Received: 20 November 2021 | Accepted: 4 August 2022

DOI: 10.1111/biom.13738

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Bayesian treatment screening and selection using subgroup-specific utilities of response and toxicity

Juhee Lee[1] | **Peter F. Thall**[2] | **Pavlos Msaouel**[3]

[1]Department of Statistics, Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, California, USA

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

[3]Departments of Genitourinary Medical Oncology and Translational Molecular Pathology, M.D. Anderson Cancer Center, Houston, Texas, USA

Correspondence
Juhee Lee, Department of Statistics, Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA.
Email: juheelee@soe.ucsc.edu

Funding information
National Cancer Institute, Grant/Award Numbers: 1R01CA261978, 5 P30 CA016672 45; U.S. Department of Defense, Grant/Award Number: Concept Award; Kidney Cancer Association, Grant/Award Number: Young Investigator Award; National Science Foundation, Grant/Award Number: DMS-1662427; American Society of Clinical Oncology/Conquer Cancer Foundation, Grant/Award Number: Career Development Award; University of Texas MD Anderson Cancer Center, Grant/Award Number: Khalifa Scholar Award

Abstract

A Bayesian design is proposed for randomized phase II clinical trials that screen multiple experimental treatments compared to an active control based on ordinal categorical toxicity and response. The underlying model and design account for patient heterogeneity characterized by ordered prognostic subgroups. All decision criteria are subgroup specific, including interim rules for dropping unsafe or ineffective treatments, and criteria for selecting optimal treatments at the end of the trial. The design requires an elicited utility function of the two outcomes that varies with the subgroups. Final treatment selections are based on posterior mean utilities. The methodology is illustrated by a trial of targeted agents for metastatic renal cancer, which motivated the design methodology. In the context of this application, the design is evaluated by computer simulation, including comparison to three designs that conduct separate trials within subgroups, or conduct one trial while ignoring subgroups, or base treatment selection on estimated response rates while ignoring toxicity.

KEYWORDS
Bayesian design, clustering, patient prognostic subgroups, treatment screening design, utility function

1 | INTRODUCTION

A new Bayesian design is proposed for randomized phase II group sequential trials in settings where it is desired to screen a set of experimental treatments, $E_1, \ldots, E_K$, compared to an active control treatment, C, based on ordinal categorical response, Y_R, and toxicity, Y_T, and patients are classified at enrollment into ordinal prognostic subgroups. A utility function of $\boldsymbol{Y} = (Y_R, Y_T)$ is used to quantify risk–benefit trade-offs. All decisions are subgroup specific, with screening rules that drop any E_k in any subgroup where it has an unacceptably low response rate or high toxicity rate. The goal is to select a best acceptable E_k, if it exists, in each subgroup, for future confirmatory evaluation in a phase III trial based on survival or progression-free survival time. This prospective evaluation of subgroup-specific effects is intended to obviate *post hoc* assessments within

selected subgroups, which may be considered data dredging, when evaluating treatments and planning future trials.

The design is motivated by a randomized phase II trial of first-line therapy using targeted agents for metastatic clear cell renal cell carcinoma (mRCC). The trial compares two new immunotherapy regimens, nivolumab plus ipilimumab (N+I) (Motzer et al., 2019) and pembrolizumab plus lenvatinib (P+L) (Motzer et al., 2021), with the active control targeted agent pazopanib (Pa) (Tannir et al., 2020). Patients with mRCC often are classified using their International Metastatic Renal-Cell Carcinoma Database Consortium (IMDC) prognostic risk scores, which are defined using biomarkers and clinical variables, including anemia, thrombocytosis, neutrophilia, hypercalcemia, performance status, and time from diagnosis to treatment. IMDC scores are used to classify patients into three prognostic subgroups: favorable (IMDC = 0), intermediate (IMDC = 1 or 2), and poor (IMDC ≥ 3) (Heng et al., 2013). The ability of these prognostic subgroups to account for outcome heterogeneity in mRCC has led to their use as stratification factors in randomized phase III trials in mRCC (Motzer et al., 2021; Rini et al., 2019). IMDC risk score also is recommended by the National Comprehensive Cancer Network (NCCN) guidelines to guide treatment selection for mRCC (Msaouel et al., 2021). However, to date there exist no randomized trial data comparing N+I and P+L to Pa to determine which of these three regimens may be the most appropriate first-line mRCC therapies for each IMDC subgroup (Adashek et al., 2020).

Our proposed design was motivated by the unmet need for a randomized comparison of the approved first-line mRCC therapies, including treatment screening and selection for each IMDC subgroup. To be practical, particularly in settings such as mRCC where pivotal phase III trials have already led to regulatory approval of the treatments of interest, the design must efficiently inform subgroup-specific treatment decisions without necessitating large sample sizes. A prior may be established using a combination of elicited values and historical data. For our application, prior information from the previously conducted phase III trials was used. Details are provided in Supporting Information B. When prior information is sparse, a weakly informative prior can be used. The design produces highly efficient and accurate screening and selection by borrowing information across subgroups through a hierarchical model including latent subgroup membership variables, illustrated in Section 5.

We index the ordered prognostic subgroups by $g \in \{1, \ldots, G\}$, with $g = 1$ denoting the best and $g = G$ the worst prognosis. While we assume ordinal subgroups to reflect the use of IMDC risk score to define subgroups in the mRCC trial, this restriction may be dropped in settings with nonordered subgroups, for example, if subgroups correspond to histologic subtypes. The randomization is restricted to balance the sample sizes of the treatments within each subgroup. In the mRCC trial, Y_T is a binary indicator of toxicity, and Y_R is ordinal with possible values 0 for progressive disease (PD), 1 for stable disease (SD), 2 for partial response (PR), and 3 for complete response (CR). All decisions are based on posterior quantities characterizing E_k-versus-C effects on $\boldsymbol{Y} = (Y_R, Y_T)$ for each g and $k = 1, \ldots, K$. To reduce notation, when no meaning is lost, we identify treatments by the integers $k = 0$ for C and $k = 1, \ldots, K$ for $E_1, \ldots, E_K$.

To enhance flexibility and borrow strength between subgroups, the design uses model-based clustering of the predefined subgroups to adaptively combine two or more adjacent subgroups if the interim or final data show that they have similar estimated outcome distributions. For example, if the distributions of $[\boldsymbol{Y}|k, g]$ and $[\boldsymbol{Y}|k, g+1]$ are determined to be similar for all treatments $k = 0, 1, \ldots, K$, then subgroups g and $g+1$ are combined to form the cluster $\{g, g+1\}$, and the model parameterization is reduced accordingly. The Bayesian model yields a posterior distribution over all possible clustering configurations. Averaging over the posterior cluster distribution, rather than selecting one optimal clustering, ensures that the design's decisions account for uncertainty in the subgroup clustering. Treatment selection is based on an elicited utility function, $U_g(\boldsymbol{Y})$, that quantifies the risk–benefit trade-off between Y_R and Y_T for each subgroup $g = 1, \ldots, G$. At the end of the trial, a best acceptable E_k is chosen for each g by maximizing the posterior predictive (PP) mean of $U_g(\boldsymbol{Y})$.

Designs that do both screening and selection of multiple experimental treatments in the same trial have been considered by many authors in a variety of settings. In the context of a simulation study comparing outcome adaptive to fair randomization in multiarm trials with a binary response outcome Y_R, Wathen and Thall (2017) listed design components that may be varied. These include outcomes, choice of interim decision rules, whether sample sizes of open treatment arms are enriched if inferior arms are dropped early, selection criteria, and whether a concurrent control arm is included as a comparator. Multiarm studies based on a binary response indicator Y_R that do screening and selection using estimates of the comparative effects $\Pr(Y_R = 1|k) - \Pr(Y_R = 1|0)$ for each $k = 1, \ldots, K$ often are called *platform trials*. Rossell et al. (2007) proposed a Bayesian decision theoretic platform design that may drop treatments or enter new treatments interimly, with final decisions of whether a treatment should be studied in a subsequent phase III trial. Other examples of platform trials include a Bayesian design for studying

molecularly targeted agents (Yuan et al., 2016), a design for evaluating combination therapies (Kaizer et al., 2018), and a method for adding new E_k's during the trial (Ventz et al., 2018), among many others.

Several phase II designs based on bivariate binary (Y_R, Y_T) have been proposed, including the single-arm phase II designs of Conaway and Petroni (1995) and Chen and Chi (2012) to evaluate one experimental treatment. Buzaianu et al. (2022) proposed a randomized phase II design to evaluate multiple experimental treatments, using stochastic curtailment to define a closed sequential procedure. However, none of these designs includes a randomized control arm, accounts for patient heterogeneity, or accommodates ordinal outcomes.

Our proposed design may be considered a platform trial, or a randomized multiarm controlled phase II trial with a bivariate ordinal outcome, that evaluates all E_k's from the start, accounts for prognostic subgroups, does adaptive subgroup clustering, includes interim rules to drop unsafe or ineffective E_k's within subgroups, and bases final treatment selections on PP mean utilities. Since each of these design elements has been used previously, the novelty of our design lies in the fact that it includes all of these features, and in the way that it combines them.

In Section 2, we describe our Bayesian probability model, prior specification, subgroup clustering method, and posterior computation. Section 3.1 describes the utility functions and trial design. A heuristic method for determining maximum sample size is described in Section 4. A simulation study to evaluate the proposed design's operating characteristics (OCs) and compare it to three simpler designs is presented in Section 5. We conclude with a brief discussion in Section 6.

2 | PROBABILITY MODEL

2.1 | Sampling distribution

Let $n(t)$ denote the number of patients accrued up to trial time t, and index patients by $i = 1, \dots, n(t)$. For the ith patient, denote treatment by $\tau_i \in \{0, 1, \dots, K\}$, subgroup by $x_i \in \{1, \dots, G\}$, and outcomes by $\boldsymbol{Y}_i = (Y_{i,T}, Y_{i,R})$, where $Y_{i,j} \in \{0, \dots, M_j - 1\}$ for $j = T, R$. In the mRCC trial, $Y_{i,T} \in \{0, 1\}$ with $M_T = 2$, and $Y_{i,R} \in \{0, 1, 2, 3\}$ with $M_R = 4$.

For convenience, we denote each model's parameter vector by θ. We specify a model for regression of $\boldsymbol{Y}_i$ on (τ_i, x_i) by defining $\boldsymbol{Y}_i$ in terms of latent real-valued bivariate normal variables, $\boldsymbol{Z}_i = (Z_{i,T}, Z_{i,R})$. We assume $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_{n(t)}$ are mutually independent given θ, and construct a model for each $\boldsymbol{Z}_i$ by introducing real-valued latent patient-specific random effect vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i,T}, \epsilon_{i,R})$, assuming $\boldsymbol{\epsilon}_i | \Omega \overset{iid}{\sim} N_2(\boldsymbol{0}, \Omega)$, where $\boldsymbol{0} = (0, 0)'$ and Ω is a random 2×2 variance–covariance matrix. For each $i = 1, \dots, n(t)$, we assume conditional independence of $Z_{i,T}$ and $Z_{i,R}$ given $\boldsymbol{\epsilon}_i$, with marginals

$$Z_{i,j} | x_i = g, \tau_i = k, \epsilon_{i,j}, \theta \overset{indep}{\sim} N(\mu_{j,k,g} + \epsilon_{i,j}, \sigma^2), \quad \text{for } j = T, R, \tag{1}$$

and fixed σ^2. We will specify priors for $\{\mu_{j,k,g}\}$ and Ω below. Patient random effects similar to $\{\boldsymbol{\epsilon}_i\}$ have been used widely to model multivariate data in a wide variety of settings. See, for example, Gorfine and Hsu (2011) or Lee et al. (2019, 2021).

For each outcome $j = T, R$ and treatment k, we use cutoffs $u^k_{j,0} < u^k_{j,1} < \cdots < u^k_{j,M_j}$ to define $Y_{i,j} = m$ if and only if $u^k_{j,m} < Z_{i,j} \le u^k_{j,m+1}$ for $y_j = 0, \dots, M_j - 1$, giving the marginal conditional distribution

$$\begin{aligned} P(Y_{i,j} = m | x_i = g, \tau_i = k, \epsilon_{i,j}, \theta) &= \Phi_1\left(u^k_{j,m+1} | \mu_{j,k,g} + \epsilon_{i,j}, \sigma^2\right) \\ &\quad - \Phi_1\left(u^k_{j,m} | \mu_{j,k,g} + \epsilon_{i,j}, \sigma^2\right), \end{aligned} \tag{2}$$

where Φ_d denotes the cumulative distribution function (cdf) of a d-variate normal distribution. Due to the assumptions on the distributions of $Z_{i,j}$, the $\boldsymbol{Y}_i$s are also mutually independent given θ, and $Y_{i,T}$ and $Y_{i,R}$ are conditionally independent given $\boldsymbol{\epsilon}_i$ and θ.

We let the cutoffs $\{u^k_{j,m}, m = 2, \dots, M_j - 1\}$ be random for flexibility, set $u^k_{j,1} = 0$ for all (j, k) to avoid nonidentifiabilty of the model, and also set $u^k_{j,0} = -\infty$ and $u^k_{j,M_j} = \infty$, so $\sum_{m=0}^{M_j - 1} P(Y_{i,j} = m | x_i = g, \tau_i = k) = 1$. We will define priors for $\boldsymbol{u} = \{u^k_{j,m}, j = T, R, \ m = 2, \dots, M_j - 1, k = 0, \dots, K\}$ for outcomes with $M_j > 2$ below. Integrating over the distribution of $\boldsymbol{\epsilon}_i$ gives the bivariate normal distribution $\boldsymbol{Z}_i | x_i = g, \tau_i = k \overset{indep}{\sim} N_2(\boldsymbol{\mu}_{k,g}, \Sigma)$ with $\Sigma = \Omega + \sigma^2 I_2$, which in turn implies that

$$\begin{aligned} P(\boldsymbol{Y}_i = \boldsymbol{y} | x_i = g, \tau_i = k, \theta) &= \Phi_2\left(u^k_{T,y_T+1}, u^k_{R,y_R+1} | \boldsymbol{\mu}_{g,k}, \Sigma\right) \\ &\quad - \Phi_2\left(u^k_{T,y_T+1}, u^k_{R,y_R} | \boldsymbol{\mu}_{g,k}, \Sigma\right) - \Phi_2\left(u^k_{T,y_T}, u^k_{R,y_R+1} | \boldsymbol{\mu}_{g,k}, \Sigma\right) \\ &\quad + \Phi_2\left(u^k_{T,y_T}, u^k_{R,y_R} | \boldsymbol{\mu}_{g,k}, \Sigma\right), \end{aligned} \tag{3}$$

for $\boldsymbol{y} = (y_T, y_R)$, where $\boldsymbol{\mu}_{k,g} = (\mu_{T,k,g}, \mu_{R,k,g})$. The $\boldsymbol{\epsilon}_i$s account for between-patient heterogeneity not explained by the (x_i, τ_i)s, and for each i, correlation between $\epsilon_{i,T}$ and $\epsilon_{i,R}$ induces association between $Z_{i,T}$ and $Z_{i,R}$, and thus association between $Y_{i,T}$ and $Y_{i,R}$.

If an alternative model including a shared scalar random effect for $Z_{i,T}$ and $Z_{i,R}$ were assumed, it would induce

a positive within-patient association, while our model induces $\text{corr}(Z_{i,\text{T}}, Z_{i,\text{R}}) = \Omega_{12}/(\sqrt{\Omega_{11}+\sigma^2}\sqrt{\Omega_{22}+\sigma^2})$, which allows both positive and negative association. Because the conditional likelihoods of $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ given ϵ_i are products of the univariate normals in (2), this construction greatly simplifies posterior simulation of θ, which can be done using a Gibbs sampler (Zeger & Karim, 1991). We provide details in Section 2.3 and Supporting Information C.

Temporarily suppress the patient index i. To characterize regression of each outcome Y_j on subgroup and treatment, we define the means of the latent variable distribution in (1) as

$$\mu_{j,k,g} = \eta_{j,k} + \alpha_{j,k,g}, \tag{4}$$

where $\eta_{j,k}$ is the intercept for treatment k and $\alpha_{j,k,g}$ is an additive treatment–subgroup interaction. This model allows a wide variety of treatment–subgroup curves of each outcome, and it provides a flexible basis for subgroup-specific treatment screening and selection. Since the prognostic subgroups are ordinal, we impose the constraints $\alpha_{\text{T},k,1} \le \ldots \le \alpha_{\text{T},k,G}$ and $\alpha_{\text{R},k,1} \ge \ldots \ge \alpha_{\text{R},k,G}$. This induces a stochastic ordering of the distribution of each Y_j in g for each k, with $\text{P}(Y_\text{T} \le y_\text{T}|g,k) \ge \text{P}(Y_\text{T} \le y_\text{T}|g',k)$ and $\text{P}(Y_\text{R} \le y_\text{R}|g,k) \le \text{P}(Y_\text{R} \le y_\text{R}|g',k)$ for $g < g'$. In settings with nonordinal subgroups, such as histological disease subtypes, these constraints may be dropped.

2.2 | Clustering subgroups

To adaptively combine adjacent subgroups that the data show have similar treatment–subgroup interactions, $\{\alpha_{j,k,g}\}$, we do model-based clustering, and assume that treatment effects are identical within each cluster. We implement this using latent cluster membership variables, $\boldsymbol{s} = (s_1, \ldots, s_G)$, where each $s_g \in \{1, \ldots, G\}$. If $s_g = s_{g'}$ for subgroups $g \ne g'$, then these subgroups belong to the same cluster. Similarly to Lee et al. (2021), since the predefined subgroups are ordinal, we set $s_1 = 1$, require $s_1 \le \ldots \le s_G$, and define a prior on $\boldsymbol{s}$ by proceeding sequentially for $g = 2, \ldots, G$. We assume that subgroup $g \ge 2$ is combined with subgroup $g-1$ in a cluster with fixed probability ξ, and is not combined with subgroup $g-1$ with probability $1-\xi$. Formally, $\text{P}(s_g = s_{g-1}|s_{g-1}) = \xi$ and $\text{P}(s_g = s_{g-1}+1|s_{g-1}) = 1-\xi$, and the prior on $\boldsymbol{s}$ is

$$p(\boldsymbol{s}|\xi) = \prod_{g=2}^{G} p(s_g|\xi, s_{g-1}) = \prod_{g=2}^{G} \xi^{\mathcal{I}(s_g = s_{g-1})}(1-\xi)^{1-\mathcal{I}(s_g = s_{g-1})}, \tag{5}$$

denoting the indicator $\mathcal{I}(A) = 1$ if A is true and 0 otherwise. This construction allows only neighboring subgroups to be combined. Let $H \le G$ denote the number of distinct clusters, with $s_G = H$. Our motivating trial has $G = 3$ predefined risk subgroups, so there are four possible cluster configurations, $\boldsymbol{s} = (1,1,1)$, $(1,1,2)$, $(1,2,2)$, or $(1,2,3)$, which define, respectively, $H = 1$, 2, 2, and 3 clusters. For example, in the cluster configuration $\boldsymbol{s} = (1,2,2)$, subgroup 1 has its own cluster $\{1\}$, with $s_1 = 1$, and subgroups 2 and 3 are combined as the cluster $\{2,3\}$, with $s_2 = s_3 = 2$.

In settings where the subgroups are not ordinal, such as disease subtypes, the ordering constraint on $\boldsymbol{s}$ should be dropped, and any clustering method may be used, such as using a Gaussian mixture model (Chapple & Thall, 2018), or a random partition (Xu et al., 2016).

To borrow strength using the clusters, given $\boldsymbol{s}$, we define cluster-specific treatment effects, $\alpha^\star_{j,k,h}$, and assume that all subgroups in a cluster have the same effects, that is, $\alpha_{j,k,g} = \alpha^\star_{j,k,h}$ for all g with $s_g = h$. This implies that, for each E_k, the distribution of $\boldsymbol{Y}$ is the same for all subgroups in a cluster, so the model dimension depends on H. For example, $\boldsymbol{s} = (1,2,2)$ gives clusters $\{1\}$ and $\{2,3\}$ with $H = 2$. The distribution of $\boldsymbol{Y}$ for subgroup 1 has $\mu_{j,k,1} = \eta_{j,k} + \alpha^\star_{j,k,1}$ for (4) since $s_1 = 1$, and the likelihood is

$$\begin{aligned}\text{P}(Y_{i,j} = y_j|\boldsymbol{s} = (1,2,2), x_i = 1, \tau_i = k, \epsilon_{i,j}, \theta) \\ = \Phi_1\left(u^k_{j,y_j+1}|\eta_{j,k} + \alpha^\star_{j,k,1} + \epsilon_{i,j}, \sigma^2\right) \\ - \Phi_1\left(u^k_{j,y_j} \ |\eta_{j,k} + \alpha^\star_{j,k,1} + \epsilon_{i,j}, \sigma^2\right).\end{aligned}$$

The $Y_{i,j}$s for subgroups $g = 2$ and $g = 3$ have the same means, $\mu_{j,k,2} = \mu_{j,k,3} = \eta_{j,k} + \alpha^\star_{j,k,2}$ since $\alpha_{j,k,1} = \alpha_{j,k,2} = \alpha^\star_{j,k,2}$, for $j = T$ or R, and the same conditional likelihoods,

$$\begin{aligned}\text{P}(Y_{i,j} = y_j|\boldsymbol{s} = (1,2,2), x_i = g, \tau_i = k, \epsilon_{i,j}, \theta) \\ = \Phi_1\left(u^k_{j,y_j+1}|\eta_{j,k} + \alpha^\star_{j,k,2} + \epsilon_{i,j}, \sigma^2\right) \\ - \Phi_1\left(u^k_{j,y_j} \ |\eta_{j,k} + \alpha^\star_{j,k,2} + \epsilon_{i,j}, \sigma^2\right).\end{aligned}$$

Our adaptive clustering method uses a distribution over $\boldsymbol{s}$ that stochastically combines adjacent prognostic subgroups having similar treatment effects. Optimal treatments are chosen for subgroups by marginalizing over the posterior of $\boldsymbol{s}$. The practical advantage is that borrowing information through clustering improves estimation of the distributions $P(\boldsymbol{Y}|k,g,\theta)$, which in turn improves reliability of subgroup-specific decision making. Alternatively, one can view each value of $\boldsymbol{s}$ as a model with a particular subgroup clustering, with (5) defining a prior distribution over all possible models. To account for uncertainty about model choice, we average over the models using the

posterior of $\boldsymbol{s}$, rather than conditioning on a selected model. This is reflected in the design's decisions for subgroup-specific treatment screening and selection.

We next define a prior on the vector of cluster-specific treatment effects, $\boldsymbol{\alpha}^\star = \{\alpha^\star_{j,k,h}\}$, conditional on $\boldsymbol{s}$. For identifiability, we set $\alpha^\star_{j,k,1} = 0$ for all j and k, so $\mu_{j,k,g} = \eta_{k,j}$ if $s_g = 1$.

Given $H > 1$ clusters, we assume normal priors with ordering constraints on $\alpha^\star_{j,k,h}$ for each $h > 1$, as follows;

$$
\begin{aligned}
&p(\alpha^\star_{\mathrm{T},k,2}, \ldots, \alpha^\star_{\mathrm{T},k,H} | \boldsymbol{s}, \bar{\alpha}_\mathrm{T}, \upsilon^2_\mathrm{T}) \\
&\quad \propto \prod_{h=2}^{H} \phi_1(\alpha^\star_{\mathrm{T},k,h} | \bar{\alpha}_{\mathrm{T},h}, \upsilon^2_\mathrm{T}) \mathrm{I}(\alpha^\star_{\mathrm{T},k,h} > \alpha^\star_{\mathrm{T},k,h-1}), \\
&p(\alpha^\star_{\mathrm{R},k,2}, \ldots, \alpha^\star_{\mathrm{R},k,H} | \boldsymbol{s}, \bar{\alpha}_\mathrm{R}, \upsilon^2_\mathrm{R}) \\
&\quad \propto \prod_{h=2}^{H} \phi_1(\alpha^\star_{\mathrm{R},k,h} | \bar{\alpha}_{\mathrm{R},h}, \upsilon^2_\mathrm{R}) \mathrm{I}(\alpha^\star_{\mathrm{R},k,h} < \alpha^\star_{\mathrm{R},k,h-1}). \quad (6)
\end{aligned}
$$

The ordering constraints on $\boldsymbol{s}$ and $\boldsymbol{\alpha}^\star_j$, imply that $\alpha_{\mathrm{T},k,1} \leq \ldots \leq \alpha_{\mathrm{T},k,G}$ and $\alpha_{\mathrm{R},k,1} \geq \ldots \geq \alpha_{\mathrm{R},k,G}$ for each treatment k.

2.3 | Prior specification and posterior computation

We complete the prior specification, to account for $\boldsymbol{\eta} = \{\eta_{j,k}\}$, $\boldsymbol{e} = \{e^k_{j,m}\}$, and Ω. For outcome j with $M_j > 2$, we assume $u^k_{j,m+1} = u^k_{j,m} + e^k_{j,m}$, $m = 1, \ldots, M_j - 2$, and let $e^k_{j,m} \overset{indep}{\sim} \mathrm{Ga}(\bar{e}_{j,m}\kappa_j, \kappa_j)$ with fixed prior mean $\bar{e}_{j,m}$ and prior variance $\bar{e}_{j,m}/\kappa_j$. For treatment- and outcome-specific intercepts, we assume $\eta_{j,k} \overset{indep}{\sim} \mathrm{N}(\bar{\eta}_j, w^2_j)$, with $\bar{\eta}_j$ and w^2_j fixed, and let $\Omega \sim$ inv-Wishart(ν, Ω_0) with $\mathrm{E}(\Omega) = \Omega_0/(\nu - 3)$.

The vector of all model parameters is $\theta = (\boldsymbol{\eta}, \boldsymbol{\alpha}^\star, \boldsymbol{e}, \Omega)$, aside from the random subgroup partition $\boldsymbol{s}$. For the renal cancer trial design, the hyperparameters $\tilde{\theta}$ characterizing the priors were established using historical data from Tannir et al. (2020) and Motzer et al. (2013, 2019, 2021), and elicited prior probabilities. General guidelines for establishing fixed hyperparameters are in Supporting Information A. Details of prior calibration for our mRCC trial application are given in Supporting Information B. The proposed procedure of calibrating the prior involves preliminary simulation studies by varying values of $\tilde{\theta}$, which also provides some empirical results on sensitivity analyses of the design's performance to the specification of $\tilde{\theta}$.

The interim data $\mathcal{D}_{n(t)}$ at trial time t include all outcomes and treatment assignments from previously enrolled patients. Given $\tilde{\theta}$ and $\mathcal{D}_{n(t)}$, the joint posterior of θ, the latent subgroup variables $\boldsymbol{s}$, and patient-specific latent random effects $\boldsymbol{\epsilon} = \{\epsilon_i, i = 1, \ldots, n(t)\}$ is

$$
\begin{aligned}
p(\theta, \boldsymbol{s}, \boldsymbol{\epsilon} | \mathcal{D}_{n(t)}, \tilde{\theta}) &\propto \prod_{i=1}^{n(t)} p(\epsilon_i | \Omega) \times \\
&\prod_{j=T,R} p(y_{i,j} | x_i, \tau_i, \epsilon_{i,j}, \theta, \boldsymbol{s}, \tilde{\theta})\, p(\theta | \boldsymbol{s}, \tilde{\theta})\, p(\boldsymbol{s} | \xi). \quad (7)
\end{aligned}
$$

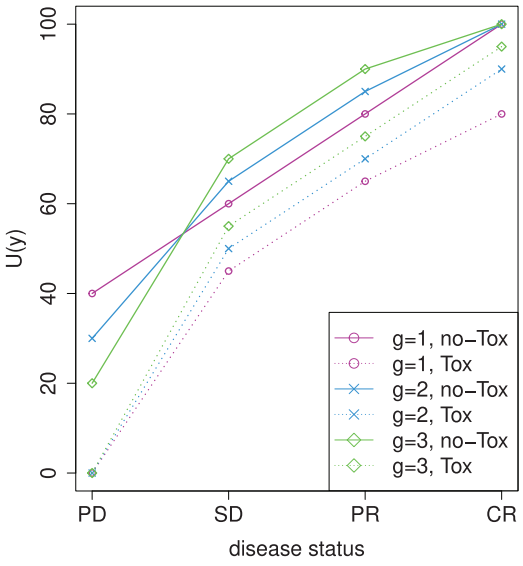
We use Markov chain Monte Carlo (MCMC) simulation to generate posterior samples by iteratively drawing $(\boldsymbol{s}, \theta, \boldsymbol{\epsilon})$, with each conditional on the values of the others at each iteration through $\boldsymbol{s}$, θ, and $\boldsymbol{\epsilon}$. Recall that the values of $\boldsymbol{s}$ define different models and the likelihood in (9) depends on $\boldsymbol{s}$, as illustrated in Section 2.2. Because the dimension of θ changes across the models defined by $\boldsymbol{s}$, we use reversible jump sampling and construct an MCMC simulation that moves among all possible models. The joint posterior of $\boldsymbol{s}$ and θ determines all decision criteria used by the design, and $p(\boldsymbol{s} | \mathcal{D}_{n(t)}, \tilde{\theta})$ is used to average over models. Computational details are given in Supporting Information C. A computer program "Treatment-Screen-Subgroup" for implementing the proposed design is available from the journal's website as Supporting Information.

3 | DECISION CRITERIA AND TRIAL DESIGN

3.1 | Utility function

The utility function accommodates the possibility that clinicians may be more willing to accept a higher risk of toxicity if disease status is likely to be improved for poor risk patients. Thus, the utility function allows risk–benefit preferences between Y_T and Y_R to differ between subgroups, with $U_g(\boldsymbol{Y})$ assigned to outcome $\boldsymbol{Y} = (Y_\mathrm{T}, Y_\mathrm{R})$ for subgroup g.

To apply the design, numerical utilities of the $M_\mathrm{T} \times M_\mathrm{R}$ elementary outcomes must be elicited from the clinical collaborators, with the numerical values reflecting the physicians' beliefs and preferences regarding patients' risk–benefit trade-offs in each subgroup. We illustrate how subgroup-specific utility functions may be established using our motivating trial. We first specify the interval [0, 100] as a convenient domain for numerical utilities, and fix $U_g(0, 3) = 100$ and $U_g(1, 0) = 0$ for all g, since these are the respective utilities for the best and worst possible outcomes. Given these values, for each subgroup g, we elicit intermediate values for the remaining outcomes. The numerical values must satisfy the consistency



		PD	SD	PR	CR
$g = 1$ (Favorable)	no Tox	40	60	80	100
	Tox	0	45	65	85
$g = 2$ (Intermediate)	no Tox	30	65	85	100
	Tox	0	50	70	90
$g = 3$ (Poor)	no Tox	20	70	90	100
	Tox	0	55	75	95

FIGURE 1 Illustration of subgroup-specific utilities U_g of a bivariate outcome $\boldsymbol{Y} = (Y_T, Y_R)$, for subgroups $g = 1, 2, 3$. $Y_T = 0$ and 1 represent no occurrence and occurrence of severe toxicity, respectively. $Y_R = 0, 1, 2$, and 3 represent PD, SD, PR, and CR, respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

conditions $U_g(y_T, y_R) < U_g(y_T, y_R + 1)$ and $U_g(0, y_R) > U_g(1, y_R)$. For the mRCC trial, there are eight elementary outcomes, so six numerical utilities were elicited for each of $G = 3$ subgroups. In therapy of mRCC, patients with favorable IMDC ($g = 1$) have more indolent disease and more time to test other subsequent therapies than patients with more aggressive disease ($g = 2$ or 3), and they are less willing to tolerate toxicity even if CR is achieved. This was reflected by calibrating the utilities across the three prognostic subgroups, as follows. We required that $U_g(y_T, y_R)$ for each (y_T, y_R) with $y_R > 0$ (SD or better) must be nondecreasing in g, to reflect the belief that having some response, that is, no PD, is more desirable for a higher risk subgroup, regardless of y_T. For the same reason, even when $y_R = 3$ (CR is achieved), having $y_T = 1$ is penalized more for a favorable risk subgroup, so $U_g(1, 3)$ increases in g. However, we let $U_g(0, 0)$ decrease in g, while $U_g(1, 0) = 0$ for all g, since having PD is less desirable for a higher risk group. Thus, a treatment with a high toxicity probability is more likely to be optimal for higher risk subgroups if the treatment has a good chance of efficacy.

The numerical utilities elicited for the mRCC trial are illustrated graphically in Figure 1, which includes a table of the utilities.

The mean utility of treating a patient in subgroup g with E_k is

$$\overline{U}_g(k|\boldsymbol{s}, \theta, \tilde{\theta}) = \sum_{y_T=0}^{M_T-1} \sum_{y_R=0}^{M_R-1} U_g(\boldsymbol{y}) \times p(\boldsymbol{y}|k, g, \boldsymbol{s}, \theta, \tilde{\theta}),$$

where $p(\boldsymbol{y}|k, g, \boldsymbol{s}, \theta, \tilde{\theta}) = \int p(\boldsymbol{y}|k, g, \boldsymbol{s}, \theta, \tilde{\theta}, \epsilon) p(\epsilon|\theta, \tilde{\theta}) d\epsilon$ is given in (3).

We use PP mean utilities as criteria for treatment selection. Given data $\mathcal{D}_{n(t)}$ at trial time t, the PP mean utility of giving treatment k to a future patient in subgroup g is

$$\begin{aligned} u_g(k|\mathcal{D}_{n(t)}) &= E\left\{\overline{U}_g(k|\boldsymbol{s}, \theta, \tilde{\theta})|\mathcal{D}_{n(t)}\right\} \\ &= \sum_{\boldsymbol{s}} \int \overline{U}_g(k|\boldsymbol{s}, \theta, \tilde{\theta}) p(\boldsymbol{s}, \theta|\mathcal{D}_{n(t)}, \tilde{\theta}) d\theta. \end{aligned} \tag{8}$$

Because there are no closed forms for the posterior distribution of $(\theta, \boldsymbol{s})$ or the PP distribution of $\boldsymbol{Y}$ for a future patient, numerical approximation is used to compute (11). We approximate $u_g(k|\mathcal{D}_{n(t)})$ using a Monte Carlo sample of $\boldsymbol{s}$ and θ values simulated from the posterior $p(\boldsymbol{s}, \theta|\mathcal{D}_{n(t)}, \tilde{\theta})$. Computational details are given in Supporting Information C.

3.2 | Treatment screening criteria

A key component of a phase II design is a rule to stop accrual to E_k that observed interim data have shown is unsafe. Our design is intended for clinical settings where there is a treatment C with established antidisease activity and safety. We include rules to stop accruing patients to an E_k in any subgroup where the data show that E_k has an unacceptably high rate of toxicity or PD compared to C. If all E_k's are found to be unacceptable for a subgroup, the design stops accrual and does not select any E_k in that subgroup. The design terminates a trial if the accrual is stopped for all subgroups. Recall that $(Y_T = 1) =$ Toxicity and $(Y_R = 0) =$ PD. We denote the probabilities of these events with treatment k in subgroup g by $\zeta_T(k, g)$ and $\zeta_{PD}(k, g)$. Any E_k not satisfying the following rules is considered unacceptable for subgroup g, and accrual to such E_k is stopped in g;

$$P(\zeta_T(k, g) < \zeta_T(0, g)|\mathcal{D}_{n(t)}) < p_T^\star \quad \text{(Safety Criterion)} \tag{9}$$

or

$$\mathrm{P}(\zeta_{\mathrm{PD}}(k,g) < \zeta_{\mathrm{PD}}(0,g)|\mathcal{D}_{n(t)}) < p_{\mathrm{R}}^{\star} \quad \text{(PD Futility Criterion)}. \tag{10}$$

Typically, small positive values between 0.01 and 0.20 are chosen for $p_{\mathrm{T}}^{\star}$ and $p_{\mathrm{R}}^{\star}$. We chose $p_{\mathrm{T}}^{\star} = p_{\mathrm{R}}^{\star} = 0.10$ for the renal cancer trial design based on preliminary simulations. The rule in (12) says that, given the data $\mathcal{D}_{n(t)}$, it is unlikely that the probability of toxicity with E_k is lower than that with C for subgroup g, and (13) says it is unlikely that the probability of PD with E_k is lower than that with C for subgroup g. At the end of the trial, in each subgroup an optimal treatment is selected from the set of acceptable E_k's. In the simpler case where Y_{R} is a binary response indicator with $\zeta_{\mathrm{R}}(k,g) = \Pr(Y_{\mathrm{R}} = 1|k,g,\boldsymbol{s},\boldsymbol{\theta})$, the stopping rule given in terms of PD in (13) may be replaced by the futility stopping criterion, similar to that used by Wathen and Thall (2017). For a fixed value of $\delta_{\mathrm{R}} \geq 0$, the futility rule is

$$\mathrm{P}(\zeta_{\mathrm{R}}(k,g) > \zeta_{\mathrm{R}}(0,g) + \delta_{\mathrm{R}}|\mathcal{D}_{n(t)}) < p^{\star}.$$

(Response Futility Criterion). (11)

Each E_k-versus-C comparison in the screening rules in (12), (13), and (14) is based on randomization between E_k and C, which ensures an unbiased comparison in each rule. This contrasts with the biased comparisons obtained if historical data on C or clinical experience were used to provide fixed numerical values, used in place of $\zeta_{\mathrm{T}}(0,g)$, $\zeta_{\mathrm{PD}}(0,g)$, and $\zeta_{\mathrm{R}}(0,g)$, to construct rules similar to (12), (13), and (14). Our design does not include any provisions for dropping the C arm because this would do away with these advantages.

3.3 | Trial conduct

The design includes L interim analyses done after successive cohorts of size $\lfloor \frac{1}{L+1} N_{\max} \rfloor$ each, with accumulated sample size $n_\ell = \lfloor \frac{\ell}{L+1} N_{\max} \rfloor$ at the ℓth analysis for $\ell = 1, \ldots, L$, with $n_{L+1} = N_{\max}$. For our motivating renal cancer trial, $L = 1$ with one interim analysis done based on data from $n_1 = \lfloor N_{\max}/2 \rfloor$ patients and the final selection performed at $n_2 = N_{\max}$. While in general the design may have $L > 1$ interim looks, $L \geq 3$ may not be logistically feasible, unless the accrual rate is very low. Given the 12-week outcome evaluation window in the mRCC trial, which is typical for solid tumors, we chose $L = 1$ to make the trial feasible. For any $L > 0$, the design parameters should be calibrated carefully, including the $p_j^{\star}$ cutoffs, using simulation to examine OCs and decide how to schedule the interim looks.

Based on (12) and (13) computed using the current data $\mathcal{D}_{n_\ell}$, for subgroup g let $\mathcal{A}_\ell(g)$ denote the set of acceptable E_k's. In each subgroup g, the randomization for cohort $\ell + 1$ is restricted to the current set $\mathcal{A}_\ell(g)$ of treatments that are acceptable for g. Denote the patient entry times by $0 \leq e_1 \leq e_2 \leq \ldots$, and let $n_{k,g}(e_i)$ denote the number of patients in subgroup g given treatment k up to trial time e_i. For each g, the randomization probabilities are proportional to $1/\{n_{k,g}(e_i) + 1\}$. For example, if $K = 2$, $n_{0,g}(e_i) = 1$, $n_{1,g}(e_i) = 0$, and $n_{2,g}(e_i) = 2$, the randomization probabilities in subgroup g are 0.27, 0.55, and 0.18 for treatments $k = 0, 1, 2$, respectively. To obtain a practical design, if g has no acceptable treatments interimly, formally if $\mathcal{A}_\ell(g) = \emptyset$, then enrollment to g is terminated, so $\mathcal{A}_{\ell'}(g) = \emptyset$ for all $\ell' > \ell$. At the end of the trial, for each g the final set $\mathcal{A}_{L+1}(g)$ is computed and the best acceptable E_k is selected. No treatment is selected for g if it has no acceptable treatments, that is, if $\mathcal{A}_{L+1}(g) = \emptyset$. Regardless of what early treatment terminations may be done, the planned overall maximum sample size $N_{\max}$ is maintained rather than being reduced. This has the advantage that, if some treatments are terminated interimly in a subgroup g, the sample sizes of the acceptable treatments in g are enriched, which improves the reliability of the final optimal treatment selection. The trial may be conducted as follows:

Steps for trial conduct

(1) Record the subgroup x_i of the ith patient enrolled at trial time e_i, and randomize fairly among $\{E_1, \ldots, E_K, C\}$ with probability proportional to $1/\{n_{k,g}(e_i) + 1\}$.
(2) At each interim analysis $\ell = 1, \ldots, L$,
   (a) if there is no acceptable E_k for subgroup g, that is, $\mathcal{A}_\ell(g) = \emptyset$, then terminate accrual to g, with $\mathcal{A}_{\ell'}(g) = \emptyset$ for all $\ell' > \ell$ and no E_k is selected in subgroup g;
   (b) if $\mathcal{A}_\ell(g) = \emptyset$ for all g, terminate the trial and do not select any E_k for any g;
   (c) for a patient in the $\ell + 1$st cohort with $x_i = g$, if $\mathcal{A}_\ell(g) \neq \emptyset$, assign the patient to treatment arm $k \in \{0\} \cup \mathcal{A}_\ell(g)$ with probability proportional to $1/\{n_{k,g}(e_i) + 1\}$.
(3) Select a final treatment for each subgroup g with $\mathcal{A}_{L+1}(g) \neq \emptyset$ based on all data $\mathcal{D}_{N_{\max}}$ subject to (12) and (13) using the criterion

$$\tau_{\mathrm{sel}}(g) = \underset{k \in \mathcal{A}_{L+1}(g)}{argmax}\, u_g(k|\mathcal{D}_{N_{\max}}),$$

with $\tau_{\mathrm{sel}}(g) = 0$ denoting the case where no E_k is selected.

4 | DETERMINING SAMPLE SIZE

We recommend determining $N_{\max}$ heuristically based on G, K, the anticipated accrual rate, resource limitations including financial costs and trial duration, and design OCs. A simulation study should be designed that includes scenarios defined in terms of fixed outcome probabilities not computed from the assumed underlying model. One also should specify, for each subgroup g, fixed values corresponding to unacceptably high $\zeta_T(k,g)$ and $\zeta_{PD}(k,g)$, and unacceptably low $\zeta_R(k,g)$. Nominally "good" OCs have reasonably large subgroup-specific probabilities of (1) screening out undesirable E_k's and (2) selecting a desirable E_k, if it exists, for each g, across the scenarios considered.

Most applications should have $G \leq 6$ subgroups and $K \leq 3$ E_k's, but in a given setting some prespecified values of G or K may not be feasible. To determine $(N_{\max}, G, K)$, the following heuristic process may be carried out. This requires preliminary simulations, and to facilitate the process one may use a small number of repetitions for each case, such as $B = 200$ per scenario, and also examine a set of scenarios smaller than the full set that will be given with the final design. First, fix G and K, with the possibility of later reducing them, if necessary. One then may specify two or three feasible $N_{\max}$ values, simulate the trial for each $N_{\max}$, and evaluate the OCs. If a value of $N_{\max}$ does not give a design with good OCs, then smaller values of G, K, or both may be considered, and the design then resimulated, with this process repeated until values of G and K giving good OCs are obtained. A final $(N_{\max}, G, K)$ may be chosen by considering the OCs of all combinations evaluated. For the mRCC trial, given $G = 3, K = 2$, and accrual rate of four patients per month, we evaluated $N_{\max} = 90$ and 180, and chose 90 on that basis since this gives a design with good OCs. If, instead, we had begun with $K = 3$ experimental treatments, then examining the four combinations of $K = 2$ or 3 and $N_{\max} = 90$ or 180 would have been appropriate.

5 | SIMULATION STUDY

5.1 | Simulation design

To evaluate the design's performance, we simulated the renal cancer trial under eight scenarios. For each scenario, we assumed three subgroups, three treatments including a control, and binary toxicity outcomes and four-level ordinal efficacy outcomes, so $G = 3$, $K = 2$, $M_T = 2$, and $M_R = 4$. We simulated each x_i from a trinomial distribution with equal probabilities, 1/3 for each subgroup. Each trial had $N_{\max} = 90$, so 10 patients were expected, on average, for each combination of (k, g). While this maximum sample size may seem large, it is needed to evaluate three treatments reliably. Moreover, if the three subgroups have equal proportions of 1/3 then the expected subsample size in each subgroup is 30, which is close to what is used conventionally. We studied the case with one interim analysis at $n(t) = \lfloor N_{\max}/2 \rfloor$, so $L = 1$. For each scenario, we specified the true clustering of the three predefined subgroups, $\boldsymbol{s}^{\text{true}} = (s_1^{\text{true}}, s_2^{\text{true}}, s_3^{\text{true}})$, the variance $\sigma^{2,\text{true}}$ of the probit scores, the covariance matrix Ω^{true} for the random effect vectors, and $\{\eta_{j,k}^{\text{true}}, j = T, R, k = 0, \ldots, K\}$, and $\alpha_{j,k,h}^{\star,\text{true}}$, $h = 2, \ldots, H^{\text{true}}$ with $\alpha_{j,k,1}^{\star,\text{true}} = 0$, $u_{R,m}^{k,\text{true}}$, $m = 2, 3$ while fixing $u_{j,0}^{k,\text{true}} = -\infty$, $u_{j,1}^{k,\text{true}} = 0$, and $u_{j,M_j}^{k,\text{true}} = \infty$. For each (k, g), we used (3) with assumed true parameter values and computed the probability $\pi_{k,g}^{\text{true}}(\boldsymbol{y})$ of each of the $M_T \times M_R = 8$ possible elementary outcomes. The true values of $\eta_{j,k}^{\text{true}}$, $e_{R,m}^{k,\text{true}}$, and $\alpha_{j,k,h}^{\star,\text{true}}$ are given in Supporting Information Table 3. The true parameter values were specified arbitrarily, not using the design's assumed model, to examine the robustness of our design. Supporting Information Table 4 illustrates $\pi_{k,g}^{\text{true}}(\boldsymbol{y})$ over all $\boldsymbol{y}$ for each k and g. We simulated $\boldsymbol{Y}_i = (Y_{i,T}, Y_{i,R})$ with probability $\pi_{k,g}^{\text{true}}(\boldsymbol{y})$ conditional on $x_i = g$ and $\tau_i = k$.

Table 1 gives true values of the marginal probabilities $\pi_T^{\text{true}}(k,g)$ of severe toxicity, $\pi_{PD}^{\text{true}}(k,g)$ of PD, and of the expected utility, $U_{k,g}^{\text{true}}$, $k > 0$, computed using $\pi_{k,g}^{\text{true}}(\boldsymbol{y})$. For subgroup g, any E_k having $\pi_T^{\text{true}}(k,g) > \pi_T^{\text{true}}(0,g)$ or $\pi_{PD}^{\text{true}}(k,g) > \pi_{PD}^{\text{true}}(0,g)$ is truly unacceptable, and the E_k having maximum $U_{k,g}^{\text{true}}$ is truly optimal. Additional details are given in Supporting Information D.

With $G = 3$, four configurations of $\boldsymbol{s}^{\text{true}}$ are possible due to subgroup ordinality. Scenario 1 has $\boldsymbol{s}^{\text{true}} = (1,1,1)$, Scenarios 2 and 3 have $\boldsymbol{s}^{\text{true}} = (1,1,2)$, Scenarios 4 and 5 have $\boldsymbol{s}^{\text{true}} = (1,2,2)$, and Scenarios 6, 7, and 8 have $\boldsymbol{s}^{\text{true}} = (1,2,3)$. Due to subgroup–treatment interactions and the subgroup-specific utility function, the pattern of true expected utilities across treatments varies with subgroups in all scenarios. In Scenario 1, E_2 is optimal for all subgroups, but the pattern of $U_{k,g}^{\text{true}}$ varies with g. In Scenarios 2 and 4–7, because acceptability of the E_k's and truly optimal treatments vary by subgroups, subgroup-specific decision making is critical. In Scenario 2, E_1 is optimal for subgroups 1 and 2, but E_2 optimal for subgroup 3. In Scenario 3, no E_k is acceptable for any subgroup, so the optimal decision is to stop the trial early and not select either E_k. In Scenario 4, no E_k's are acceptable for subgroup 1, while both E_1 and E_2 are acceptable and E_1 is optimal for subgroups 2 and 3. In Scenario 8, all E_k's are acceptable for all subgroups, and the $U_{k,g}^{\text{true}}$s are very similar across treatments within each subgroup, with E_2 having slightly higher $U_{k,g}^{\text{true}}$ for all g.

We call the proposed design "Sub." In the simulation study, we considered three comparators. "Sep" runs a

TABLE 1 Simulation results

Treatment Arms			C	E_1	E_2	C	E_1	E_2
			Scenario 1 ($\boldsymbol{s}^{\text{true}} = (1,1,1)$)			Scenario 2 ($\boldsymbol{s}^{\text{true}} = (1,1,2)$)		
$\pi_{\text{T}}^{\text{true}}(k,g)$		$g=1$	0.20	0.15	0.15	0.20	0.10	0.15
		$g=2$	0.20	0.15	0.15	0.20	0.10	0.15
		$g=3$	0.20	0.15	0.15	0.46	0.44	0.25
$\pi_{\text{PD}}^{\text{true}}(k,g)$		$g=1$	0.25	0.20	0.10	0.20	0.05	0.15
		$g=2$	0.25	0.20	0.10	0.20	0.05	0.15
		$g=3$	0.25	0.20	0.10	0.46	0.44	0.34
$U_{k,g}^{\text{true}}$		$g=1$		65.65	**73.77**		**78.02**	69.62
		$g=2$		67.96	**77.59**		**81.86**	73.02
		$g=3$		70.02	**80.67**		46.40	**57.69**
$p_{k,g}^{\text{sel}}$	Sub	$g=1$	0.04	0.05	**0.91**	0.02	**0.88**	0.10
		$g=2$	0.05	0.06	**0.90**	0.02	**0.89**	0.09
		$g=3$	0.04	0.09	**0.87**	0.03	0.22	**0.76**
	Sep	$g=1$	0.19	0.25	**0.56**	0.08	**0.85**	0.08
		$g=2$	0.19	0.24	**0.57**	0.07	**0.86**	0.07
		$g=3$	0.18	0.25	**0.57**	0.11	0.18	**0.72**
	Comb	all g	0.05	0.04	**0.92**	0.03	**0.42**	**0.55**
	Eff	$g=1$	0.04	0.10	**0.85**	0.02	**0.84**	0.14
		$g=2$	0.05	0.11	**0.85**	0.02	**0.84**	0.14
		$g=3$	0.04	0.12	**0.84**	0.03	0.25	**0.72**
$p_{k,g}^{\text{safe}}$	Sub	$g=1$		0.92	0.94		0.96	0.92
	& Eff	$g=2$		0.91	0.94		0.97	0.93
		$g=3$		0.91	0.94		0.86	0.94
	Sep	$g=1$		0.75	0.61		0.87	0.74
		$g=2$		0.75	0.61		0.88	0.74
		$g=3$		0.75	0.61		0.69	0.86
	Comb	all g		0.89	0.95		0.96	0.96
$n_{k,g}^{\text{trt}}$	Sub	$g=1$	10.17	9.50	9.93	10.22	9.92	9.54
	& Eff	$g=2$	10.13	9.47	9.88	10.40	9.93	9.72
		$g=3$	10.03	9.47	9.77	10.08	9.46	9.88
	Sep	$g=1$	10.33	8.87	9.79	10.46	9.66	9.12
		$g=2$	10.34	8.85	9.80	10.50	9.65	9.19
		$g=3$	10.33	8.88	9.81	10.32	8.69	10.13
	Comb	$g=1$	8.65	8.08	8.68	8.68	8.29	8.69
		$g=2$	10.87	10.69	11.75	10.91	11.08	11.73
		$g=3$	9.83	9.49	10.03	9.79	9.86	10.02

(Continues)

TABLE 1 (Continued)

Treatment Arms			C	E_1	E_2	C	E_1	E_2
			Scenario 3 ($\boldsymbol{s}^{\text{true}} = (1,1,2)$)			Scenario 4 ($\boldsymbol{s}^{\text{true}} = (1,2,2)$)		
$\pi_{\text{T}}^{\text{true}}(k,g)$		$g=1$	0.10	*0.25*	*0.30*	0.10	*0.20*	*0.25*
		$g=2$	0.10	*0.25*	*0.30*	0.49	0.28	0.58
		$g=3$	0.15	*0.38*	*0.44*	0.49	0.28	0.58
$\pi_{\text{PD}}^{\text{true}}(k,g)$		$g=1$	0.15	*0.25*	*0.20*	0.15	*0.30*	*0.30*
		$g=2$	0.15	*0.25*	*0.20*	0.54	0.39	0.44
		$g=3$	0.22	*0.34*	*0.32*	0.54	0.39	0.44
$U_{k,g}^{\text{true}}$		$g=1$		58.24	63.44		60.83	59.52
		$g=2$		59.90	65.85		**54.77**	46.66
		$g=3$		52.76	55.95		**54.04**	45.32
$p_{k,g}^{\text{sel}}$	Sub	$g=1$	**0.77**	0.10	0.13	**0.30**	0.56	0.14
		$g=2$	**0.79**	0.09	0.12	0.04	**0.91**	0.05
		$g=3$	**0.80**	0.11	0.09	0.03	**0.87**	0.10
	Sep	$g=1$	**0.70**	0.12	0.18	**0.63**	0.31	0.07
		$g=2$	**0.73**	0.10	0.18	0.09	**0.77**	0.14
		$g=3$	**0.82**	0.14	0.04	0.09	**0.73**	0.19
	Comb	all g	**0.85**	0.08	0.06	**0.04**	**0.89**	0.07
	Eff	$g=1$	**0.77**	0.10	0.13	**0.30**	0.50	0.20
		$g=2$	**0.79**	0.09	0.12	0.04	**0.74**	0.23
		$g=3$	**0.80**	0.10	0.11	0.03	**0.74**	0.24
$p_{k,g}^{\text{safe}}$	Sub	$g=1$		*0.16*	*0.14*		*0.63*	*0.44*
	& Eff	$g=2$		*0.14*	*0.12*		0.95	0.69
		$g=3$		*0.16*	*0.11*		0.96	0.69
	Sep	$g=1$		*0.20*	*0.19*		*0.33*	*0.15*
		$g=2$		*0.15*	*0.19*		0.87	0.71
		$g=3$		*0.16*	*0.06*		0.88	0.71
	Comb	all g		*0.11*	*0.07*		*0.93*	*0.69*
$n_{k,g}^{\text{trt}}$	Sub	$g=1$	8.34	*7.24*	*7.09*	10.30	*9.78*	*8.50*
	& Eff	$g=2$	8.14	*7.19*	*7.03*	10.51	10.27	8.83
		$g=3$	8.07	*6.99*	*6.85*	10.54	10.23	8.78
	Sep	$g=1$	8.73	*7.33*	*7.69*	8.70	*7.57*	*7.22*
		$g=2$	8.77	*7.34*	*7.75*	10.64	9.82	8.77
		$g=3$	8.00	*6.86*	*7.04*	10.65	9.83	8.76
	Comb	$g=1$	7.14	*6.37*	*6.26*	9.03	*8.75*	*7.75*
		$g=2$	8.71	*7.45*	*7.40*	11.71	11.62	10.19
		$g=3$	7.89	*6.99*	*6.77*	10.30	10.31	8.90

(Continues)

TABLE 1 (Continued)

Treatment Arms			C	E_1	E_2	C	E_1	E_2
			Scenario 5 ($\boldsymbol{s}^{\text{true}} = (1,2,2)$)			Scenario 6 ($\boldsymbol{s}^{\text{true}} = (1,2,3)$)		
$\pi_{\text{T}}^{\text{true}}(k,g)$		$g=1$	0.20	0.15	0.15	0.05	*0.20*	*0.15*
		$g=2$	0.41	0.39	0.22	0.44	0.37	0.44
		$g=3$	0.41	0.39	0.22	0.54	*0.66*	*0.68*
$\pi_{\text{PD}}^{\text{true}}(k,g)$		$g=1$	0.20	0.05	0.15	0.05	0.20	0.20
		$g=2$	0.28	0.35	0.18	0.44	0.32	0.46
		$g=3$	0.28	0.35	0.18	0.54	0.78	0.70
$U_{k,g}^{\text{true}}$		$g=1$		**79.50**	69.62		66.44	66.36
		$g=2$		55.26	**69.48**		**57.92**	47.89
		$g=3$		55.23	**71.83**		20.13	24.83
$p_{k,g}^{\text{sel}}$	Sub	$g=1$	0.03	**0.71**	0.26	**0.57**	0.23	0.20
		$g=2$	0.03	0.04	**0.93**	0.18	**0.65**	0.17
		$g=3$	0.03	0.05	**0.92**	**0.47**	0.23	0.30
	Sep	$g=1$	0.20	**0.64**	0.16	**0.72**	0.14	0.14
		$g=2$	0.09	0.03	**0.89**	0.12	**0.79**	0.09
		$g=3$	0.08	0.03	**0.88**	**0.69**	0.14	0.17
	Comb	all g	0.03	**0.06**	**0.92**	**0.23**	**0.56**	0.22
	Eff	$g=1$	0.03	**0.71**	0.27	**0.57**	0.22	0.21
		$g=2$	0.03	0.13	**0.84**	0.18	**0.59**	0.23
		$g=3$	0.03	0.12	**0.86**	**0.47**	0.22	0.31
$p_{k,g}^{\text{safe}}$	Sub	$g=1$		0.92	0.93		*0.28*	*0.28*
	& Eff	$g=2$		0.79	0.96		0.74	0.63
		$g=3$		0.78	0.96		*0.31*	*0.38*
	Sep	$g=1$		0.65	0.74		*0.17*	*0.17*
		$g=2$		0.51	0.91		0.86	0.70
		$g=3$		0.51	0.91		*0.19*	*0.19*
	Comb	all g		0.90	0.96		*0.64*	*0.56*
$n_{k,g}^{\text{trt}}$	Sub	$g=1$	10.23	9.67	9.78	9.66	*8.32*	*8.06*
	& Eff	$g=2$	10.45	9.23	10.17	10.20	9.30	8.78
		$g=3$	10.34	9.08	10.12	9.82	*7.97*	*8.32*
	Sep	$g=1$	10.17	9.40	9.11	8.59	*6.93*	*7.57*
		$g=2$	10.31	8.29	10.29	10.44	9.48	8.89
		$g=3$	10.33	8.29	10.29	9.40	*7.14*	*8.17*
	Comb	$g=1$	8.86	7.89	8.88	8.71	*7.64*	*7.34*
		$g=2$	11.16	10.41	12.10	11.02	9.76	9.55
		$g=3$	10.03	9.31	10.29	9.64	*8.90*	*8.45*

(Continues)

TABLE 1 (Continued)

Treatment Arms			C	E_1	E_2	C	E_1	E_2
			Scenario 7 ($\boldsymbol{s}^{\text{true}} = (1,2,3)$)			Scenario 8 ($\boldsymbol{s}^{\text{true}} = (1,2,3)$)		
$\pi_{\text{T}}^{\text{true}}(k,g)$		$g=1$	0.10	*0.20*	*0.20*	0.15	0.15	0.10
		$g=2$	0.54	0.32	*0.61*	0.22	0.22	0.15
		$g=3$	0.80	*0.90*	0.70	0.22	0.22	0.15
$\pi_{\text{PD}}^{\text{true}}(k,g)$		$g=1$	0.10	*0.30*	*0.30*	0.20	0.15	0.20
		$g=2$	0.49	0.44	*0.54*	0.32	0.30	0.32
		$g=3$	0.80	*0.84*	0.68	0.51	0.49	0.51
$U_{k,g}^{\text{true}}$		$g=1$		63.82	62.84		66.70	**68.39**
		$g=2$		**53.76**	40.81		58.59	**61.27**
		$g=3$		12.42	**26.92**		46.19	**47.16**
$p_{k,g}^{\text{sel}}$	Sub	$g=1$	**0.49**	0.32	0.19	0.10	0.29	**0.61**
		$g=2$	0.10	**0.73**	0.17	0.09	0.34	**0.57**
		$g=3$	0.10	0.14	**0.76**	0.08	0.39	**0.53**
	Sep	$g=1$	**0.86**	0.10	0.04	0.15	0.42	**0.43**
		$g=2$	0.12	**0.76**	0.12	0.25	0.35	**0.40**
		$g=3$	0.15	0.10	**0.75**	0.29	0.32	**0.39**
	Comb	all g	**0.09**	**0.33**	**0.58**	0.10	0.25	**0.65**
	Eff	$g=1$	**0.49**	0.27	0.25	0.10	0.34	**0.56**
		$g=2$	0.10	**0.53**	0.37	0.09	0.32	**0.59**
		$g=3$	0.10	0.16	**0.74**	0.08	0.32	**0.60**
$p_{k,g}^{\text{safe}}$	Sub	$g=1$		*0.40*	*0.34*		0.81	0.84
	& Eff	$g=2$		0.85	*0.72*		0.81	0.83
		$g=3$		*0.60*	0.84		0.82	0.84
	Sep	$g=1$		*0.12*	*0.06*		0.70	0.71
		$g=2$		0.85	*0.62*		0.62	0.62
		$g=3$		*0.49*	0.80		0.61	0.60
	Comb	all g		*0.79*	*0.81*		0.80	0.83
$n_{k,g}^{\text{trt}}$	Sub	$g=1$	9.78	*8.75*	*8.24*	10.07	9.30	9.30
	& Eff	$g=2$	10.28	9.80	*8.98*	10.30	9.45	9.55
		$g=3$	10.38	*8.89*	9.77	10.15	9.29	9.45
	Sep	$g=1$	8.55	*7.04*	*7.39*	10.26	8.77	9.36
		$g=2$	10.69	9.69	*8.67*	9.84	8.46	9.26
		$g=3$	10.57	*7.69*	10.28	10.15	8.57	9.36
	Comb	$g=1$	8.98	*7.99*	*8.13*	*8.74*	7.99	8.31
		$g=2$	11.41	10.47	*10.89*	*11.03*	10.52	11.11
		$g=3$	10.01	*9.48*	9.42	9.86	9.39	9.59

Note: $p_{k,g}^{\text{safe}}$ = P(declare E_k safe for subgroup g), $p_{k,g}^{\text{sel}}$ = P(select E_k as optimal for subgroup g), $k = 1, 2$, and $p_{0,g}^{\text{sel}}$ = P(do not choose any E_k as optimal in subgroup g). $n_{k,g}^{\text{trt}}$ = mean number of patients treated with k in subgroup g. Values for *truly unacceptable* and **true optimal** treatments are given in *red italics* and **blue bold**. This table appears in color in the electronic version of this article, and any mention of color refers to that version.

separate trial for each subgroup, "Comb" ignores patient subgroups and makes the same decisions for all patients combined, and "Eff" selects the optimal E_k in each subgroup by maximizing the PP probability of PR or CR. While it may appear that comparison to Eff is unfair, we include it because most phase II designs and platform trials are based on one binary efficacy outcome. Moreover, this comparison assesses the benefit of basing each subgroup-specific treatment selection on utility functions of $(Y_{\text{T}}, Y_{\text{R}})$ rather than Y_{R} alone. While Comb and Sep are simpler than Sub, because they use a utility based on bivariate ordinal $(Y_{\text{T}}, Y_{\text{R}})$ both designs still are more

sophisticated than most randomized phase II screening designs used in practice, as described in Section 1. Comb and Sep are based on the same assumed model used for Sub, but with the key simplification of (4) that no subgroup effects are included. Thus, for these designs $\mu_{j,k} = \tilde{\eta}_{j,k}$ with prior $\tilde{\eta}_{j,k} \overset{indep}{\sim} \mathrm{N}(\bar{\eta}'_j, w_j^2)$, where $\bar{\eta}'_j$ is specified using the elicited probabilities.

The Sep design runs separate trials in the three subgroups, and no information is borrowed between trials. Since $N_{\max} = 90$ in Sub, to ensure a fair comparison, each subgroup-specific trial in Sep has $N_{\max} = 30$ patients, with an interim analysis performed at $n_1 = 15$. Since Comb ignores subgroups, for this design we used $U_2(\boldsymbol{Y})$ as the common utility function. Under Comb, (1) if an E_k is found to be unacceptable then no later patient will be treated with E_k regardless of their subgroup, (2) if all E_k's are identified as unacceptable then the trial is terminated, and (3) a treatment is selected as optimal for all subgroups. The Eff design assumes the same model as Sub, but selects the optimal treatment for subgroup g while ignoring toxicity, using the probability of CR or PR as the criterion,

$$\tau_{\mathrm{sel}}(g) = \underset{k \in \mathcal{A}_{L+1}(g)}{argmax}\ \mathrm{P}(y_{\mathrm{R}} = 2 \text{ or } 3 | k, g, \mathcal{D}_{N_{\max}}).$$

We evaluated the designs using the following subgroup-specific criteria. In subgroup g,

(1) $p^{\mathrm{safe}}_{k,g}$ = probability of declaring E_k safe compared to C, for $k = 1, \ldots, K$;
(2) $p^{\mathrm{sel}}_{k,g}$ = probability of selecting E_k as optimal, for $k = 1, \ldots, K$;
(3) $p^{\mathrm{sel}}_{0,g}$ = probability of not selecting any E_k as optimal;
(4) $n^{\mathrm{trt}}_{k,g}$ = mean number of patients in subgroup g treated with k, for $k = 0, 1, \ldots, K$.

While $p^{\mathrm{safe}}_{k,g}$ and $p^{\mathrm{sel}}_{k,g}$ vary with g under Sub, Sep, and Eff, they are the same for all g under Comb. Since Sub and Eff are the same except that they use different criteria for optimal treatment selection, they have the same values of $p^{\mathrm{safe}}_{k,g}$ and $n^{\mathrm{trt}}_{k,g}$, but different values of $p^{\mathrm{sel}}_{k,g}$ and $p^{\mathrm{sel}}_{0,g}$. Index the simulated trials under each design by $b = 1, \ldots, B$. For the bth trial, let $\tau^{(b)}_{sel}(g) \in \{0, 1, \ldots, K\}$ denote the treatment selected for subgroup g, with $\tau^{(b)}_{sel}(g) = 0$ if no E_k is selected, $w^{(b)}_k(g) = 1$ if treatment E_k is identified as safe for subgroup g and 0 if not, and $N^{(b)}$ the total number of patients treated. For each scenario and design, we summarized the simulation results by the following subgroup-specific sample proportions:

$$p^{\mathrm{safe}}_{k,g} = \frac{1}{B}\sum_{b=1}^{B} w^{(b)}_k(g),\ k = 1, \ldots, K,$$

$$p^{\mathrm{sel}}_{k,g} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{I}(\tau^{(b)}_{sel}(g) = k),\ k = 0, \ldots, K,$$

$$n^{\mathrm{trt}}_{k,g} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N^{(b)}} \mathcal{I}(\tau^{(b)}_i = k \text{ and } x^{(b)}_i = g),\ k = 0, \ldots, K.$$

5.2 | Simulation results

A total of $B = 1000$ trials with $N_{\max} = 90$ were simulated under each scenario. The simulation results are summarized in Table 1. Recall that, in subgroup g, an E_k with $\pi^{\mathrm{true}}_{\mathrm{T}}(k, g) > \pi^{\mathrm{true}}_{\mathrm{T}}(0, g)$ or $\pi^{\mathrm{true}}_{\mathrm{PD}}(k, g) > \pi^{\mathrm{true}}_{\mathrm{PD}}(0, g)$ is truly unacceptable, and the E_k with maximum $U^{\mathrm{true}}_{k,g}$ is truly optimal. Larger differences $\pi^{\mathrm{true}}_j(k, g) - \pi^{\mathrm{true}}_j(0, g)$ for $j = T$ or PD, or the largest $U^{\mathrm{true}}_{k,g}$ minus the second largest, are more meaningful.

Overall, for each subgroup, the Sub design reliably identifies E_k's that are either excessively toxic or have low efficacy compared to C, with small $p^{\mathrm{safe}}_{k,g}$ obtained for truly unacceptable E_k's. When either of $\pi^{\mathrm{true}}_{\mathrm{T}}(k, g)$ or $\pi^{\mathrm{true}}_{\mathrm{PD}}(k, g)$ for E_k is substantively larger than the corresponding value for C, $p^{\mathrm{safe}}_{k,g}$ is particularly small. When E_k is truly unacceptable for all subgroups, $p^{\mathrm{safe}}_{k,g}$ is small for all subgroups, in part because the model improves reliability by borrowing information across clustered subgroups through $\alpha^{\star}_{j,k,s_g}$. For example, under Scenario 3, where all E_k's are unacceptable for all subgroups, Sub yields at most 16% for $p^{\mathrm{safe}}_{k,g}$ for all (E_k, g) and selects no E_k with probabilities 0.77, 0.79, and 0.80 for the three subgroups. In Scenario 8, where $\pi^{\mathrm{true}}_{\mathrm{T}}(k, g) - \pi^{\mathrm{true}}_{\mathrm{T}}(0, g)$ or $\pi^{\mathrm{true}}_{\mathrm{PD}}(k, g) - \pi^{\mathrm{true}}_{\mathrm{PD}}(0, g)$ for E_k are very small or 0, $p^{\mathrm{safe}}_{k,g}$ is at least 0.81 for all E_k and g.

In Scenario 4, E_1 and E_2 are truly unacceptable in subgroup 1 but acceptable for subgroups 2 and 3. Sub incorrectly identifies those arms as acceptable with probabilities 0.63 and 0.44 for subgroup 1, but the differences $\pi^{\mathrm{true}}_{\mathrm{T}}(1, 1) - \pi^{\mathrm{true}}_{\mathrm{T}}(0, 1) = 0.20 - 0.10 = 0.10$ and $\pi^{\mathrm{true}}_{\mathrm{T}}(3, 1) - \pi^{\mathrm{true}}_{\mathrm{T}}(0, 1) = 0.25 - 0.10 = 0.15$ are small, with differences 0.15 for the corresponding PD probabilities. It is unrealistic to expect a screening rule with overall $N_{\max} = 90$ to reliably detect such small differences in a setting with three treatments and three subgroups. However, the $p^{\mathrm{safe}}_{k,g}$ values 0.63 and 0.44 for $k = 1$ and 2 in subgroup $g = 1$ obtained with $N_{\max} = 90$ drop to 0.42 and 0.25, respectively, for the larger sample size $N_{\max} = 180$ (Supporting Information Table 7), illustrating that the reliability of the design's screening rules increases with sample size.

The Sub design selects truly optimal safe treatments with high probabilities. When an arm is optimal in more than one subgroup, $p^{\mathrm{sel}}_{k,g}$ is especially high. For example, in Scenario 1 where E_2 is optimal for all subgroups, Sub

selects E_2 as optimal with probabilities 0.91, 0.90, and 0.87 for subgroups 1, 2, and 3, respectively. In Scenario 2, where E_1 is optimal for subgroups 1 and 2, E_1 is chosen as optimal with probabilities 0.88 and 0.89 for those subgroups. On the other hand, when a subgroup's true optimal treatment is different from that of the other subgroups and the other subgroups have the same optimal treatment arm, $p^{\text{sel}}_{k,g}$ is smaller for the subgroup having a different optimal treatment arm. In Scenario 5 where E_1 is truly optimal for subgroup 1, and E_2 is optimal for subgroups 2 and 3 in Scenario 5, Sub selects the true optimal E_k with probabilities 0.71 for subgroup 1, and 0.93 and 0.92 for subgroups 2 and 3. In Scenario 8, although $U^{\text{true}}_{k,g}$ is similar for all treatments in each subgroup, Sub selects E_2 as optimal with probabilities 0.61, 0.57, and 0.53 for subgroups 1, 2, and 3.

The mean sample sizes $n^{\text{trt}}_{k,g}$ in subgroup g show that the design reliably identifies unacceptable E_k's during a trial and assigns fewer patients to truly unacceptable E_k's, as seen in Scenarios 3, 4, and 6. When any E_k is truly acceptable in those scenarios, more patients were treated at the truly safe E_k or C. Recall that the true values of the parameters in the simulation setup are arbitrarily specified and very different from the prior means. Thus, in terms of all criteria, Sub is robust in that it performs well in a variety of scenarios not matching any particular model.

Probabilities of identifying an E_k as safe and of treatment selection for the comparators, Sep, Comb, and Eff, also are summarized in Table 1. Sub has greatly superior performance compared to these comparators, or very similar performance, in nearly all scenarios. Sep performs similarly to Sub, or slightly better in some scenarios, for example, subgroup 1 in Scenario 4, subgroup 2 in Scenario 6, and Scenario 7. In Scenario 7, where not choosing any E_k, choosing E_1, and choosing E_2 are optimal for subgroups 1–3, respectively, Sub chooses them as optimal with probabilities 0.49, 0.73, and 0.76, compared to 0.86, 0.76, and 0.75 with Sep. However, when more than one subgroup has the same true optimal treatment, Sub often performs much better than Sep. For example, in Scenario 1, where E_2 is truly optimal for all subgroups, $p^{\text{sel}}_{g,2}$ is 0.87 to 0.91 under Sub versus 0.56 to 0.57 under Sep. In Scenario 2, where E_1 is truly optimal in subgroups 1 and 2, while E_2 as truly optimal in subgroup 3, Sub has $p^{\text{sel}}_{k,g}$ of the truly optimal treatments 0.88, 0.89, and 0.76 versus 0.85, 0.86, and 0.72 under Sep, for subgroups 1, 2, and 3. Moreover, when the truly optimal E_k is not selected, Sub tends to select the second optimal E_k more often than Sep. While Sep is far less reliable than Sub in terms of correct selection, Sep is safe, as seen in Scenario 3, where Sep selects no E_k with probabilities 0.70, 0.73, and 0.82 for subgroups 1, 2, and 3.

Comb behaves very poorly when truly optimal treatments differ between subgroups. For example, in Scenario 2, Comb selects E_2 as optimal with probabilities 0.55 for all subgroups, while the true optimal treatment for subgroups 1 and 2 is E_1. In Scenario 4, all E_k's are truly unacceptable for subgroup 1, Comb incorrectly selects E_1 for subgroup 1 with probability 0.89. This because E_1 is truly optimal for subgroups 2 and 3 and Comb ignores subgroups. Moreover, E_1 and E_2 are identified as acceptable with probabilities 0.93 and 0.69 for all subgroups including subgroup 1, when they are truly unacceptable for subgroup 1.

Eff performs similarly to Sub in most scenarios, but in scenarios where true response probabilities are similar across subgroups and toxicity probabilities are substantively different, the comparative performance of Eff is very poor. In Scenario 4, the difference $\pi^{\text{true}}_{\text{PD}}(2,g) - \pi^{\text{true}}_{\text{PD}}(2,g) = 0.44 - 0.39 = 0.05$ for subgroups 2 and 3, but $\pi^{\text{true}}_{\text{T}}(2,g) = 0.58$ versus $\pi^{\text{true}}_{\text{T}}(1,g) = 0.28$. This results in E_1 being truly optimal by a wide margin for both subgroups. Sub accounts for the difference in $\pi^{\text{true}}_{\text{T}}(k,g)$ through the utility function, with true utilities 54.77 for E_1 versus 46.66 for E_2 in subgroup 2, and 54.04 for E_1 versus 45.32 for E_2 in subgroup 3. Thus, Sub correctly selects E_1 as optimal with probabilities 0.91 for $g = 2$ and 0.87 for $g = 3$. Because Eff ignores the much greater toxicity probability of E_2, it incorrectly selects E_2 as optimal with probability 0.74 in $g = 2$ and in $g = 3$. The extremely poor performance of Eff when some treatments have unacceptable $\pi^{true}_{\text{T}}(k,g)$ shows the advantage of using $U_g(Y_{\text{R}}, Y_{\text{T}})$, compared to the conventional practice of using the response rate.

We also examined the effect of increasing $N_{\max}$ to 180, summarized in Supporting Information Table 3. Sub and Sep have improved performances in all scenarios, with smaller differences in $p^{\text{sel}}_{k,g}$, $p^{\text{safe}}_{k,g}$, and $n^{\text{trt}}_{k,g}$. The performances of Comb and Eff do not improve with larger $N_{\max}$, because they ignore information about either patient heterogeneity or treatment toxicity.

As an additional comparator, following a reviewer's suggestion, we included the two-stage phase II trial design of Conaway and Petroni (1995) (CP), which has a bivariate binary response and toxicity outcome. The CP design differs from the Sub design in that it (1) is singe arm and evaluates one experimental treatment, (2) bases decisions on a test of two-dimensional hypotheses with assumed fixed null response and toxicity probabilities $p_{\text{R},0}$ and $p_{\text{T},0}$, and (3) assumes patient homogeneity. Similar phase II designs based on response and toxicity are given by Chen and Chi (2012) and Buzaianu et al. (2022), who also consider multiple E_k's and apply stochastic curtailment. We also added an additional scenario, Scenario 9, to show the effects of differences between the proposed Sub design and existing phase II designs. The simulations, given in

Supporting Information Section E, show that the CP design performs poorly in many scenarios. Ignoring subgroups causes the CP design to make the same decision for subgroups that may have very different treatment effects. Additionally, the CP design's assumed fixed $(p_{R,0}, p_{T,0})$, used as a basis for treatment evaluation in the test of hypotheses, may differ substantially from the empirical probabilities seen in the control arm used as comparator by the Sub design. These limitations also exist for the other phase II designs noted above.

6 | DISCUSSION

The Sub design is necessarily complex because it evaluates multiple treatments, accounts for subgroup effects, does subgroup-specific screening and selection, and does adaptive subgroup clustering. Advantages of this complexity are shown by the simulations. Across nine scenarios, the Sub design is generally superior to all competitors considered, and greatly superior in many cases. The subjectivity of the utility function values is an advantage of the method, since the utility makes risk–benefit trade-offs explicit and provides a basis for including both good and bad outcomes in the selection criterion.

Practical requirements of the Sub design include eliciting subgroup-specific utilities and many prior parameters, specifying complex simulation scenarios, and doing simulations to calibrate prior hyperparameters and determine feasible N_{max}, K, and G that give a design with good OCs. The Sub design replaces KG conventional single-arm phase II trials, one for each (k, g) combination, or K conventional trials if heterogeneity is ignored. Thus, while $N_{max} = 90$ or 180 may seem large, for the mRCC trial Sub replaces three to nine conventional trials. Once the computer program for applying the screening and selection rules is in place, trial conduct is similar to that of a conventional randomized multiarm trial.

The Sub design might be generalized to allow new E_k's to be introduced interimly. While such an extension might be logistically convenient, the design would need to account for biasing effects due comparing nonconcurrent treatment arms, as discussed by Karrison et al. (2003) and illustrated by Freidlin and Korn (2021).

ACKNOWLEDGMENTS

Juhee Lee's research was supported by NSF grant DMS-1662427. Pavlos Msaouel was supported by a Young Investigator Award from the Kidney Cancer Association, a Career Development Award from the American Society of Clinical Oncology/Conquer Cancer Foundation, a Concept Award from the United States Department of Defense, and by the MD Anderson Khalifa Scholar Award. Peter Thall's research was supported by NIH/NCI grants 1R01CA261978 and 5 P30 CA016672 45. The authors thank Mark Conaway for providing optimal parameters for his phase II design to use in the simulations, and two referees for their detailed and constructive comments.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are computer simulated. The computer code that simulates data sets is available from the journal's website as supporting information.

ORCID

Juhee Lee https://orcid.org/0000-0002-9787-3830
Peter F. Thall https://orcid.org/0000-0002-7293-529X
Pavlos Msaouel https://orcid.org/0000-0001-6505-8308

REFERENCES

Adashek, J.J., Genovese, G., Tannir, N.M. & Msaouel, P. (2020) Recent advancements in the treatment of metastatic clear cell renal cell carcinoma: a review of the evidence using second-generation p-values. *Cancer Treatment and Research Communications*, 23, 100166.

Buzaianu, E.M., Chen, P. & Hsu, L. (2022) A curtailed procedure for selecting among treatments with two Bernoulli endpoints. *Sankhya B*, 84, 320–339.

Chapple, A.G. & Thall, P.F. (2018) Subgroup-specific dose finding in phase I clinical trials based on time to toxicity allowing adaptive subgroup combination. *Pharmaceutical Statistics*, 17, 734–749.

Chen, C.-T. & Chi, Y. (2012) Curtailed two-stage designs with two dependent binary endpoints. *Journal of Biopharmaceutical Statistics*, 11, 57–62.

Conaway, M.R. & Petroni, G.R. (1995) Bivariate sequential designs for phase II trials. *Biometrics*, 51, 656–664.

Freidlin, B. & Korn, E. (2021) Platform trials—beware the noncomparable control group. *New England Journal of Medicine*, 384, 1572–1573.

Gorfine, M. & Hsu, L. (2011) Frailty-based competing risks model for multivariate survival data. *Biometrics*, 67, 415–426.

Heng, D.Y., Xie, W., Regan, M.M., Harshman, L.C., Bjarnason, G.A., Vaishampayan, U.N. et al. (2013) External validation and comparison with other models of the International Metastatic Renal-Cell Carcinoma database consortium prognostic model: a population-based study. *The Lancet Oncology*, 14, 141–148.

Kaizer, A.M., Hobbs, B.P. & Koopmeiners, J.S. (2018) A multi-source adaptive platform design for testing sequential combinatorial therapeutic strategies. *Biometrics*, 74, 1082–1094.

Karrison, T., Huo, D. & Chappell, R. (2003) A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials*, 24, 506–522.

Lee, J., Thall, P.F. & Msaouel, P. (2021) Precision Bayesian phase I-II dose-finding based on utilities tailored to prognostic subgroups. *Statistics in Medicine*, 40, 5199–5217.

Lee, J., Thall, P.F. & Rezvani, K. (2019) Optimizing natural killer cell doses for heterogeneous cancer patients based on multiple event times. *Journal of the Royal Statistical Society: Series C*, 68, 809–828.

Motzer, R., Alekseev, B., Rha, S.-Y., Porta, C., Eto, M., Powles, T. et al. (2021) Lenvatinib plus pembrolizumab or everolimus for advanced renal cell carcinoma. *New England Journal of Medicine*, 384, 1289–1300.

Motzer, R.J., Hutson, T.E., Cella, D., Reeves, J., Hawkins, R., Guo, J. et al. (2013) Pazopanib versus sunitinib in metastatic renal-cell carcinoma. *New England Journal of Medicine*, 369, 722–731.

Motzer, R.J., Rini, B.I., McDermott, D.F., Frontera, O.A., Hammers, H.J., Carducci, M.A. et al. (2019) Nivolumab plus ipilimumab versus sunitinib in first-line treatment for advanced renal cell carcinoma: extended follow-up of efficacy and safety results from a randomised, controlled, phase 3 trial. *The Lancet Oncology*, 20, 1370–1385.

Msaouel, P., Lee, J. & Thall, P.F. (2021) Making patient-specific treatment decisions using prognostic variables and utilities of clinical outcomes. *Cancers*, 13, 2741.

Rini, B.I., Plimack, E.R., Stus, V., Gafanov, R., Hawkins, R., Nosov, D. et al. (2019) Pembrolizumab plus axitinib versus sunitinib for advanced renal-cell carcinoma. *New England Journal of Medicine*, 380, 1116–1127.

Rossell, D., Mueller, P. & Rosner, G. (2007) Screening designs for drug development. *Biostatistics*, 8, 595–608.

Tannir, N.M., Msaouel, P., Ross, J.A., Devine, C.E., Chandramohan, A., Gonzalez, G.M.N. et al. (2020) Temsirolimus versus pazopanib (tempa) in patients with advanced clear-cell renal cell carcinoma and poor-risk features: a randomized phase II trial. *European Urology Oncology*, 3, 687–694.

Ventz, S., Cellamare, M., Parmigiani, G. & Trippa, L. (2018) Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics*, 19, 199–215.

Wathen, J.K. & Thall, P.F. (2017) A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials*, 14, 432–440.

Xu, Y., Trippa, L., Mueller, P. & Ji, Y. (2016) Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences*, 8, 159–180.

Yuan, Y., Guo, B., Munsell, M., Lu, K. & Jazaeri, A. (2016) MIDAS: a practical Bayesian design for platform trials with molecularly targeted agents. *Statistics in Medicine*, 35, 3892–3906.

Zeger, S.L. & Karim, M.R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

SUPPORTING INFORMATION

Web Appendices A–E, and Tables referenced in Section 1-Section 5 are available with this paper at the Biometrics website on Wiley Online Library. A computer program "Treatment-Screen-Subgroup" that reproduces the findings of this paper is available from the journal's website as supporting information.

How to cite this article: Lee, J., Thall, P.F. & Msaouel, P. (2023) Bayesian treatment screening and selection using subgroup-specific utilities of response and toxicity. *Biometrics*, 79, 2458–2473. https://doi.org/10.1111/biom.13738