $Y$ indicates a nonfatal severe adverse event (SAE). Thus, death is accounted for by $T$ and not by $Y$. Our motivation was the desire to provide a feasible design for a trial of four chemotherapy combinations for treating pediatric patients with choroid plexus tumors (CPTs). CPTs include three histological subtypes: choroid plexus carcinoma (CPC), atypical plexus papilloma (APP) and choroid plexus papilloma (CPP). A typical CPT patient is a very young child, with 61% less than three years of age. Because these tumors are too rare for standard protocol designs, to date no clinical trial specific for CPTs has been completed (Wolff et al., 2002). In the planned trial, the standard treatment is $S$=carboplatin+cyclophosphamide+etoposide+vincristine, and the three experimental treatments are $E_1$=doxorubicin+cisplatinum+actinomycin+etoposide, $E_2$=high dose methotrexate and $E_3$=temozolomide+irinotecan. The outcomes are $T$=EFS time, with an event defined as disease progression or death, and $Y$=I(severe toxicity). The baseline covariates are Age and indicators of metastatic disease, whether surgery achieved a complete resection (CR), and the unfavorable CPC histology. The goal is to increase $T$ on average while controlling the probability of toxicity, compared to standard therapy, while also accounting for patient heterogeneity.

We propose a hybrid two-stage select-and-test design based on $(T,Y)$ and baseline covariates. The logic of our design mimics that of the TSE design, with design parameters chosen to minimize expected sample size subject to overall Type I and Type II error constraints. The main differences between our procedure and the TSE design are that (i) we account for patient heterogeneity, (ii) we base the tests on two-dimensional covariate-adjusted treatment effect parameters, and (iii) we take a Bayesian geometric approach in which test statistics are constructed from posterior probabilities of two-dimensional parameter sets, rather than using conventional $Z$-scores.

## 2. Overview of the methodology

### 2.1. The basic idea

To establish the ideas underlying the geometric construction, for simplicity we temporarily consider the special case of two treatments, with $j=0$ and 1 for $S$ and $E$, and we ignore covariates. The parameters $\theta_{Y,j}$ and $\theta_{T,j}$ represent average behavior of two different outcomes. In the CPT trial, the physician decided to use $\theta_{Y,j}$=the probability of toxicity and $\theta_{T,j}=\Pr(\mathrm{T}>24 \text{ months})$ with treatment $j$ as the parameters, where $T$ is the time to treatment failure, defined as disease progression or death. Alternatively, the physician might use the mean of median of $T$ as $\theta_{T,j}$. Thinking of $\boldsymbol{\theta}_j=(\theta_{Y,j},\theta_{T,j})$ as a two-dimensional parameter that describes the average behavior of the outcomes $(Y,T)$ with treatment $j$, the problem of comparing $E$ to $S$ becomes that of comparing $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_0$.

In such settings, there are many methods for dimension reduction to obtain a one-dimensional decision criterion. Our approach exploits a Bayesian formulation and subjective goals of the physician, using the following construction. Let $\boldsymbol{\mu}_0=(\mu_{Y,0},\mu_{T,0})$ be the prior mean of $\boldsymbol{\theta}_0$ based on experience with $S$. We elicit a set of fixed target pairs, $\{\boldsymbol{\theta}^{e,1},\ldots,\boldsymbol{\theta}^{e,m}\}$, for $\boldsymbol{\theta}_E$ that the physician considers equally desirable improvements over $\boldsymbol{\mu}_0$. We then form the polygonal line $\mathcal{L}$ connecting the elicited values, and define the target parameter set, $\Theta_{\mathcal{L}}$, of all $\boldsymbol{\theta}$ pairs at least as desirable as a pair on $\mathcal{L}$. The test is based on the posterior probability of $\Theta_{\mathcal{L}}$ compared to the posterior probability of the null set $\Theta_0$ where $\theta_Y \geq \mu_{Y,0}$ and $\theta_T \leq \mu_{T,0}$. Accounting for covariate effects requires the additional steps of fitting regression models, obtaining covariate-adjusted treatment effects, and transforming $\Theta_0$ and $\Theta_{\mathcal{L}}$. This adds technical complications to the construction, but the essential idea is still to base tests on posterior probabilities of two-dimensional parameter sets. For the CPT trial, the target line and target parameter set in terms of $(\theta_{Y,j},\theta_{T,j})$ are given in Fig. 1. In other settings with two outcomes, this construction may be carried out in the same way, but using different parameters. For example, in a phase I/II dose-finding trial based on indicators of toxicity and response, $\theta_{T,j}$ would be replaced by the response probability.

### 2.2. Model and parameters

We define the distribution of $T$ conditional on $Y$ to allow the possibility that toxicity may affect the event time. Let $\xi$ denote the vector of model parameters. For a patient with baseline covariates $\boldsymbol{Z}=(Z_1,\ldots,Z_q)$ given treatment $j=0,1,\ldots,J$, denote the probability of toxicity by $\theta_{Y,j}(\boldsymbol{Z},\xi)=\Pr(Y=1|j,\boldsymbol{Z},\xi)$ and the conditional probability density function and survivor function of $T$ given $\boldsymbol{Z}$ and $Y$ by $f_j(t|\boldsymbol{Z},Y,\xi)$ and $\overline{F}_j(t|\boldsymbol{Z},Y,\xi)=\Pr(T>t|j,\boldsymbol{Z},Y,\xi)$. Denoting the time to the event or independent right censoring by $T^o$ and $\epsilon=I(T^o=T)$, the likelihood function of the observed outcome $(T^o,\epsilon,Y)$ given $\boldsymbol{Z}$ and $j$ is

$$\mathcal{L}(T^o,\epsilon,Y|j,\boldsymbol{Z},\xi)=f_j(T^o|\boldsymbol{Z},Y,\xi)^{\epsilon}\overline{F}_j(T^o|\boldsymbol{Z},Y,\xi)^{1-\epsilon}\theta_{Y,j}(\boldsymbol{Z},\xi)^{Y}\{1-\theta_{Y,j}(\boldsymbol{Z},\xi)\}^{1-Y}. \qquad (1)$$

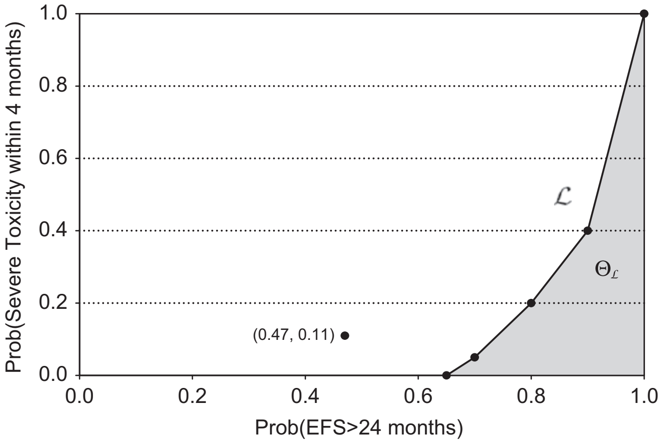

**Fig. 1.** Historical mean pair $\boldsymbol{\mu}_{\theta,0} = (0.47,0.11)$, elicited targets pairs $\{\boldsymbol{\theta}^{e,1},\ldots,\boldsymbol{\theta}^{e,5}\}$, target line $\mathcal{L}$, and target set $\boldsymbol{\Theta}_{\mathcal{L}}$ in the natural parameter domain, with $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{Z})$ evaluated at the reference covariate vector, corresponding to a 2-year-old patient with non-metastatic disease, complete resection and CPC histology.

Let $\theta_{T,j}(\boldsymbol{Z},Y,\xi)$ be a parametric function quantifying the effect of treatment $j$ on $T$ given $Y$ and $\boldsymbol{Z}$. This may be a mean, a median, or $\overline{F}_j(t^*|\boldsymbol{Z},Y,\xi)$ for fixed time $t^*$. Thus, $\boldsymbol{\theta}_j(\boldsymbol{Z},Y,\xi) = (\theta_{T,j}(\boldsymbol{Z},Y,\xi),\theta_{Y,j}(\boldsymbol{Z},\xi))$ is a two-dimensional parametric function. The treatment effect parameters that are basis of the test are defined as follows. Let $\boldsymbol{Z}_Y$ and $\boldsymbol{Z}_T$ be vectors of variables obtained from $\boldsymbol{Z}$, of dimensions $q_Y$ and $q_T$. We assume parametric regression models under which $\theta_{T,j}(\boldsymbol{Z},Y,\xi)$ is a function of the linear term

$$\eta_{T,j}(\boldsymbol{Z}_T,Y,\xi) = \gamma_{T,j} + \beta_{T,1}Z_{T,1} + \cdots + \beta_{T,q_T}Z_{T,q_T} + \beta_{T,q_T+1}Y = \gamma_{T,j} + \boldsymbol{\beta}_T\boldsymbol{Z}_T + \beta_{T,q_T+1}Y, \tag{2}$$

and $\theta_{Y,j}(\boldsymbol{Z},\xi)$ is a function of the linear term

$$\eta_{Y,j}(\boldsymbol{Z}_Y,\xi) = \gamma_{Y,j} + \beta_{Y,1}Z_{Y,1} + \cdots + \beta_{Y,q_Y}Z_{Y,q_Y} = \gamma_{Y,j} + \boldsymbol{\beta}_Y\boldsymbol{Z}_Y, \tag{3}$$

where $\boldsymbol{\beta}_Y$ and $\boldsymbol{\beta}_T$ are covariate effects. Expression (2) specifies the conditional distribution of $[T|j,Y,\boldsymbol{Z}]$, with $\beta_{T,q_T+1}$ the effect of toxicity on the distribution of $T$. Expression (3) specifies the marginal distribution of $[Y|j,\boldsymbol{Z}]$. The parameter $\gamma_{T,j}$ is the effect of treatment $j$ on $T$ after adjusting for $Y$ and $\boldsymbol{Z}$, and $\gamma_{Y,j}$ is the effect of treatment $j$ on $Y$ after adjusting for $\boldsymbol{Z}$. Thus, the parameter vector $\vec{\gamma} = (\gamma_0,\gamma_1,\ldots,\gamma_J)$ may be considered a $2(J+1)$-dimensional, covariate-adjusted generalization of the $(J+1)$-dimensional parameter vector $\vec{\boldsymbol{\theta}} = (\theta_0,\theta_1,\ldots,\theta_J)$ given earlier in the univariate setting. For each $E_j$, we define the two-dimensional $E_j$-versus-$S$ treatment effect $\boldsymbol{\delta}_j = \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_0 = (\gamma_{T,j}-\gamma_{T,0},\gamma_{Y,j}-\gamma_{Y,0}) = (\delta_{T,j},\delta_{Y,j})$, so that larger $\delta_{T,j}$ and smaller $\delta_{Y,j}$ are more desirable. The test statistics will be defined in terms of $2J$-dimensional parameter vector $\vec{\boldsymbol{\delta}} = (\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_J)$.

### 2.3. Steps of the construction

We first list the steps required to construct the test statistics. Additional details are given in Sections 3 and 4. Let $\boldsymbol{\Theta}$ denote the set of all possible values taken on by the $\boldsymbol{\theta}_j(\boldsymbol{Z},Y,\xi)$ pairs, and denote the set of all possible $\boldsymbol{\delta}_j$'s by $\boldsymbol{\Delta}$. Since $\theta_{T,j}(\boldsymbol{Z},Y,\xi)$ is defined conditional on the toxicity outcome $Y$, in the following construction targets are elicited for each value $y^* = 0,1$ of $Y$ and averaged to obtain a set of targets that do not depend on $Y$.

1. Ask the physician to specify a reference covariate vector $\boldsymbol{z}^*$. For each value $y^*$=0, 1 of $Y$, denote the prior mean of $\boldsymbol{\theta}_0(\boldsymbol{z}^*,y^*,\xi)$ by $\boldsymbol{\mu}_{\theta,0}(y^*) = (\mu_{\theta,T,0}(y^*),\mu_{\theta,Y,0})$. For a fixed toxicity probability $\theta_Y^*$, denote $\mu_{\theta,T,0} = \mu_{\theta,T,0}(1)\theta_Y^* + \mu_{\theta,T,0}(0)(1-\theta_Y^*)$ and $\boldsymbol{\mu}_{\theta,0} = (\mu_{\theta,T,0},\mu_{\theta,Y,0})$.

2. For each $y^*$=0 or 1, elicit a set of target $\boldsymbol{\theta}$ pairs $\{\boldsymbol{\theta}^{(e,1)}(y^*),\ldots,\boldsymbol{\theta}^{(e,m)}(y^*)\}$ in $\boldsymbol{\Theta}$ from the physician that correspond to equally desirable improvements over $\boldsymbol{\mu}_{\theta,0}(y^*)$. The two sets should be elicited so that $\boldsymbol{\theta}^{(e,r)}(1)$ corresponds to $\boldsymbol{\theta}^{(e,r)}(0)$ for each $r = 1,\ldots,m$. Obtain a single set of mean target pairs by forming the weighted averages $\boldsymbol{\theta}^{(e,r)} = \boldsymbol{\theta}^{(e,r)}(1)\theta_Y^* + \boldsymbol{\theta}^{(e,r)}(0)(1-\theta_Y^*)$, for $r = 1,\ldots,m$. Construct the polygonal *target line*, $\mathcal{L}$ that connects the mean elicited target pairs $\vec{\boldsymbol{\theta}}^e = (\boldsymbol{\theta}^{(e,1)},\ldots,\boldsymbol{\theta}^{(e,m)})$.

3. The *target set* $\boldsymbol{\Theta}_{\mathcal{L}} = \bigcup'_{\theta\in\mathcal{L}}\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_T \geq \theta'_T \text{ and } \theta_Y \leq \theta'_Y\}$ consists of all $\boldsymbol{\theta}$ pairs at least as desirable as a pair on $\mathcal{L}$. Fig. 1 illustrates $\boldsymbol{\mu}_{\theta,0}$,$\mathcal{L}$, and $\boldsymbol{\Theta}_{\mathcal{L}}$ for the CPT trial.

4. In $\boldsymbol{\Delta}$, denote the line obtained from $\mathcal{L}$ by $\mathcal{D}_{\mathcal{L}}$, the two-dimensional alternative treatment effect parameter set obtained from $\boldsymbol{\Theta}_{\mathcal{L}}$ by $\boldsymbol{\Delta}_{\mathcal{L}}$, and the null set of all $\boldsymbol{\delta} = (\delta_T,\delta_Y)$ in $\boldsymbol{\Delta}$ with $\delta_T \leq 0$ and $\delta_Y \geq 0$ by $\boldsymbol{\Delta}_0$. Fig. 2 illustrates $\mathcal{D}_{\mathcal{L}}$,$\boldsymbol{\Delta}_0$, and $\boldsymbol{\Delta}_{\mathcal{L}}$ for the CPT trial.

5. The test statistics used in each stage of the trial are the posterior probability ratios

$$R_j(data) = \frac{\Pr(\boldsymbol{\delta}_j \in \Delta_{\mathcal{L}}|data)}{\Pr(\boldsymbol{\delta}_j \in \Delta_0|data)}, \quad j = 1,\ldots J. \tag{4}$$

Under the construction, larger $R_j(data)$ corresponds to greater superiority of $E_j$ over $S$.

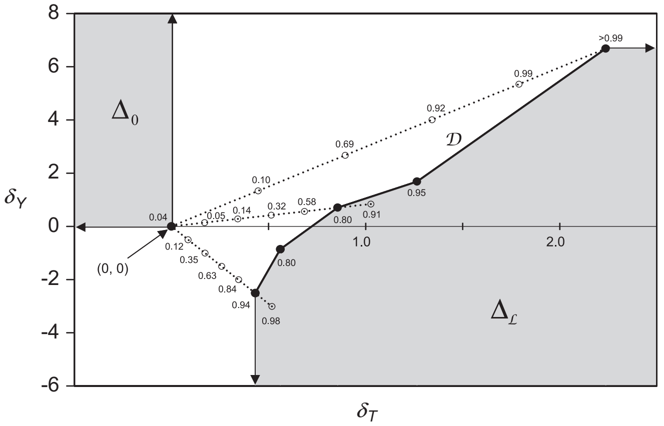

**Fig. 2.** Transformed elicited targets, target line and target set in the $\mathbf{\Delta}$ domain of comparative, covariate-adjusted treatment effects. Values in this figure were obtained from those in Fig. 1 by applying the transformation $\boldsymbol{\theta} \longrightarrow \boldsymbol{\delta} = \tau(\boldsymbol{\theta}) - \boldsymbol{\mu}_{\gamma,0}$. The null set is $\mathbf{\Delta}_0 = \{\boldsymbol{\delta} : \delta_T \le 0, \delta_Y \ge 0\}$.

**Table 1**
Historical population percent and elicited prior event-free survival (EFS) time probabilities for each of 16 patient prognostic subgroups defined in terms of age, whether the tumor was metastatic at first diagnosis (MET), whether surgery achieved a complete resection (CR), and the unfavorable histology choroid plexus carcinoma (CPC).

| Age $\geq 3$ | MET | CR | CPC | Popn. percent | Elicited prior Pr(EFS > $t$ years) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | $t$=0.5 | $t$=1.0 | $t$=2.0 | $t$=3.0 |
| No | No | Yes | Yes | 18.9 | 0.68 | 0.58 | 0.50 | 0.45 |
| No | No | Yes | No | 0 | 0.89 | 0.85 | 0.83 | 0.81 |
| No | No | No | Yes | 26.2 | 0.47 | 0.23 | 0.18 | 0.15 |
| No | No | No | No | 1.7 | 0.68 | 0.66 | 0.60 | 0.50 |
| No | Yes | Yes | Yes | 3.9 | 0.65 | 0.42 | 0.32 | 0.25 |
| No | Yes | Yes | No | 2.4 | 0.86 | 0.65 | 0.54 | 0.50 |
| No | Yes | No | Yes | 6.3 | 0.44 | 0.12 | 0.08 | 0.04 |
| No | Yes | No | No | 1.2 | 0.65 | 0.45 | 0.41 | 0.38 |
| Yes | No | Yes | Yes | 20.1 | 0.87 | 0.82 | 0.79 | 0.70 |
| Yes | No | Yes | No | 0 | 0.94 | 0.91 | 0.89 | 0.82 |
| Yes | No | No | Yes | 9.9 | 0.73 | 0.65 | 0.50 | 0.43 |
| Yes | No | No | No | 2.2 | 0.77 | 0.69 | 0.56 | 0.50 |
| Yes | Yes | Yes | Yes | 0.5 | 0.84 | 0.53 | 0.50 | 0.46 |
| Yes | Yes | Yes | No | 2.2 | 0.90 | 0.70 | 0.60 | 0.55 |
| Yes | Yes | No | Yes | 3.9 | 0.70 | 0.48 | 0.35 | 0.28 |
| Yes | Yes | No | No | 0.7 | 0.74 | 0.62 | 0.36 | 0.29 |

### 2.4. Model and parameters for the CPT trial

The CPT trial design initially was motivated by the physician's desire to use ($Y$,$T$) as the outcome, and to study several experimental regimens. The physician chose to use the probability of toxicity and Pr($T > 24$ months) as the reference parameters, and considered $T$ to be independent of $Y$, so the formulation was simpler than that given above. Specifically, $\beta_{T,q_T+1} \equiv 0$ in (2), and only one set of target pairs was elicited in step 2 of the construction. The covariates used in the CPT trial, given in Table 1, are well-known to be related to event-free survival time for this disease and were specified by the physician. Denoting $I_{\pm}(A) = +1$ if event $A$ occurs and −1 if not, to obtain $\eta_{T,j}(\boldsymbol{Z}_T, Y, \xi)$ and thus $\gamma_{T,j}$, we defined the four binary covariates $Z_{T,1} = I_{\pm}(\text{Age} \geq 3), Z_{T,2} = I_{\pm}(\text{metastatic disease}), Z_{T,3} = I_{\pm}$ (surgery achieved a CR), and $Z_{T,4} = I_{\pm}(\text{histology=CPC})$. Using these, the physician specified the reference patient to be 2 years old, with non-metastatic disease, surgical CR, and CPC histology, denoted by $\boldsymbol{z}^* = (-1, -1, +1, +1)$. In contrast, the probability of toxicity only varied with Age, and was modeled using the logistic fractional polynomial function $\eta_{Y,j} = \text{logit}\{\theta_{Y,j}(\text{Age}, \xi)\} = \gamma_{Y,j} + \beta_{Y,1}\ \text{Age}^{1/2} + \beta_{Y,2}\ \log(\text{Age})$. Details of how this function was determined from elicited values are provided in Section 6.

While eliciting the target pairs for the CPT trial from the physician, we used a graphical representation of preliminary versions of Fig. 1, and modified the figure adaptively as he adjusted some values based on the figure. The process for obtaining the $\boldsymbol{\theta}$ target pairs thus was straightforward. Initially, the shape of $\mathcal{L}$ was surprising. When we questioned this, the physician explained that this $\mathcal{L}$ very accurately reflects his strong belief that a very high risk of toxicity is an acceptable trade-off for a high probability of achieving a longer event-free survival time in CPT patients.

In general, $T$ is not reduced to the binary indicator of the event $(T>t^*)$, and all right-censored event time data $(T^o,\epsilon)$ are included in the likelihood and used to compute the posteriors that determine the test statistic. In the CPT trial, while $\theta_{T,j}(\boldsymbol{Z},Y,\xi)=\overline{F}_j(t^*|\boldsymbol{Z},Y,\xi)$ for reference time $t^*=24$ months is used as the basis for defining targets and hypotheses (described below in Section 3), $T$ is not reduced to the binary indicator of the event $(T>24)$.

## 3. A geometry for comparing treatments

In this section, we provide details of the method for constructing two-dimensional parameter sets that provide a geometric basis for treatment comparison. Given null pair $\boldsymbol{\mu}_{\theta,0}$, a set of elicited target pairs $\vec{\boldsymbol{\theta}}^e$ and target set $\Theta_{\mathcal{L}}$ are *admissible* if (i) $\boldsymbol{\mu}_{\theta,0}\notin\Theta_{\mathcal{L}}$ and (ii) $\theta_Y$ is a strictly increasing function of $\theta_T$ for $\boldsymbol{\theta}\in\mathcal{L}$. We require these admissibility conditions to avoid cases that do not make sense. These conditions can be ensured easily during the target pair elicitation process by showing the physician a plot of $\boldsymbol{\mu}_{\theta,0}$ and the elicited targets. The sample size, Type I error, and generalized power of the test, defined below in Section 5, depend on the shape of $\mathcal{L}$ and the distance between $\boldsymbol{\mu}_{\theta,0}$ and $\mathcal{L}$. Similar to a conventional test of a one-dimensional parameter, in the present setting the sample size required to achieve given Type I error and generalized power would be larger for $\mathcal{L}$ closer to $\boldsymbol{\mu}_{\theta,0}$. Condition (ii) rules out the possibility that some portions of $\mathcal{L}$ may be either horizontal or vertical line segments. For example, if $\mathcal{L}$ were allowed to contain a horizontal line segment, then there would be parameter pairs, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, both on that horizontal segment such that $\theta_T<\theta_T'$ but $\theta_Y=\theta_Y'$. This would say that, despite the fact that $\boldsymbol{\theta}'$ has greater efficacy than $\boldsymbol{\theta}$ and the two pairs have identical toxicity, they are considered equally desirable, which is nonsense. In particular, "staircase" shaped $\mathcal{L}$ are inadmissible.

This construction is illustrated for the CPT trial by Fig. 1. While we assume that $T$ may depend on $Y$, it may be argued either that toxicity $(Y=1)$ decreases the mean EFS time since CPT patients who experience toxicity may have their chemotherapy dose reduced or, instead, that toxicity increases mean EFS time due to a positive association between toxicity and cancer cell killing by the chemotherapy. For the CPT trial, it was not necessary to average two sets of targets, described earlier, since the physician considered $T$ to be independent of $Y$. The historical mean was $\boldsymbol{\mu}_{\theta,0}=(0.47,0.11)$, and the five target pairs were $\boldsymbol{\theta}^{e,1}=(0.65,0.01)$, $\boldsymbol{\theta}^{e,2}=(0.70,0.05)$, $\boldsymbol{\theta}^{e,3}=(0.80,0.20)$, $\boldsymbol{\theta}^{e,4}=(0.90,0.40)$, $\boldsymbol{\theta}^{e,5}=(0.99,0.99)$, considered equally desirable improvements over (0.47, 0.11). Fig. 1 shows the targets, polygonal target line $\mathcal{L}$, target set $\Theta_{\mathcal{L}}$ and null pair $\boldsymbol{\mu}_{\theta,0}$. The trade-off between EFS and toxicity allows Pr(toxicity $|\boldsymbol{z}^*$) to increase substantially over the historical mean 0.11 provided that Pr($T>24$ months$|\boldsymbol{z}^*$) is sufficiently large compared to the null value 0.47. The extreme case $\boldsymbol{\theta}^{e,5}=(0.99,0.99)$ was considered to be as desirable as $\boldsymbol{\theta}^{e,2}=(0.70,0.05)$ where Pr($T>24$ months) has an improvement from 0.47 to 0.70 and Pr(toxicity $|\boldsymbol{z}^*$) drops from 0.11 to 0.05. This was due to the great importance placed on improving EFS time, and the fact that death is accounted for by $T$ and not included in the definition of toxicity.

Fig. 2 shows the structure obtained from that in Fig. 1 by mapping each $\boldsymbol{\theta}$ to a corresponding pair $\boldsymbol{\delta}$ in the covariate-adjusted $E_j$-versus-$S$ effect domain. The null pair $\boldsymbol{\mu}_{\theta,0}=(0.47,0.11),\mathcal{L}$, and $\boldsymbol{\Theta}_{\mathcal{L}}$ in the $\boldsymbol{\Theta}$ domain in Fig. 1 are mapped to $\mathbf{0}=(0,0),\mathcal{D}$, and $\boldsymbol{\Delta}_{\mathcal{L}}$, respectively, in the $\boldsymbol{\Delta}$ domain in Fig. 2. While the vertical line from $\boldsymbol{\delta}^{e,1}$ to $(\delta_T^{e,1},-\infty)$ and the horizontal line from $\boldsymbol{\delta}^{e,5}$ to $(+\infty,\delta_Y^{e,5})$ in Fig. 2 both are portions of the boundary of the transformed target set $\boldsymbol{\Delta}_{\mathcal{L}}$, these lines are *not* portions of the transformed target line $\mathcal{D}=\tau(\mathcal{L})-\boldsymbol{\mu}_{\gamma,0}$. In $\boldsymbol{\Delta}$, all pairs $\boldsymbol{\delta}$ on the transformed target boundary $\mathcal{D}$ are equally desirable. However, pairs on either the vertical or horizontal lines on the boundary of $\boldsymbol{\Delta}_{\mathcal{L}}$ are *not* considered to be as desirable as pairs on $\mathcal{D}$, since this would not make sense.

## 4. A hybrid select-and-test design

### 4.1. Priors

Our formulation utilizes available information on covariate effect parameters, obtained by elicitation or from historical data. In contrast, we assume non-informative iid priors on the treatment effect pairs $\gamma_0,\gamma_1,\ldots,\gamma_J$ because comparative $E_j$-versus-$S$ decisions will be based on the posteriors of $\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_J$. While it may be argued that the prior on $\gamma_0$ should be informative to reflect historical experience with $S$, we do not make this assumption in order to avoid confounding treatment effects based on the trial data with effects of latent variables that may differ in distribution between the trial and historical data. Under the assumption of no treatment–covariate interactions, any differences in the covariate effects $(\boldsymbol{\beta}_T,\boldsymbol{\beta}_Y)$ between prior experience and the trial would be manifested approximately additively in the linear terms (3) and (2) for all $j=0,1,\ldots,J$. Hence such differences would introduce additional variability but each $\boldsymbol{\delta}_j$ will be approximately unbiased. Our formulation ensures that the treatment comparisons will be dominated by the data from the trial, with at most numerically trivial effects due to the non-informative priors on the $\gamma_j$'s.

### 4.2. Hypotheses and trial design

Fig. 2 shows that the null set $\boldsymbol{\Delta}_0$ is the upper left quadrant of $\mathcal{R}^2$, and that $\boldsymbol{\Delta}_0$ and $\boldsymbol{\Delta}_{\mathcal{L}}$ are disjoint. The *global null hypothesis* is $H_0:\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_J\in\boldsymbol{\Delta}_0$. Since larger $\gamma_{T,j}$ and smaller $\gamma_{Y,j}$ are more desirable, under $H_0$ no $E_j$ is better than $S$ since $\gamma_{T,j}\le\gamma_{T,0}$ and $\gamma_{Y,j}\ge\gamma_{Y,0}$. For each $j=1,\ldots J$, we define the *jth alternative hypothesis* $H_{a,j}:\boldsymbol{\delta}_j\in\boldsymbol{\Delta}_{\mathcal{L}}$. While $H_{a,j}$ says that $E_j$ provides an improvement over $S$, it does not say anything about $\boldsymbol{\delta}_{j'}$ for any $j'\neq j$, and in particular $H_{a,1},\ldots,H_{a,J}$ are not disjoint. The *global alternative hypothesis* that at least one $E_j$ provides an improvement over $S$ is $H_a=\bigcup_{j=1}^{J}H_{a,j}$.

During the trial, each patient's $\boldsymbol{Z}$ vector is recorded at enrollment, and a treatment $j$ is randomly chosen using the Pocock-Simon method (1975) to balance on $\mathbf{Z}$. The treatment is administered and the patient's outcome $(Y,T^o,\epsilon)$ is observed. In stage 1, let $data_1$ denote the data available when the decision is made and denote the index of the empirically best experimental treatment at the end of stage 1 by $v = \text{argmax}\{j=1,\ldots,J : R_j(data_1)\}$. Let $M_1$ be the number of events observed and $r_1$ the decision cut-off for the test statistic $R_v(data_1)$. When the decision is made in stage 2, denote the data from all patients in arms $E_v$ and $S$ from both stages by $data_2$, let $M_2$ denote the total number of events observed from both treatment arms $E_v$ and $S$, and let $r_2$ denote the decision cut-off for the test statistic $R_v(data_2)$. Given $(M_1,M_2,r_1,r_2)$, the trial is conducted as follows:

*Stage* 1. Randomize patients among $E_1,\ldots,E_J$ and $S$ until $M_1$ events have been observed. The stage 1 decision is made when $Y$ has been evaluated for all stage 1 patients enrolled up to time $t_1$. If $R_v(data_1) > r_1$ then continue to stage 2; otherwise, terminate the trial and accept $H_0$.

*Stage* 2. Randomize patients between $E_v$ and $S$ until $M_2$ additional events from patients in these two arms have been observed. The final decision is made when $Y$ has been evaluated for all patients. If $R_v(data_2) > r_2$ then accept $H_{a,v}$; otherwise accept $H_0$.

Since at most one $E_j$ may be selected in stage 1, the design allows the $J+1$ possible conclusions $H_{a,1},\ldots,H_{a,J}$, that of the $E_j$'s is sufficiently promising for stage 2 evaluation, or $H_0$, that no $E_j$ is promising and the trial should be stopped. If the trial continues to stage 2, the data from patients enrolled in stage 1 are utilized when computing the stage 2 statistic $R_v(data_2)$ for the final decision. In particular, some of the $M_2$ additional events observed in stage 2 may come from patients enrolled in arms $E_v$ or $S$ in stage 1. Let $t_Y^*$ be the length of the time period from the start of therapy during which toxicity is monitored, which is 4 months for the CPT trial. Because $Y$ is a binary variable based on a 4-month observation period, $Y=1$ may be scored at any time toxicity occurs up to $t_Y^*$ but $Y=0$ may be scored only after the patient has been observed for a period of length $t_Y^*$ without toxicity. The design makes decisions only after $Y$ has been evaluated for all patients in each stage to avoid over-estimating the $\theta_{Y,j}$'s. The waiting period needed to evaluate $Y$ for all patients may be of length up to $t_Y^*$, with the stage 1 decisions made at trial time $t_1 + t_Y^*$. Let $t_1$ denote the time, from the start of the trial, when the $M_1^{st}$ event is observed in stage 1. Ideally, patient enrollment should be suspended at trial time $t_1$ in stage 1 in order to evaluate $Y$ for the last patient accrued, since it may be considered ethically undesirable to continue giving patients experimental treatments that soon may be considered inferior to $S$. If it is not feasible to suspend accrual between stages, accrual is continued and additional failures $T^o = T$ in that period may cause the planned $M_1$ to be overrun slightly by the time $data_1$ is used to make the stage 1 decision. This effect should be negligible in most settings, and it may be accounted for when computing the design's operating characteristics and deriving design parameters, which was done with the CPT trial. Computational details are provided below, in Section 4.5. For the stage 2 decision, the values of $(T^o,\epsilon)$ for stage 1 patients in arms $E_v$ or $S$ for whom $\epsilon = 0$ at the stage 1 decision are updated at the final decision time to be either the patient's extended follow up time $T^o$ if the event still has not occurred $(\epsilon = 0)$ or the patient's observed event time if $T^o = T$ $(\epsilon = 1)$.

### 4.3. *Operating characteristics*

While the design uses Bayesian decision criteria, in practice it must have good frequentist properties. To define overall Type I and Type II error, we generalize the approach taken by TSE in the simpler case of one-dimensional parameters $\theta_0,\theta_1,\ldots,\theta_J \in \mathcal{R}^1$, without covariates. As a frame of reference for what follows, we first briefly review the TSE formulation. The $E_j$-versus-$S$ effects are $\delta_j = \theta_j - \theta_0$ for $j=1,\ldots,J$, the null set is $\boldsymbol{\Delta}_0 = (-\infty,\ 0]$ and the alternative set is $\boldsymbol{\Delta}_a = [\delta^*,\ +\infty)$ for fixed $\delta^* > 0$. The global null hypothesis is $H_0 : \delta_1 = \cdots = \delta_J = 0$, the $j$th alternative is $H_{a,j} : \delta_j \in \boldsymbol{\Delta}_a$, and the global alternative is $H_a = \bigcup_{j=1}^J H_{a,j}$. The Type I error is the probability of concluding $H_a$ when $H_0$ is true. The *generalized power* (GP) is the probability that, for some $j$ such that $H_{a,j}$ is true, one correctly concludes $H_{a,j}$. TSE show that, subject to the requirements (i) $\delta_1,\ldots\delta_J \in \boldsymbol{\Delta}_0 \cup \boldsymbol{\Delta}_a$ and (ii) at least one $\delta_j \in \boldsymbol{\Delta}_a$, a parameter set $\{\delta_1,\ldots,\delta_J\}$ minimizing the GP has exactly one $\delta_j = \delta^*$ and $\delta_{j'} = 0$ for all $j' \neq j$. This is called a *least favorable configuration* (LFC), and it may occur in $J$ ways, one for each $E_j$. The requirement (i) is imposed because no statistical procedure can reliably distinguish between fixed parameter values that are arbitrarily close. Since the indexing of the $J$ experimental treatments is arbitrary, by symmetry the GP is the probability of concluding $H_{a,J}$ under the LFC where $\delta_1 = \cdots = \delta_{J-1} = 0$ and $\delta_J = \delta^*$. The formulation given by TSE is slightly more general in that they consider two improvement values, a clinically insignificant value $\delta_1^* > 0$ and a clinically significant value $\delta_2^* > \delta_1^*$. The above account simplifies the formulation by setting $\delta_1^* = 0$ and writing $\delta_2^* = \delta^*$.

We define Type I error and GP similarly, but now the hypotheses refer to the parameter sets $\boldsymbol{\Delta}_0$ and $\boldsymbol{\Delta}_{\mathcal{L}}$ in $\mathcal{R}^2$. For each $j=1,\ldots,J$, the event that the design concludes $H_{a,j}$ is

$$A_j = \{v(data_1) = j,\ R_j(data_1) > r_1 \text{ and } R_j(data_2) > r_2\}, \tag{5}$$

and the event that the global alternative $H_a$ is accepted is $A = \bigcup_{j=1}^J A_j = \{R_v(data_1) > r_1 \text{ and } R_v(data_2) > r_2\}$. The event that the design accepts $H_0$ is

$$A_0 = \{R_v(data_1) \le r_1\} \cup \{R_v(data_1) > r_1 \text{ and } R_v(data_2) \le r_2\}. \tag{6}$$

Denote a $2J$-vector of $J$ fixed treatment effect pairs by $\vec{\boldsymbol{\delta}}^* = (\boldsymbol{\delta}_1^*,\ldots,\boldsymbol{\delta}_J^*)$. We make a strong distinction between such fixed parameter vectors, which will be used to evaluate the design's properties, and the random parameter vector $\vec{\boldsymbol{\delta}}$ under the

Bayesian model. In our setting, the GP is the probability $\phi(\vec{\boldsymbol{\delta}}^*)$ of correctly concluding $H_{a,j}$ for a true parameter vector $\vec{\boldsymbol{\delta}}^*$ for which $\boldsymbol{\delta}_j^* \in \mathbf{\Delta}_{\mathcal{L}}$. In order to define practical criteria for making decisions and deriving design parameters the following theorem is needed, which may be considered a two-dimensional version of the theorem in (TSE) (1988). Similar to the one-dimensional case, we impose the requirement that all $\boldsymbol{\delta}_j^*$ must be in either $\mathbf{\Delta}_{\mathcal{L}}$ or $\Delta_0$ since no statistical procedure can reliably distinguish between parameters that are arbitrarily close.

**Theorem.** *If* (i) $\boldsymbol{\delta}_1^*, \ldots \boldsymbol{\delta}_J^* \in \mathbf{\Delta}_{\mathcal{L}} \cup \Delta_0$ *and* (ii) *at least one* $\boldsymbol{\delta}_j^* \in \mathbf{\Delta}_{\mathcal{L}}$, *then the GP is minimized if there is exactly one j such that* $\boldsymbol{\delta}_j^* \in \mathcal{D}$ *and* $\boldsymbol{\delta}_{j'}^* = \mathbf{0}$ *for all* $j' \neq j$.

A heuristic proof is given in the Appendix.

The theorem does not specify the location of $\boldsymbol{\delta}_j^*$ on $\mathcal{D}$ since this depends on the particular geometry of $\mathcal{D}$, which is determined by $\boldsymbol{\mu}_{\theta,0}$ and $\vec{\boldsymbol{\theta}}^e$. Since the indices of $E_1, \ldots, E_J$ are arbitrary, let $E_J$ be the treatment providing the improvement, so that $(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^*)$ is the vector of $J$ fixed parameter pairs with $\boldsymbol{\delta}_J^* = \boldsymbol{\delta}^*$ and all other pairs $\mathbf{0}$. The GP can be no smaller than $\min_{\boldsymbol{\delta}^* \in \mathcal{D}} \phi(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^*)$, and we will call such a vector that minimizes the GP a LFC. Similarly, the Type I error is the maximum over $\boldsymbol{\delta}^* \in \mathbf{\Delta}_0$ of the probability of rejecting $H_0$, given by $\alpha^* = \sum_{j=1}^{J} \Pr(A_j \,|\, \vec{\boldsymbol{\delta}}^* = \vec{\mathbf{0}}) = J \times \Pr(A_J \,|\, \vec{\boldsymbol{\delta}}^* = \vec{\mathbf{0}})$, where $\vec{\mathbf{0}}$ is the $2J$-vector with all entries 0. Let $\overline{\alpha}$ be the design's largest allowed probability of a Type I error and $\overline{\phi}$ the smallest allowed probability of making a correct decision under $H_a$. Let $\zeta(\vec{\boldsymbol{\delta}}^*) = \Pr\{R_v(data_1) \le r_1 \,|\, \vec{\boldsymbol{\delta}}^*\}$ denote the probability of early termination after stage 1, and let $\overline{\zeta}$ be the maximum allowed value of $\zeta$ under an LFC. In order to feasibly obtain an optimal design, $\phi$ and $\zeta$ are evaluated at the $m$ transformed target pairs $\{\boldsymbol{\delta}^{e,1}, \ldots, \boldsymbol{\delta}^{e,m}\}$ that determine $\mathcal{D}$. The design requires the Type I error constraint $\alpha^* \le \overline{\alpha}$, the incorrect early stopping probability constraint

$$\zeta^* =_{def} \max_{r=1,\ldots,m} \{\zeta(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^{e,r})\} \le \overline{\zeta} \tag{7}$$

and the GP constraint

$$\phi^* =_{def} \min_{r=1,\ldots,m} \{\phi(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^{e,r}), r = 1, \ldots, m\} \ge \overline{\phi}. \tag{8}$$

The design parameters are thus $\{\overline{\alpha}, \overline{\phi}, \overline{\zeta}, M_1, M_2, r_1, r_2\}$.

### 4.4. Deriving optimal design parameters

While $\{\overline{\alpha}, \overline{\phi}, \overline{\zeta}\}$ are predetermined, $\{M_1, M_2, r_1, r_2\}$ must be derived. As with any group sequential design based on event times where decisions are made when specified numbers of events are observed, the sample sizes $N_1$ and $N_2$ in stages 1 and 2 are random quantities determined by the design's required event counts, $M_1$ and $M_2$, and the underlying accrual and event rates. The total sample size, $N$, is either $N_1$ or $N_1 + N_2$, with mean $\mathrm{E}(N) = \mathrm{E}(N_1) + (1-\zeta)\mathrm{E}(N_2)$, which may be computed under the null $\vec{\boldsymbol{\delta}}^* = \mathbf{0}$ or at an LFC. Subject to the constraints $\phi^* \ge \overline{\phi}$, $\alpha^* \le \overline{\alpha}$ and $\zeta^* \le \overline{\zeta}$, we derive the $\{M_1, M_2, r_1, r_2\}$ values that minimize the equally weighted mean overall sample size $\overline{\mathrm{E}}(N) = \frac{1}{2}\mathrm{E}_{H_0}(N) + \frac{1}{2}\mathrm{E}_{LFC}(N)$, with the second expectation computed under the fixed parameter vector $(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^*)$ that minimizes the GP.

### 4.5. Numerical methods

For each fixed $\boldsymbol{\delta}^*$, we obtained the optimal $\{M_1, M_2, r_1, r_2\}$ for the CPT trial by searching in three nested loops, with $M_1$ in the outer loop, $M_2$ nested within $M_1$, and $(r_1, r_2)$ nested within $M_2$. For each $\{M_1, M_2, r_1, r_2\}$ examined, given the assumed true value $\vec{\boldsymbol{\delta}}^*$, we computed $\phi^*$, $\zeta^*$, $\alpha^*$ and the objective function $\overline{\mathrm{E}}(N)$ by generating 2000 data sets from the likelihood with parameters corresponding to $(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\delta}^*)$. During the search, a design was considered possible if it satisfied the three error constraints. Initially, a small number of $(M_1, M_2)$ pairs were considered. For each $(M_1, M_2)$ pair, a local grid search was done in $(r_1, r_2)$ starting with a coarse grid having increments 0.10 to identify possible solutions, then picking the possible $(r_1, r_2)$ pair giving the smallest value of $\overline{\mathrm{E}}(N)$, then localizing the grid around that pair and refining the grid increments to 0.01. This was repeated with the grid of $(r_1, r_2)$ values refined to increments of 0.001. For each evaluation of $(\phi^*, \zeta^*, \alpha^*, \overline{\mathrm{E}}(N))$ at a pair of $(M_1, M_2)$ values, if there were no $(r_1, r_2)$ pairs giving a possible solution, $M_2$ was incremented to $M_2 + 1$, with this limited to at most 20 steps. Similarly, given $M_1$, if no $M_2$ yielding a possible solution was found then $M_1$ was incremented to $M_1 + 1$, with this repeated until a possible solution was found. Once a set of possible designs were obtained, the $(M_1, M_2)$ grid was localized at the pair $(M_1, M_2)^o$ minimizing $\overline{\mathrm{E}}(N)$ and an exhaustive search was done in a contiguous set of pairs around $(M_1, M_2)^o$.

To simulate each data set when applying the above algorithm, each patient's $\mathbf{Z}$ was sampled from the historical covariate distribution ("Popn. Percent" in Table 1) and a treatment $j$ was then randomly chosen based on $\mathbf{Z}$ using the Pocock–Simon method (1975). The toxicity indicator $Y$ was simulated first using the probability $\theta_{Y,j}(\mathbf{Z}, \boldsymbol{\xi})$, substituting the prior means of $\gamma_{Y,0}$ and the covariate effects $\boldsymbol{\beta}_Y$. The specified $\delta_{Y,j}^*$ determined the fixed intercept as $\gamma_{Y,j}^* = \delta_{Y,j}^* + E(\gamma_{Y,0})$. The event time $[T \,|\, \mathbf{Z}, Y]$ then was simulated under the lognormal model using the prior means of $\gamma_{T,0}, \boldsymbol{\beta}_T, \beta_{T,q_T+1}$ and $\sigma_T^2$, with

$\gamma^*_{T,j} = \delta^*_{T,j} + \mathrm{E}(\gamma_{T,0})$. For each simulated data set, the stage 1 posterior probability ratios $R_1(data_1), \ldots, R_J(data_1)$ were computed, the best stage 1 treatment index $\nu(data_1)$ was identified, the stage 2 test statistic $R_\nu(data_2)$ was computed, and the decision rules were then evaluated. Our construction involves a large number of nuisance parameters. While misspecifying these values may result in error inflation, we have used prior means, since these were the best available numerical values.

The values of $R_j(data_1)$ and $R_j(data_2)$ were computed numerically using Markov chain Monte Carlo with Gibbs sampling (Robert and Cassella, 1999). For prior $p(\xi)$, each generated chain $\{\xi^{(1)}, \ldots, \xi^{(K)}\}$ had $K = 10{,}000$ samples distributed in proportion to the posterior integrand, $\mathrm{Lik}(data,\xi) \times p(\xi)$. The first sample vector $\xi^{(1)}$ in each chain was set equal to the mode of the integrand, found using the Nelder–Mead algorithm (1965).

## 5. Establishing priors for the CPT trial

Since clinical experience with CPT (Wrede et al., 2005, 2007) indicates that the hazard function for EFS is certainly non-constant and very likely non-monotone, we assumed that $[T|\boldsymbol{Z},j]$ is lognormally distributed with $\mathrm{E}\{\log(T)|\boldsymbol{Z},j\} = \gamma_{T,j} + \boldsymbol{\beta}_T\boldsymbol{Z}_T$ and $\mathrm{var}\{\log(T)|\boldsymbol{Z},j\} = \sigma^2_T$. The parameter subvector for the distribution of $T$ is thus $\xi_T = (\gamma_{T,0}, \gamma_{T,1}, \gamma_{T,2}, \gamma_{T,3}, \boldsymbol{\beta}_T, \sigma_T)$.

To determine numerical values of the hyperparameters $\tilde{\xi}_T$ of the prior $p(\xi_T|\tilde{\xi}_T)$, we proceeded as follows. Denote a normal distribution with mean $\mu$ and variance $\sigma^2$ by $\mathrm{N}(\mu,\sigma^2)$. Non-informative iid N(0,100) priors were assumed for the eight treatment effect parameters $(\gamma_{T,j}, \gamma_{Y,j})$ for $j=0,1,2,3$, and for the effect $\beta_{T,5}$ of $Y$ on $T$. To obtain an informative prior on $\boldsymbol{\beta}_T$, values of the EFS probabilities $\overline{F}(t|\boldsymbol{Z}_T,\xi)$ were elicited for each of the 16 configurations of $(Z_{T,1}, \ldots, Z_{T,4})$ and $t$=0.50, 1.0, 2.0 and 3.0 years, with fixed $z^*_{T,5} = y^* = 0$. The elicited values are given in Table 1. We assumed that $\beta_{T,1}, \ldots, \beta_{T,4}$ are normally distributed with common variance, and that $\log(\sigma_T)$ is normal. Indexing the elicitation times by $r = 1,2,3,4$ and the covariate vectors by $l = 1, \ldots, 16$, and regarding each prior mean EFS probability

$$\tilde{\overline{F}}_{r,l}(\tilde{\xi}_T) = E\{\overline{F}(t_r,\boldsymbol{Z}_l|\xi_T)|\tilde{\xi}_T\} = \int \overline{F}(t_r,\boldsymbol{Z}_l|\xi_T)p(\xi_T|\tilde{\xi}_T)\,d\xi_T \tag{9}$$

as a function of $\tilde{\xi}_T$, a generalization of the method of Thall and Cook (2004) was used to solve for $\tilde{\xi}_T$. This was done by treating the elicited probabilities $\overline{F}^{(e)}(t_r|\boldsymbol{Z}_l)$ as observations with means $\tilde{\overline{F}}_{r,l}(\tilde{\xi}_T)$ in a nonlinear regression model parameterized by $\tilde{\xi}_T$ and solving for the values of $\tilde{\xi}_T$ that minimized $\sum_{r,l}\{\overline{F}^{(e)}(t_r|\boldsymbol{Z}_l) - \tilde{\overline{F}}_{r,l}(\tilde{\xi}_T)\}^2$. The resulting normal priors of the covariate parameters have means $\mathrm{E}(\beta_{T,1}, \beta_{T,2}, \beta_{T,3}, \beta_{T,4}) = (0.44,-0.41,0.56,-0.53)$ with common variance $0.15^2$, and the prior on $\sigma^2_T$ is given by $\log(\sigma_T) \sim \mathrm{N}(-0.08,0.14^2)$. These were used as the priors for the model parameters when deriving the optimal designs.

For toxicity, based on clinical experience in treating CPTs only $Z_{T,1} = \mathrm{Age}$ was used as a covariate in the definition of $\theta_{Y,j}$. Toxicity data taking the specific form of $Y$ were not available, however. Consequently, to establish a model for $\theta_{Y,j}(Z_{T,1},\gamma_{Y,0},\boldsymbol{\beta}_Y)$ and an informative prior on $\boldsymbol{\beta}_Y$, toxicity probabilities for each of nine patient ages were elicited independently from each of three oncologists with extensive experience treating patients with CPTs. The sets of elicited toxicity probabilities, shown in Fig. 1, were remarkably consistent in that all three physicians gave values that described the same non-monotone function of Age in which the probability of toxicity is high for infants, very low for patients 3–50 years of age, and moderately higher for older patients. While most CPT patients are very young, the age domain used in the elicitation is appropriate for the trial because its eligibility is not restricted to children, and CPT patients can be as old as 80 years. Denoting the elicited probability at the $r$th age from the $l$th physician by $\theta^{(e)}_{Y,l,r}$, these values were fit as a function of Age by considering mixed models with fractional polynomial linear terms (Royston and Altman, 1994) taking the general form $\mathrm{logit}\{\theta^{(e)}_{Y,l,r}\} = \gamma_{Y,0} + \beta_{Y,1}\,\mathrm{Age}^{p_1} + \beta_{Y,2}\,\mathrm{Age}^{p_2} + \rho^{(l)} + \epsilon_{l,r}$ where $\gamma_{Y,0},\beta_{Y,1},\beta_{Y,2}$ are fixed parameters, $\rho^{(1)},\rho^{(2)},\rho^{(3)}$ are iid $\mathrm{N}(0,\sigma^2_\rho)$ physician effects, and the $\epsilon_{l,r}$'s are iid $\mathrm{N}(0,\sigma^2_\epsilon)$ residuals. One best fitting model to use as a basis for constructing a design was determined following the approach recommended by Sauerbrei and Royston (1999). This was done by varying the exponents $p_1$ and $p_2$ exhaustively over the set $\{-2,-1,-0.5,0,0.5,1,2,3\}$, where the exponent "0" corresponds to log(Age). The 28 models of the above form with $p_1 \neq p_2$, and the eight models not including the term $\beta_{Y,2}\,\mathrm{Age}^{p_2}$, were considered. Each of these 36 models was fit both with and without physician effects, for a total of 72 models. The Bayesian Information Criterion (BIC, Schwarz, 1978) values ranged from 85.5 to 140.2 for the 72 models, with the best fitting model given by the linear term including $\mathrm{Age}^{1/2}$ and log(Age). Fig. 3 gives the plots of these three sets of elicited toxicity probabilities, along with the posterior mean fitted curve under the best fitting fractional polynomial model. To obtain the informative distributions on $\beta_{Y,1}$ and $\beta_{Y,2}$ needed for the trial design, a Bayesian version of this model was fit, dropping the physician effects, with $\gamma_{Y,0},\beta_{Y,1},\beta_{Y,2}$ following N(0,100) priors and $\sigma^2_\rho$ following an inverse gamma prior with mean 1 and variance 1000. The estimated linear term of the fitted model obtained by replacing each parameter with its posterior mean, each subscripted with its standard deviation, is

$$\mathrm{logit}\{\hat{\theta}_Y(\mathrm{Age},\gamma_{Y,0},\beta_{Y,1},\beta_{Y,2})\} = -2.26_{0.46} + 0.78_{0.19}\,\mathrm{Age}^{1/2} - 1.33_{0.27}\,\log(\mathrm{Age}).$$

The posteriors $\beta_{Y,1} \sim \mathrm{N}(0.78,0.19^2)$ and $\beta_{Y,2} \sim \mathrm{N}(-1.33,0.27^2)$ obtained from this fit were used as their priors in formulating the trial design. The fitted models from the elicited data for $Y$ and $T$ gave prior means $\boldsymbol{\mu}_{\theta,0} = (0.47,0.11)$. That is, the mean

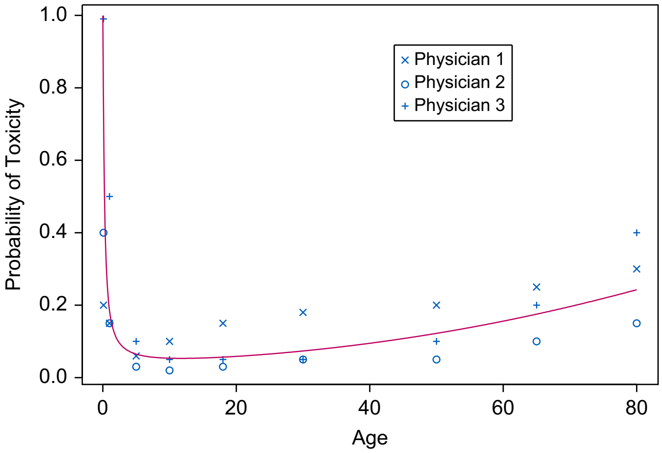

**Fig. 3.** Elicited values of Pr(toxicity) as a function of age, with the posterior mean of the fitted curve under the Bayesian fractional polynomial model.

probability that $T > 24$ months for the reference patient $\boldsymbol{z}^*$ was 0.47, and based on the elicited age-dependent toxicity probabilities the mean probability that a 2-year-old patient would suffer toxicity was 0.11.

## 6. Design parameters for the CPT trial

### 6.1. Optimal designs

The optimal designs with $\overline{\alpha} = 0.05$, $\overline{\phi} = 0.80$ and $\overline{\zeta} = 0.05$ were derived for $J$=1,2, and 3. The case $J$=1 is a two-arm, two-stage design that uses the same geometric construction but only compares one $E$=$E_1$ to $S$ based on $\boldsymbol{\delta} = \boldsymbol{\delta}_1$. In this case, the null hypothesis is $H_0 : \boldsymbol{\delta} \in \boldsymbol{\Delta}_0$ and there is one alternative, $H_{a,1} = H_a : \boldsymbol{\delta} \in \boldsymbol{\Delta}_{\mathcal{L}}$. This case was included as a basis for comparison to assess the effects of having to perform a selection in stage 1.

The optimal designs are given in Table 2. It is important to bear in mind that the test cutoffs $r_1$ and $r_2$ pertain to the posterior probability ratios $R(data_1)$ and $R(data_2)$, and not to conventional $Z$-scores. The numbers of events required to make decisions in each stage, $M_1$ and $M_2$, increase substantially with $J$ due to the facts that, with larger $J$, more treatments are evaluated and the stage 1 decision includes selection among the $E_j$'s as well as comparison to $S$. This produces increases in overall expected sample size and trial duration with $J$, although these values also vary greatly with the case under which they were computed. The GP, $\phi^*$, follows a U-shaped pattern as the true $\boldsymbol{\theta}^*$ moves along $\mathcal{L}$, as illustrated by Fig. 2, with $\phi^*$ smallest at $\boldsymbol{\theta}^{e,2} = (0.70, 0.05)$ and the GP values of $\boldsymbol{\theta}^{e,2}$, with $\boldsymbol{\theta}^{e,3}$ very close to each other in each case. This makes intuitive sense in terms of the centered targets $\boldsymbol{\delta}^{e,j} = \boldsymbol{\theta}^{e,j} - (0.47, 0.11)$, shown in Fig. 2, since $\boldsymbol{\delta}^{e,2}$ is closest to (0,0).

In addition to the GP values $\phi^*$ at the null and the five target pairs, Fig. 2 also shows values of $\phi^*$ for $\boldsymbol{\delta}$ pairs on lines connecting (0,0) to three of the targets, and for also values of $\boldsymbol{\delta}$ in the interior of $\boldsymbol{\Delta}_{\mathcal{L}}$. Considered together, the numerical values of $\phi^*$ in Fig. 2 illustrate that $\phi^*(\boldsymbol{\delta})$ increases monotonically as $\boldsymbol{\delta}$ moves away from the null pair (0,0) toward any target pair. The response surface determined by $\phi^*$ as a function of $\boldsymbol{\delta}$ may be considered a generalization of a conventional power curve based on a single parameter.

### 6.2. Robustness

To examine robustness of the method to the assumed lognormal distribution for $T$, we evaluated the design's behavior with $T$ contaminated by multiplicative noise. For the design with $J$=3, in each case of true $\boldsymbol{\theta}^*$ in Table 2 we simulated samples of the event times with each $T_i$ replaced by $G_iT_i$, where the $G_i$'s were independent gamma random variables with mean 1 and variance $\sigma_G^2$. This had the effects of both changing the distribution of $T_i$ and increasing its variability. Denoting $E(G_i) = \mu_G$ and $E(T_i) = \mu_T$, since

$$\mathrm{var}(G_iT_i) = \sigma_T^2\sigma_G^2 + \sigma_T^2\mu_G^2 + \sigma_G^2\mu_T^2 = \sigma_T^2(\sigma_G^2 + \mu_G^2 + \sigma_G^2\mu_T^2/\sigma_T^2),$$

the effect of multiplying $T_i$ by $G_i$ is to increase the standard deviation $\sigma_T$ by the multiplicative factor $\sigma_{GT}/\sigma_T = (\sigma_G^2 + \mu_G^2 + \sigma_G^2\mu_T^2/\sigma_T^2)^{1/2}$. This factor depends not only on the variances $\sigma_T^2$ and $\sigma_G^2$ but also the values of $\mu_G^2$ and $\mu_T^2$. Table 3 gives the values of $\phi^*$, PET, and E($N$) for $\sigma_G = 0$, corresponding to the assumed lognormal distribution, and $\sigma_G = 0.2$, 0.4, and 0.6, corresponding to successively larger amounts of contamination. As shown in Table 3, for $\sigma_G = 0.4$ or 0.6 the value of $\sigma_{GT}/\sigma_T$ may be quite large, and it varies substantially depending on the value of the true target pair $\boldsymbol{\theta}^*$ where the design is evaluated. The Type I error and GP hold up quite well even when there is substantial contamination, with $\phi^*$ decreasing substantially only at Targets 2, 3, or 4 for large values $\sigma_{GT}/\sigma_T \geq 1.39$. Thus, the method appears to be robust to all but extreme departures from the lognormal assumption.

**Table 2**
The optimal design parameters $(M_1,M_2,r_1,r_2)$ minimize $\overline{\mathrm{E}}(N)$ subject to $\alpha^* \le 0.05$, $\zeta^* \le 0.05$ and $\phi^*(\vec{\boldsymbol{\delta}}^*) \ge 0.80$ for all $\vec{\boldsymbol{\delta}}^* \in \mathcal{D}$. Each target has true $\vec{\boldsymbol{\theta}}^* = (\boldsymbol{\theta}_0^*,\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_J^*)$ with $\boldsymbol{\theta}_J^* = \boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_j^* = \boldsymbol{\theta}_0^* = (0.47,0.11)$ for all $j<J$. PET=Pr(Early termination after stage 1) and E(Dur)=expected trial duration, in years, assuming 30 patients per year accrual.

| Case | True $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1$ | $\phi^*$ | PET | E($N_1$) | E($N_2$) | E($N$) | $\overline{\mathrm{E}}(N)$ | E(Dur) |
|---|---|---|---|---|---|---|---|---|
| $J=1: M_1=14,\ M_2=17,\ r_1=0.510,\ r_2=14.790$ | | | | | | | | |
| Null | (0.47, 0.11) | 0.05 | 0.61 | 57.4 | 17.2 | 64.1 | – | 2.5 |
| Target 1 | (0.65, 0.01) | 0.94 | 0.01 | 62.0 | 30.2 | 91.9 | 78.0 | 2.9 |
| Target 2 | (0.70, 0.05) | 0.80 | 0.05 | 63.0 | 30.3 | 91.7 | 77.9 | 2.9 |
| Target 3 | (0.80, 0.20) | 0.83 | 0.05 | 66.5 | 32.2 | 97.2 | 80.6 | 3.0 |
| Target 4 | (0.90, 0.40) | 0.99 | $<0.01$ | 69.5 | 35.1 | 104.5 | 84.3 | 3.1 |
| Target 5 | (0.99, 0.99) | $>0.99$ | $<0.01$ | 77.1 | 41.2 | 118.0 | 91.2 | 3.4 |
| Case | True $\boldsymbol{\theta}^* = \boldsymbol{\theta}_2$ | $\phi^*$ | PET | E($N_1$) | E($N_2$) | E($N$) | $\overline{\mathrm{E}}(N)$ | E(Dur) |
| $J=2: M_1=19,\ M_2=45,\ r_1=0.431,\ r_2=16.199$ | | | | | | | | |
| Null | (0.47, 0.11) | 0.05 | 0.36 | 66.1 | 53.2 | 99.9 | – | 3.7 |
| Target 1 | (0.65, 0.01) | 0.96 | 0.01 | 69.4 | 67.2 | 135.7 | 117.8 | 4.3 |
| Target 2 | (0.70, 0.05) | 0.82 | 0.03 | 70.2 | 68.0 | 135.9 | 117.9 | 4.3 |
| Target 3 | (0.80, 0.20) | 0.82 | 0.05 | 72.2 | 70.1 | 139.0 | 119.5 | 4.3 |
| Target 4 | (0.90, 0.40) | 0.98 | 0.01 | 75.2 | 75.7 | 149.9 | 124.9 | 4.5 |
| Target 5 | (0.99, 0.99) | $>0.99$ | $<0.01$ | 79.3 | 87.2 | 166.5 | 133.2 | 4.8 |
| Case | True $\boldsymbol{\theta}^* = \boldsymbol{\theta}_3$ | $\phi^*$ | PET | E($N_1$) | E($N_2$) | E(N) | $\overline{\mathrm{E}}(N)$ | E(Dur) |
| $J=3: M_1=28,\ M_2=67,\ r_1=0.605,\ r_2=17.376$ | | | | | | | | |
| Null | (0.47, 0.11) | 0.04 | 0.31 | 80.5 | 79.0 | 134.7 | – | 4.8 |
| Target 1 | (0.65, 0.01) | 0.94 | 0.02 | 83.5 | 90.8 | 172.1 | 153.4 | 5.4 |
| Target 2 | (0.70, 0.05) | 0.80 | 0.05 | 84.3 | 91.8 | 171.2 | 153.1 | 5.4 |
| Target 3 | (0.80, 0.20) | 0.80 | 0.05 | 86.0 | 95.5 | 176.8 | 155.7 | 5.5 |
| Target 4 | (0.90, 0.40) | 0.95 | 0.01 | 88.4 | 102.1 | 189.1 | 161.9 | 5.7 |
| Target 5 | (0.99, 0.99) | $>0.99$ | $<0.01$ | 92.2 | 118.0 | 210.0 | 172.4 | 6.1 |

**Table 3**
Robustness of the optimal design for $J=3$ experimental treatments to multiplicative contamination of $T$ by a gamma noise variable $G$ with mean 1 and standard deviation (sd) $\sigma_G$. The observed event time is $GT$, and the ratio $\sigma_{GT}/\sigma_T = \sigma_{GT}/0.93$ is the multiplicative factor by which the sd of $T$ is increased by $G$.

| Case | $\sigma_G$ | $\sigma_{GT}/\sigma_T$ | $\phi^*$ | PET | E($N$) |
|---|---|---|---|---|---|
| Null | 0 | 1.00 | 0.04 | 0.31 | 134.7 |
| $\mu_T=0.62$ | 0.2 | 1.03 | 0.05 | 0.31 | 134.0 |
| | 0.4 | 1.11 | 0.05 | 0.28 | 136.3 |
| | 0.6 | 1.23 | 0.06 | 0.29 | 132.5 |
| Target 1 | 0 | 1.00 | 0.94 | 0.02 | 172.1 |
| $\mu_T=1.052$ | 0.2 | 1.04 | 0.93 | 0.02 | 171.3 |
| | 0.4 | 1.17 | 0.92 | 0.03 | 168.7 |
| | 0.6 | 1.35 | 0.90 | 0.03 | 165.4 |
| Target 2 | 0 | 1.00 | 0.80 | 0.05 | 134.7 |
| $\mu_T=1.181$ | 0.2 | 1.05 | 0.79 | 0.05 | 134.0 |
| | 0.4 | 1.19 | 0.75 | 0.06 | 136.3 |
| | 0.6 | 1.39 | 0.73 | 0.07 | 132.5 |
| Target 3 | 0 | 1.00 | 0.80 | 0.05 | 176.8 |
| $\mu_T=1.476$ | 0.2 | 1.07 | 0.79 | 0.04 | 176.5 |
| | 0.4 | 1.25 | 0.73 | 0.07 | 170.4 |
| | 0.6 | 1.51 | 0.65 | 0.08 | 166.1 |
| Target 4 | 0 | 1.00 | 0.95 | 0.01 | 189.1 |
| $\mu_T=1.885$ | 0.2 | 1.10 | 0.95 | 0.05 | 188.0 |
| | 0.4 | 1.35 | 0.94 | 0.02 | 185.1 |
| | 0.6 | 1.68 | 0.89 | 0.03 | 179.2 |
| Target 5 | 0 | 1.00 | $>0.99$ | $<0.01$ | 210.0 |
| $\mu_T=2.857$ | 0.2 | 1.19 | $>0.99$ | $<0.01$ | 208.9 |
| | 0.4 | 1.63 | $>0.99$ | $<0.01$ | 205.9 |
| | 0.6 | 2.18 | 0.99 | $<0.01$ | 201.0 |

## 7. Discussion

The proposed design is a hybrid in that Bayesian machinery is used to obtain decision rules, while the design's parameters are derived to ensure that it has specified frequentist properties. The methodology requires a substantial effort to elicit target parameter pairs and priors, and the process of deriving design parameters is computationally intensive. While these may be considered limitations, the two-dimensional covariate-adjusted target set $\Delta_{\mathcal{L}}$ provides a much more refined goal than the conventional approach of increasing median EFS by a given amount while ignoring toxicity and patient covariates. From a practical perspective, the result of applying this design is a sample size small enough to allow the CPT trial to be completed in a realistic timeframe.

There are many other possible approaches to the problem we have addressed here. With regard to controlling error rates, a referee has pointed out that one may take an alternative approach since, if the hypothesis for treatments excluded at the interim analysis are retained, one can consider the disjunctive power to reject a false hypothesis and the familywise error rates to reject a true hypothesis. Aside from the preliminary selection, one might simultaneously use a noninferiority test for safety and a superiority test for efficacy, or use a positive quadrant test, as given by Jennison and Turnbull (1993). However, these methods do not account explicitly for the relative importance of disease-free survival time and toxicity, which was our primary motivation for using the geometric construction, based on elicited target parameter pairs. Alternatively, one might use a utility score to combine the two outcomes, although this would require dealing with the fact that $T$ has domain $[0,\infty)$. One might also use adaptive combination tests (cf. Bretz et al., 2009), which allow a variety of interim decision rules, including Bayesian rules. Our proposed methodology controls the Type I error under the global null, $H_0$, which is a strong requirement since it may be the case that some $\delta_j = 0$ but others not. In this regard, adaptive combination tests guarantee control of the Type I error for any configuration of nulls. As a future exercise, it would be worthwhile to investigate how such alternative approaches compare to that given here.

Our method relies on the strong assumption that there are no treatment–covariate interactions. We assume this in order to control the number of parameters so that we may obtain a feasible design that can actually be used. The model might be extended to include treatment–covariate interactions, for example in terms of subgroups $r = 1,\ldots,K$ determined by $\mathbf{Z}$, by replacing the linear terms with the $2(J+1)K$ treatment-subgroup effects $\eta_{Y,j,r}(\xi)$ and $\eta_{T,j,r}(\xi)$. The number of parameter pairs would then be $K(J+1)$, far too large to implement the methodology practically, since the potential hypotheses and tests would be far more complex. Even with only $K=2$ subgroups, with $J=3$ there would be a total of eight treatment-subgroup parameter pairs.

If the trial's entry criteria are sufficiently restrictive that it is reasonable to assume patients are homogeneous with respect to the outcome $(Y,T)$, then the parametric functions would take the simpler form $\boldsymbol{\theta}_j(Y,\xi) = (\theta_{T,j}(Y,\xi),\theta_{Y,j}(\xi))$ for all patients in treatment arm $j$. In this case, $\eta_{Y,j}(\mathbf{Z},\xi) = \gamma_{Y,j}$ and $\eta_{T,j}(\mathbf{Z},\xi) = \gamma_{T,j} + \beta Y$, and the target pairs that determine the alternative region $\boldsymbol{\Theta}_{\mathcal{L}}$ would be elicited without reference to covariates. Finally, if it is realistic to assume that $T$ is independent of $Y$, as in the CPT trial, then $\eta_{T,j}(\mathbf{Z},\xi) = \gamma_{T,j}$.

## Appendix

For outcomes from the likelihood (1) with parameters corresponding to fixed $\vec{\boldsymbol{\delta}}^*$, the posterior of $\vec{\boldsymbol{\delta}}$ under the Bayesian model has mean $\vec{\boldsymbol{\delta}}^*$, aside from negligible effects of the non-informative iid priors on the treatment effect pairs $\boldsymbol{\gamma}_{j,0}$, $j = 0,1,\ldots,J$, which will be ignored for simplicity. If $\vec{\boldsymbol{\delta}}^*$ contains more than one $\boldsymbol{\delta}_j^*$ that is in $\boldsymbol{\Delta}_{\mathcal{L}}$ then $\phi(\vec{\boldsymbol{\delta}}^*)$ can be reduced by moving each $\boldsymbol{\delta}_j^*$ for which this is the case from $\boldsymbol{\Delta}_{\mathcal{L}}$ to $\boldsymbol{\Delta}_0$ until exactly one $\boldsymbol{\delta}_j^* \in \boldsymbol{\Delta}_{\mathcal{L}}$ remains. If this single $\boldsymbol{\delta}_j^*$ is in the interior of $\boldsymbol{\Delta}_{\mathcal{L}}$ or falls on either the vertical or horizontal lines on $\partial\boldsymbol{\Delta}_{\mathcal{L}}$, then $\phi(\vec{\boldsymbol{\delta}}^*)$ can be reduced by either decreasing $\delta_{j,1}^*$ or increasing $\delta_{j,2}^*$. Consequently, $\boldsymbol{\delta}_j^* \in \mathcal{L}$. Similar reasoning shows that $\phi(\vec{\boldsymbol{\delta}}^*)$ is minimized if the other $J-1$ $\boldsymbol{\delta}_j^*$'s in $\boldsymbol{\Delta}_0$ all equal to $\mathbf{0}$.

## References

Bechhofer, R.E., Santner, T.J., Goldsman, D.M., 1995. Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons. John Wiley & Sons, New York.

Bretz, F., Schmidli, H., Konig, F., Racine, A., Maurer, W., 2006. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biometrical Journal 48, 623–634.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., Posch, M., 2009. Adaptive designs for confirmatory clinical trials. Statistics in Medicine 28, 1181–1217.

Cheung, Y-K., 2008. Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control. Biometrics 64, 940–949.

Jennison, C., Turnbull, B.W., 1993. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. Biometrics 49, 741–752.

Jennison, C., Turnbull, B., 2006. Confirmatory seamless phase II/III clinical trials with hypotheses selection at the interim: opportunities and limitations. Biometrical Journal 48, 650–655.

Lan, K.K.G., DeMets, D.L., 1983. Discrete sequential monitoring boundaries for clinical trials. Biometrika 70, 659–663.

Liu, Q., Pledger, G.W., 2005. Phase 2 and 3 combination designs to accelerate drug development. Journal of the American Statistical Association 100, 493–502.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Computer Journal 7, 308.

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

Pocock, S.J., 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika 64, 191–199.

Pocock, S.J., Simon, R.M., 1975. Sequential treatment assignment with balancing for prognostic covariates in the controlled clinical trial. Biometrics 31, 103–115.

Robert, C.P., Cassella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

Royston, P., Altman, D.G., 1994. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Journal of the Royal Statistical Society, Series C 43, 429–467.

Rubinstein, L.V., Korn, E.L., Freidlin, B., Hunsberger, S., Ivy, S.P., Smith, M.A., 2004. Design issues of randomized phase II trials and a proposal for phase II screening trials. Journal of Clinical Oncology 23, 7199–7206.

Sauerbrei, W., Royston, P., 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society, Series A 162, 71–94.

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

Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Stallard, N., Todd, S., 2003. Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine 22, 689–703.

Stallard, N., Friede, T., 2008. Flexible group-sequential designs for clinical trials with treatment selection. Statistics in Medicine 27, 6209–6227.

Simon, R., Thall, P.F., Ellenberg, S.S., 1994. New designs for the selection of treatments to be tested in randomized clinical trials. Statistics in Medicine 13, 417–429 (discussion pp. 447–451).

Thall, P.F., 2008. A review of phase 2–3 clinical trial designs. Lifetime Data Analysis 14, 37–53.

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

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

Wang, S.K., Tsiatis, A.A., 1978. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 43, 193–199.

Wolff, J.E., Sajedi, M., Brant, R., Coppes, M.J., Egeler, R.M., 2002. Choroid plexus tumours. British Journal of Cancer 87, 1086–1091.

Wrede, B., Liu, P., Ater, J., Wolff, J.E., 2005. Second surgery and the prognosis of choroid plexus carcinoma—results of a meta-analysis of individual cases. Anticancer Research 25, 4429–4433.

Wrede, B., Liu, P., Wolff, J.E., 2007. Chemotherapy improves the survival of patients with choroid plexus carcinoma: a meta-analysis of individual cases with choroid plexus tumors. Journal of Neurooncology 85, 345–351.