The Canadian Journal of Statistics
Vol. 50, No. 2, 2022, Pages 417–436
La revue canadienne de statistique

Estimating design operating characteristics in Bayesian adaptive clinical trials

Shirin GOLCHI*

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada H3A 1G1

Key words and phrases: Bayesian test statistic; constrained design; COVID-19; ordinal-scale outcome; proportional-odds model; sampling distribution; trial simulation.

MSC 2020: 62C10; 62F03; 62F15; 62G07; 62L15; 62M30; 62P10.

Abstract: Bayesian adaptive designs have gained popularity in all phases of clinical trials with numerous new developments in the past few decades. During the COVID-19 pandemic, the need to establish evidence for the effectiveness of vaccines, therapeutic treatments, and policies that could resolve or control the crisis emphasized the advantages offered by efficient and flexible clinical trial designs. In many COVID-19 clinical trials, because of the high level of uncertainty, Bayesian adaptive designs were considered advantageous. Designing Bayesian adaptive trials, however, requires extensive simulation studies that are generally considered challenging, particularly in time-sensitive settings such as a pandemic. In this article, we propose a set of methods for efficient estimation and uncertainty quantification for design operating characteristics of Bayesian adaptive trials. Specifically, we model the sampling distribution of Bayesian probability statements that are commonly used as the basis of decision making. To showcase the implementation and performance of the proposed approach, we use a clinical trial design with an ordinal disease-progression scale endpoint that was popular among COVID-19 trials. However, the proposed methodology may be applied generally in the clinical trial context where design operating characteristics cannot be obtained analytically. *The Canadian Journal of Statistics* 50: 417–436; 2022 © 2022 Statistical Society of Canada

Résumé: Les plans adaptatifs bayésiens ont gagné en popularité dans toutes les phases d'essais cliniques grâce à d'importants développements réalisés au cours des dernières décennies. Pendant la pandémie COVID-19, la nécessité d'établir des preuves de l'efficacité des vaccins, des traitements thérapeutiques et des politiques susceptibles de résoudre ou de contrôler la crise a mis en évidence les avantages offerts par des plans d'essais cliniques efficaces et flexibles. En raison du niveau élevé d'incertitude présent dans de nombreux essais cliniques COVID-19, les plans adaptatifs bayésiens ont été considérés comme avantageux. Cela dit, la conception d'essais adaptatifs bayésiens nécessite de vastes études de simulation qui sont généralement considérées comme difficiles, en particulier dans des contextes sensibles au facteur temps comme lors d'une pandémie. Les auteurs de cet article proposent un ensemble de méthodes d'estimation efficace et de quantification de l'incertitude pour la conception d'essais adaptatifs bayésiens. En particulier, une modélisation de la distribution d'échantillonnage des énoncés de probabilité bayésienne est proposée. Cette dernière est couramment requise lors de la prise de décisions. Pour illustrer la mise en œuvre et la performance de l'approche proposée, les auteurs ont utilisé un plan d'essai clinique avec un critère d'évaluation ordinal de l'évolution de la maladie, plan relativement populaire dans les essais COVID-19. Aussi, la méthodologie proposée est assez générale pour être appliquée dans le contexte d'essais cliniques dont les caractéristiques opérationnelles du plan correspondant ne peuvent pas être obtenues de manière analytique. *La revue canadienne de statistique* 50: 417–436; 2022 © 2022 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author: shirin.golchi@mcgill.ca*

1. INTRODUCTION

Adaptive designs are one category of flexible alternatives to conventional fixed-size, randomized clinical trials (RCTs). In adaptive designs, decisions to stop or adapt may be made at interim analyses according to accumulating evidence, which, in some cases, can result in reduced sample sizes and cost. In addition, participants benefit from an increased chance of receiving a beneficial treatment. While adaptive designs in clinical trials go beyond Bayesian adaptive trials (Pallmann et al., 2018; Burnett et al., 2020), Bayesian methods in adaptive trials have become popular. The main reasons for this popularity are that the Bayesian framework naturally accommodates sequential analysis of accumulating data and that the validity of Bayesian probability statements is not affected by small sample sizes or by the incorporation of stopping rules (Berry et al., 2010).

There have been major developments in the design of Bayesian adaptive trials during the past few decades (Berry, 1989; Berry & Eick, 1995; Carlin, Kadane & Gelfand, 1998). Specifically, these designs have been widely applied to drug development (Müller et al., 2006; Thall, Cook & Estey, 2006) and disease-specific fields such as cancer trials (Buzdar et al., 2005; Zhou et al., 2008; Barker et al., 2009; Biswas et al., 2009).

Despite their growing use of Bayesian methods in clinical trials, proposed Bayesian adaptive designs are primarily assessed by regulatory agencies on the basis of their frequentist operating characteristics such as power and false-positive rate. The Food and Drug Administration (FDA), for example, emphasizes the importance of simulation studies in evaluating operating characteristics in adaptive trials of drugs and biologicals (FDA, 2019).

Various stopping rules may be incorporated into a Bayesian adaptive trial design with multiple interim analyses. Common decision rules include stopping the trial for efficacy or futility, or eliminating an arm if the probability of effectiveness is small for the corresponding treatment. Stopping rules are most commonly defined based on a Bayesian test statistic derived from the posterior distribution of model parameters. These include the posterior or posterior predictive probability of effectiveness, i.e., the alternative hypothesis. Efficacy or futility is then defined on the basis of high and low thresholds for these Bayesian probability statements. Another common adjustment is response adaptive randomization, where allocation ratios are adjusted with respect to posterior or posterior predictive probability statements.

Similar to the critical region of a frequentist test statistic, thresholds for Bayesian probabilities can be specified to achieve satisfactory design operating characteristics (DOCs). Power, for instance, may be defined as the probability of observing a high posterior probability of effectiveness given an assumed effect size. The sampling distribution of the posterior or posterior predictive probability statement (Bayesian test statistic) is therefore required to obtain power. Unlike in the classical hypothesis testing framework, however, the sampling distribution of a Bayesian probability statement is not available in closed form. Therefore, DOCs are typically estimated via Monte Carlo methods, i.e., by simulating the trial and the sampling distribution of the Bayesian test statistic for a given set of parameters.

Simulation studies for the design of Bayesian adaptive trials can be time consuming since the combination of plausible ranges of model parameters, including effect size and baseline measure assumptions, together with possible design parameters, including efficacy and futility thresholds, sample sizes, and frequency of interim analyses, can result in a large number of simulation scenarios. In many cases, a complex trial design and/or analysis model without analytically tractable posteriors requires a significant amount of computation—involving methods such as Markov chain Monte Carlo (MCMC) approaches—for a single trial simulation, which is multiplied by the number of simulation scenarios as well as by the number of iterations.

The methods proposed in the present article are motivated by a clinical trial design exercise in the context of the COVID-19 pandemic. A brief introduction to this clinical trial is provided in Section 2. In early discussions among the team of investigators, the ordinal-scale disease progression endpoint, as defined by the World Health Organization (WHO, 2020),

was considered as the primary endpoint of the trial with the goal of evaluating the effect of the intervention in reducing the severity of COVID-19. The proportional odds (PO) ordinal logistic regression was considered as the analysis model, which is one of the simplest statistical models used for the analysis of the ordinal-scale outcome (Harrell & Lindsell, 2020). Bayesian inference for the PO model, however, requires MCMC sampling since the analytic form of the posterior is not available. Therefore, assessing DOCs for Bayesian trials with the ordinal-scale outcome requires extensive simulations and carries a significant computational burden. The computational requirements, together with the uncertainty of the risk associated with the levels of the ordinal-scale outcome, resulted in a simplification of the design that used a binary endpoint. This shifted the focus to the risk of infection rather than disease severity, which was a necessary change to meet deadlines for funding applications and regulatory approvals dictated by the time-sensitive situation.

As a motivating example, we consider the hypothetical case with the ordinal-scale disease progression outcome and the PO model used in the design of the above COVID-19 trial. The link between COVID-19 infection and disease severity may have been strong prior to vaccination. In the absence of a strong association, however, changing the primary outcome for computational reasons would not have been acceptable and could have impacted the patient outcomes and the efficiency or appeal of the study. The present work will facilitate choosing the clinically correct endpoint by addressing computational feasibility.

In this article, we propose a set of methods to overcome the computational hurdles in evaluating DOCs as well as assessing the sensitivity of DOCs to the model and design parameters in Bayesian adaptive trials. The methodology can be employed, in general, for the design of clinical trials where power analysis and assessment of DOCs rely on simulations rather than analytic forms for the sampling distribution of the test statistic. The proposed approach is to estimate the sampling distribution of Bayesian probability statements (i.e., the test statistics) through the model parameter space, which is then used to provide computationally efficient estimates of DOCs for a wide range of assumptions and decision parameters without the need for additional simulations. A simple parametric density estimation approach is employed, where the sampling distribution of the probability statement is assumed to follow a beta distribution whose parameters are modelled as Gaussian processes (GPs) with a distance-based correlation structure across the parameter space. The GPs are trained over an initial set of simulated distributions for the test statistic at a select set of parameter values.

The novelty of the present article is in proposing a set of methodologies for the design of Bayesian adaptive trials beyond simple Monte Carlo simulations for the assessment of DOCs. Currently, methods for estimating the operating characteristics for a given set of parameters rely solely on simulations and are not accompanied by adequate uncertainty quantification. This article sets the foundation for the development of methods that facilitate the use of Bayesian measures for decision making in clinical trials while satisfying traditional requirements.

As mentioned above, the ordinal-scale outcome together with the PO model is used to showcase the implementation and performance of the proposed approach. One interesting but challenging aspect of using the ordinal-scale outcome is incorporating the uncertainty associated with the risks for disease severity levels in the trial design. This is addressed by exploring the DOCs over a range of plausible risk vectors. The vector of probabilities must add up to 1, which results in a simplex. Moreover, in GP-based models, satisfactory prediction performance throughout the input (parameter) space is achieved by spreading the initial set of target function evaluations uniformly using a space-filling design (O'Hagan, 1978; McKay, Beckman & Conover, 1979; Sacks et al., 1989; Johnson, Moore & Ylvisaker, 1990; Morris & Mitchell, 1995). Space-filling designs over nonrectangular, constrained spaces have recently received considerable attention (Lin, Sharpe & Winker, 2010; Lekivetz & Jones, 2014; Mak & Joseph, 2018). Generating a space-filling design on a simplex, however, is not straightforward.

We use the method of Lekivetz & Jones (2014), which generates a covering sample of the target space, clusters the sample points, and selects the cluster centroids as the design points. We propose a sampling technique for generating an initial sample on a simplex as an efficient and effective alternative to simple Monte Carlo.

The remainder of this article is organized as follows. The clinical trial design that motivates the development of the proposed approach is described in Section 2. In Section 3, we introduce a method for emulating and predicting the sampling distribution of a Bayesian test statistic using a beta prior distribution with GP parameters. We apply the proposed approach to the motivating example in Section 4. We also discuss the design of the training set for this application and assessment of the estimation using cross-validation. Section 5 follows with a discussion of the limitations of the proposed approach and future extensions.

2. MOTIVATING EXAMPLE

This work was motivated by the design of a clinical trial investigating the effectiveness of high-dose vitamin D in preventing severe cases of COVID-19 (prevention of COVID-19 with high-dose oral vitamin D supplemental therapy in essential healthcare teams–PROTECT). Given the goal of the study, the ordinal-scale disease progression outcome recommended by WHO (2020) was of interest. The clinical progression scale has 10 disease progression levels from uninfected to dead. The challenge in the use of a 10-level endpoint in clinical trials is the lack of granular data to inform the base rate assumptions as well as the dimensionality of the parameter space in statistical inference. Therefore, simplifications of the 10-level ordinal scale that merge the levels may be used. Table 1 shows the definition of the five-level ordinal-scale disease progression endpoint with the categories uninfected, mild disease, moderate disease, severe disease, and dead.

While the ordinal-scale endpoint was considered in early stages of the design of PROTECT, owing to the scale of simulation studies required to assess the DOCs and the level of uncertainty in the base risks associated with each outcome level, the focus was changed to the risk of infection and a binary primary outcome (incidence of laboratory-confirmed COVID-19 infection) that would accommodate closed-form posterior updates. The PROTECT study adopted a Bayesian adaptive design that allowed for an interim analysis when 75% of (approximately 1000) patients completed their follow-up time (16 weeks). The total sample size was estimated via simulations to achieve at least 80% probability of success. The trial could be concluded for efficacy at the

TABLE 1: Levels and description of the ordinal-scale disease severity endpoint.

Patient state	Descriptor	Level
Uninfected	Uninfected; no viral RNA detected	0
Ambulatory mild disease	Asymptomatic; viral RNA detected	1
	Symptomatic; independent	
	Symptomatic; assistance needed	
Hospitalized; moderate disease	Hospitalized; no oxygen therapy	2
	Hospitalized; oxygen by mask or nasal prongs	
Hospitalized; severe disease	Hospitalized; oxygen by NIV or high flow	3
	Intubation and mechanical ventilation	
Dead	Dead	4

interim analysis in case of sufficient evidence in favour of effectiveness. The control risk of infection was also monitored, and the data up to the interim analysis would be used to obtain the posterior predictive power at the final analysis. If the posterior predictive power was below the 80% target, the study would be prolonged for up to 24 weeks of follow-up in order to achieve a higher chance of success. The interim analysis and the adaptive components were essential because of the quickly evolving situation of the pandemic. The PROTECT study was funded by the CIHR COVID-19 Rapid Research Funding Opportunity approved by Health Canada and started recruiting in February 2021 (ClinicalTrials.gov Identifier: NCT04483635). However, the launch of the study coincided with the widespread distribution of vaccines to healthcare workers, which resulted in a significant drop in recruitment rates and, eventually, termination of the trial due to lack of feasibility.

In Section 4, we consider the hypothetical case where the ordinal-scale outcome is used as the primary outcome in the PROTECT design and argue that the proposed techniques in this article can improve the design of adaptive clinical trials by enabling the use of outcomes of interest with more realistic statistical models while accounting for the uncertainty associated with parameter assumptions in a computationally efficient framework.

We focus on the probability of stopping early for superiority at the interim analysis with 75% of participant outcomes as the DOCs of interest in order to illustrate the implementation and performance of the proposed approach. We emphasize that implementation and performance do not necessarily depend on the specific trial design or the specific DOCs of interest.

3. METHODOLOGY

Statistical success or significance in Bayesian adaptive clinical trials is commonly defined on the basis of posterior or posterior predictive probability of the alternative hypothesis (Berry et al., 2010). Stopping decisions are made according to the same criterion. For example, the decision of stopping the trial for efficacy at an interim analysis may be made if the posterior probability of effectiveness exceeds a prespecified upper probability threshold U (e.g., $U = 0.95$), that is, if

$$\pi(\mathbf{y}_t) = P(H_A|\mathbf{y}_t) > U, \tag{1}$$

where H_A is the alternative hypothesis, which is often formulated as the treatment having at least a certain magnitude of effect, and $\mathbf{y}_t$ denotes the participants' outcomes accumulated up to the decision time t. Therefore, $\pi(\mathbf{y}_t)$ can be viewed as a test statistic, and U is the critical value with respect to which statistical significance is determined. We use π as the general notation for a Bayesian test statistic used in a generic Bayesian adaptive trial.

Within the conventional RCT framework, the known sampling distribution of the test statistic under the null and alternative hypotheses are used to specify critical values that achieve target DOCs such as a 5% false-positive rate and 80% power. The sampling distribution of a posterior probability statement π, however, is not available in general. Therefore, DOCs are typically estimated using Monte Carlo simulations for a given set of model parameters that correspond to points under the null and alternative hypotheses as well as a range of critical values.

Evaluating π for every simulated set of $\mathbf{y}$ may require MCMC sampling except for simple models where a conjugate prior is available. While this is not generally a hurdle in Bayesian inference given the vast number of sampling algorithms, approximation methods, and computational advancements in the literature, involved Bayesian computation within a simulation study can be impractical. For example, estimating DOCs in a trial design with I (interim and final) analyses, n_T sets of model parameter values, and L decision thresholds via M simulation iterations and J MCMC iterations per evaluation of π has a computational complexity of $\mathcal{O}(n_T IMJL)$. In the following, we propose an approach that reduces this computational

complexity to $\mathcal{O}(n_t IMJ)$, where $n_t \ll n_T$ and the range of decision threshold values does not contribute to the computational complexity.

Denote the model parameter space, including the plausible range of all baseline and treatment effect parameters, by Θ. The null and alternative hypotheses define a partition of the parameter space, $\Theta = \Theta_0 \cup \Theta_A$ (where $H_0 : \theta \in \Theta_0$ and $H_A : \theta \in \Theta_A$). In simulation studies designed to assess the operating characteristics of a clinical trial design, a variety of parameter values within Θ that correspond to various parameter points under the null and alternative hypotheses need to be explored. In addition to power and the false-positive rate, a variety of other DOCs are of interest in adaptive designs. For example, power at a given interim analysis, i.e., the probability of stopping the trial early due to a correct efficacy conclusion, is

$$P_y(\pi_{\text{int}} > U | \theta^* \in \Theta_A), \tag{2}$$

where π_{int} is the Bayesian test statistic at the interim point and the subscript y denotes the probability under the data or sampling distribution. Therefore, the sampling distribution of π_{int} over the parameter space Θ is key to assessing DOCs.

We propose a model for the sampling distribution of π that allows us to "learn" the distribution function of π over Θ through the empirical sampling distribution obtained at a small set of parameter values. We assign the following prior distribution to π at a given point $\theta \in \Theta$:

$$\pi | \theta \sim \text{Beta}\,(a(\theta), b(\theta)) \tag{3}$$

with

$$a(\theta) \sim \mathcal{GP}(\mu_a, \rho_a(\theta, \theta')) \quad \text{and} \quad b(\theta) \sim \mathcal{GP}(\mu_b, \rho_b(\theta, \theta')), \tag{4}$$

where $\mathcal{GP}(\mu, \rho(\theta, \theta'))$ denotes a GP with a constant mean μ and a covariance function $\rho(\theta, \theta')$. The mean and covariance parameters of the GP prior distribution are trained independently for the shape and scale parameters of the beta distribution in (3).

The GP priors in (4) are based on the assumption that the parameters of the target distribution $f(\pi|\theta)$ are smooth in Θ. This, in turn, results in predictions for DOCs that are also smooth in Θ. We take the covariance functions of the GP processes in (4) to be the squared exponential covariance function, which is one of the most common choices in GP modelling (Rasmussen & Williams, 2006). This choice of covariance function assumes infinite differentiability with respect to θ. This may prove an unrealistic assumption in some applications and may give rise to convergence issues when estimating GP parameters. However, we do not encounter such issues in the application described in Section 4.

The GPs in (4) are trained over the parameter space using estimates of $a(\theta)$ and $b(\theta)$ at select $\theta \in \Theta$. Let $\boldsymbol{\theta}_t$ denote the training set of size n_t. The trial is simulated for M iterations according to the specified design for the parameter values in $\boldsymbol{\theta}_t$, which results in a Monte Carlo sample of the distribution of π. A beta distribution is fit over the Monte Carlo sample at each training point using the method of moments to obtain estimates of $a(\theta)$ and $b(\theta)$ that match the mean and variance of the empirical distribution. These estimates, $\hat{\mathbf{a}} = \hat{a}(\boldsymbol{\theta}_t)$ and $\hat{\mathbf{b}} = \hat{b}(\boldsymbol{\theta}_t)$, are then used as realizations to obtain the posterior GPs for any parameter values throughout Θ:

$$a(\theta^*)|\hat{\mathbf{a}} \sim \mathcal{N}\left(\mu_a^{\text{post}}, V(a(\theta^*))\right) \quad \text{and} \quad b(\theta^*)|\hat{\mathbf{b}} \sim \mathcal{N}\left(\mu_b^{\text{post}}, V(b(\theta^*))\right), \tag{5}$$

where

$$\mu_a^{\text{post}} = \mu_a + \mathbf{k}_a^\top \left(K_a + \sigma_a^2 I\right)^{-1} \mathbf{a}, \quad \mu_b^{\text{post}} = \mu_b + \mathbf{k}_b^\top \left(K_a + \sigma_b^2 I\right)^{-1} \mathbf{b},$$

$$V(a(\theta^*)) = \mathbf{k}_a(\theta^*, \theta^*) + \mathbf{k}_a^\top \left(K_a + \sigma_a^2 I\right)^{-1} \mathbf{k}_a,$$

and

$$V(b(\theta^*)) = \mathbf{k}_b(\theta^*, \theta^*) + \mathbf{k}_b^\top \left(K_b + \sigma_b^2 I\right)^{-1} \mathbf{k}_b.$$

The $\mathbf{k}$s are vectors of size n_t whose components are $\rho(\theta_t, \theta^*)$, the Ks are covariance matrices whose components are $\rho(\theta_t, \theta_{t'})$, and the σ^2s are the observation variances estimated for each GP model.

Once the posterior GPs are obtained, the sampling distribution of π is fully specified (predicted) at any point in Θ, and any DOCs of interest may be evaluated as a tail probability of the beta distribution. For example, the interim power in (2) is the $100U\%$ upper tail probability of a beta distribution whose parameters are given as in (5). The posterior uncertainty of these parameters will translate into uncertainty estimates for the corresponding tail probability, i.e., the DOCs estimates.

The specification of the training set $\boldsymbol{\theta}_t$ is important in the performance of the proposed model, as in any predictive-modelling framework. There exists a vast literature on GP modelling and the design of computer experiments regarding techniques for constructing training sets on a parameter (input) space. Specifically, space-filling designs are recommended to optimize predictive performance (O'Hagan, 1978; McKay, Beckman & Conover, 1979; Sacks et al., 1989; Johnson, Moore & Ylvisaker, 1990; Tang, 1993; Morris & Mitchell, 1995; Ye, 1998; Santner, Williams & Notz, 2003; Jin, Chen & Sudjianto, 2005; Joseph & Gul, 2015). For a comprehensive review, see Joseph (2016) and the references therein. In the next section, we apply the proposed approach to a specific DOC estimation example, discuss the construction of a space-filling design on a nonrectangular, nonconvex parameter space, and propose a simple design algorithm.

4. APPLICATION TO THE PROTECT DESIGN

In this section, we consider the design of the PROTECT study and a hypothetical scenario where the ordinal-scale disease progression endpoint is the primary outcome. The primary analysis is performed using a PO model inspired in Harrell & Lindsell (2020). This reference also provides a comprehensive discussion of the design of Bayesian adaptive trials with ordinal-scale outcomes.

For the DOC of interest, we will focus on the probability of stopping early for efficacy or futility at the interim analysis with 75% of participant outcomes. Without loss of generality, we consider a four-level outcome. We apply the model proposed in the previous section to estimate the data distribution of the posterior probability of effectiveness throughout the model parameter subspace, which corresponds to a set of credible base risks and effect assumptions under the null and alternative hypotheses. Then, we estimate the probability of stopping early for a number of decision thresholds.

An individual outcome is assumed to follow a multinomial distribution $Y \sim \text{Multinom}(1, \mathbf{p} = (p_1, \ldots, p_4))$, where p_k $(k = 1, \ldots, 4)$ is the risk associated with the kth level of disease severity, and $\sum_{k=1}^4 p_k = 1$. The PO model is a logistic regression of the tail probability

$$P(Y \geq k|A) = \frac{1}{1 + \exp[-(\alpha_k + \beta A)]} \quad \text{for } k = 2, 3, 4. \tag{6}$$

Note that $P(Y \geq 1|A) = 1$ and

$$\alpha_k = -\log \frac{\sum_{i=1}^{k-1} p_i}{1 - \sum_{i=1}^{k-1} p_i}.$$

In Equation (6), A is the treatment indicator, i.e., $A = 1$ indicates that the patient has received the treatment and $A = 0$ indicates assignment to the control arm. The ratio of the odds of disease

severity then represents the effect of the treatment:

$$\text{OR} = \frac{P(Y \geq k \,|\, A = 1)}{P(Y \geq k \,|\, A = 0)} = \exp(\beta).$$

This simple parametrization reduces the treatment effect to a single parameter. While this might not be a realistic modelling framework for ordinal outcomes in general, it is a welcome simplification in the absence of prior information on the mechanism of the effect with respect to different levels of disease severity. For an example of a more robust variation of the PO model, see Murray et al. (2018).

For the remainder of the article, we focus on $\mathbf{p}$ and OR rather than α_k and β as the model parameters since the formulation of the hypotheses and the baseline assumptions are made for these parameter transformations. Specifically, the null and alternative hypotheses are $H_0 : \text{OR} \geq 1$ and $H_A : \text{OR} < 1$.

The trial hypothesis may be defined according to a minimum clinically important effect, e.g., $\text{OR} < 0.9$. However, even a small effect was clinically important in the PROTECT study. The methods described in Section 3 apply to generic H_A.

For making Bayesian inference, however, prior distributions are specified for α_k and β, and the inference is based on the posterior distribution of these parameters given the observed data $\mathbf{y}$. This posterior distribution is analytically intractable and needs to be approximated or estimated by MCMC. We use the R package developed in James (2020), which uses the Hamiltonian Monte Carlo implemented in Stan to sample from the posterior distribution of the PO model parameters. The posterior distributions of $\mathbf{p}$ and OR are obtained as transformations of the posterior samples of α_k and β, respectively.

The posterior probability of effectiveness is then

$$\pi = P(H_A \,|\, \mathbf{y}) = P(\text{OR} < 1 \,|\, \mathbf{y}).$$

We focus on the probability of stopping the trial early for efficacy or futility at the interim analysis. Superiority and futility decisions are made if $\pi > U$ or $\pi < \ell$, where U and ℓ are the prespecified upper and lower probability thresholds, respectively. The DOCs of interest are therefore the probabilities of these events under the data (sampling) distribution for a variety of model parameter values under the null and alternative hypotheses. For example, the false-positive rate is $P_{\mathbf{y}}(\pi > U | \mathbf{p}_T = \mathbf{p} *, \text{OR}_T \geq 1)$, where $\mathbf{p}_T$ and OR_T are the assumed "true" values of the model parameters.

As explained in Section 3, the key to estimating probabilities of this form is to estimate the sampling distribution of π throughout the parameter space given an initial set of samples drawn from this sampling distribution via trial simulation at a training set. Denote the transformed parameter space for the PO model by $\Theta = \mathcal{P} \times O$, where $\mathcal{P}$ is the four-dimensional simplex of all plausible values of the vector of probabilities $\mathbf{p}$, and O is the interval of OR values that can be realistically assumed as the effect of the intervention.

Generating a space-filling design on Θ is challenging because $\mathcal{P}$ is a simplex defined by the constraints $\sum_{k=1}^{4} p_k = 1$ and $p_{lk} < p_k < p_{uk}$, where the p_{lk}s and p_{uk}s determine realistic ranges of values for the risk associated with each category. In Section 4.1, we describe a simple design algorithm that generates a space-filling design on $\mathcal{P}$. The design over $\mathcal{P}$ is then combined with a set of equally spaced OR values to provide training and test sets over the parameter space Θ.

4.1. Space-Filling Design Over the Simplex Parameter Subspace

Most space-filling design algorithms, such as those mentioned above, assume a rectangular input space. However, the problem of generating designs on nonrectangular spaces has recently

received more attention (Lin, Sharpe & Winker, 2010; Lekivetz & Jones, 2014; Mak & Joseph, 2018). We use the method of Lekivetz & Jones (2014), which generates and clusters a covering sample of the target space. The cluster centroids are used as the design points. This strategy ensures that no two points are too close to each other. Generating a uniform covering sample on a simplex, however, is not straightforward. We propose a sequential Monte Carlo sampling technique that generates an initial sample on the simplex as an efficient and effective alternative to simple Monte Carlo.

Consider the constrained space $\mathcal{P} \subseteq [0,1]^4$. Generating a uniform sample over $\mathcal{P}$ is equivalent to sampling from

$$\pi^{\top}(\mathbf{p}) = \frac{\mathcal{U}(\mathbf{p})\mathbf{1}_{\mathcal{P}}(\mathbf{p})}{\int_{\mathcal{P}} \mathcal{U}(\mathbf{p})d\mathbf{p}},$$

where $\mathcal{U}(\cdot)$ is the density function of a uniform distribution with the domain $[0,1]^4$ and $\mathbf{1}_{\mathcal{P}}(\mathbf{p})$ is an indicator function, which is equal to 1 if $\mathbf{p} \in \mathcal{P}$ and 0 otherwise.

We use the sequentially constrained Monte Carlo (SCMC) sampling approach proposed in Golchi & Campbell (2016) to sample from $\pi^{\top}(\mathbf{p})$. Specifically, we define the deviation of a given point $\mathbf{p}$ from the constraints that define the design region $\mathcal{P}$ as

$$C_{\mathcal{P}}(\mathbf{p}) = \left(\left| \sum_{k=1}^{4} p_k - 1 \right|, \left(p_k - p_{uk}\right)_{k=1}^{4}, \left(p_{lk} - p_k\right)_{k=1}^{4} \right).$$

For any point $\mathbf{p} \in \mathcal{P}$, the first component of the deviation vector is zero and the rest are negative. For any point $\mathbf{p} \notin \mathcal{P}$, $C_{\mathcal{P}}(\mathbf{p})$ measures the deviation from each of the constraints.

A probabilistic version of the constraint indicator $\mathbf{1}_{\mathcal{P}}(\mathbf{p})$ is $\prod \Phi(-\tau C_{\mathcal{P}}(\mathbf{p}))$, where Φ is the cumulative distribution function for the standard normal distribution and the parameter τ controls the slope of the probit function. The number of terms in the product is equal to the number of constraints defined by $C_{\mathcal{P}}(\mathbf{p})$, which is 9 in the present example. We have that

$$\lim_{\tau \to \infty} \prod \Phi(-\tau C_{\mathcal{P}}(\mathbf{p})) \propto \mathbf{1}_{\mathcal{P}}(\mathbf{p}).$$

From a uniform sample of size N on the unit hypercube $[0,1]^4$, the goal is to filter this sample towards the simplex $\mathcal{P}$ through a sequence of increasingly constrained densities. The SCMC sequence of densities is

$$\pi^{t}(\mathbf{p}) \propto \mathcal{U}(\mathbf{p}) \prod \Phi(-\tau_t C_{\mathcal{P}}(\mathbf{p})),$$

where $t = 0, \ldots, T$ and $0 = \tau_0 < \tau_1 < \cdots < \tau_T \to \infty$.

An effective sequence of the constraint parameters (τ_ts) can be achieved adaptively (Jasra, Stephens & Doucet, 2011). At each step, the next value in the sequence is determined such that the effective sample size (ESS) does not fall below a certain threshold, e.g., half of the sample size, $N/2$. This is done by numerically solving for τ_t in

$$\mathrm{ESS} = \frac{\left(\sum_{n=1}^{N} w_n^t(\tau_t)\right)^2}{\sum_{n=1}^{N} \left(w_n^t(\tau_t)\right)^2} = \frac{N}{2},$$

where

$$w_n^t(\tau_t) = \frac{\prod \Phi(-\tau_t C_{\mathcal{P}}(\mathbf{p}_n^{t-1}))}{\prod \Phi(-\tau_{t-1} C_{\mathcal{P}}(\mathbf{p}_n^{t-1}))}$$

and $\tau_T = 10^6$.

Figure 1 shows the two-dimensional marginal samples generated over $\mathcal{P}$ using the SCMC sampling scheme. The lower and upper limits for $\mathbf{p}$ are arbitrarily specified as $\mathbf{p}_l = (0.5, 0.05, 0.01, 0.005)$ and $\mathbf{p}_u = (0.9, 0.30, 0.05, 0.025)$, respectively. In practice, any information on credible values for the base risk of each disease severity level should be used to specify these thresholds. Note that $\mathbf{p}_l$ and $\mathbf{p}_u$ need not belong to $\mathcal{P}$, i.e., the limit vectors need not satisfy the constraint that $\sum_{k=1}^{4} p_k = 1$.

Once a covering sample over the target space is generated, we use the method proposed in Lekivetz & Jones (2014), which uses K-means clustering and selects the cluster centroids as the design points. This sampling-based approach aims to generate a design in which no two points are too close to each other without relying on a distance measure such as the Euclidean distance, which does not realistically represent distances in a constrained subspace. Various other methods have been developed in the experimental design literature that can be employed here (Welch et al., 1996; Draguljić, Santner & Dean, 2012; Joseph & Gul, 2015; Gomes, Claeys-Bruno & Sergent, 2018). Simplicity of implementation motivates our choice here.

4.2. Predicted Probability of Stopping Early

We consider the PROTECT study design with an interim sample size of 1000 and estimate the probabilities of stopping early at the interim analysis for efficacy and futility for a range of baseline risks $\mathbf{p}$ and OR. We can similarly estimate power (the probability of efficacy at the interim or final analysis) or any other DOCs for more complex designs with multiple interim analyses and adaptive features. Design complexities will need to be implemented only in initial simulations and do not have any implications on the prediction methods thereafter.

The goal is to predict the probabilities that

$$P(\text{OR} < 1|\mathbf{y}) > U \tag{7}$$

and

$$P(\text{OR} < 1|\mathbf{y}) < \ell \tag{8}$$

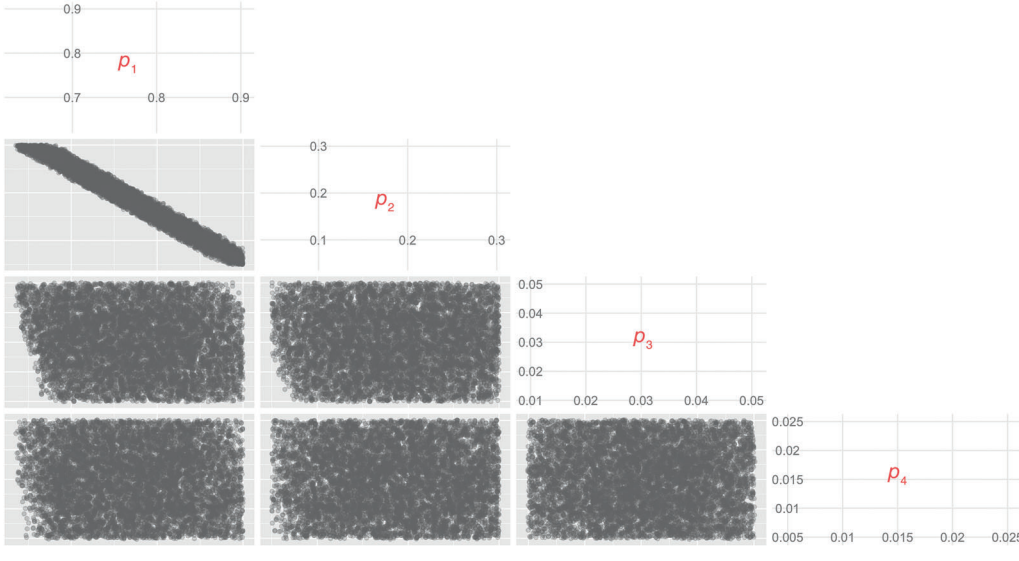

FIGURE 1: Two-dimensional marginals of the sample generated over the simplex $\mathcal{P}$.

over the input space $\Theta = \mathcal{P} \times O$, where

$$\mathcal{P} = \left\{ \mathbf{p} : \sum_{k=1}^{4} p_k = 1, 0.5 < p_1 < 0.9, 0.05 < p_2 < 0.3, 0.01 < p_3 < 0.05, \right.$$
$$\left. 0.005 < p_4 < 0.025 \right\}.$$

For the training set, a space-filling design of size 20 is generated over $\mathcal{P}$ using the methods described in Section 4.1. The Cartesian product of this set and a grid of size 4 over $O = (0.7, 1)$ is used as the final design of size 80 over Θ. Trial data are then simulated from the PO model described in Section 4 for each of the $(\mathbf{p}, \text{OR})$ pairs in the training set. For every simulated trial, the probability of superiority of the treatment (given in Equation (1)) is obtained to estimate the sampling distribution of π. A beta density function is fit to the Monte Carlo samples of π at every $\theta = (\mathbf{p}, \text{OR})$ to obtain estimates of $a(\theta)$ and $b(\theta)$. GPs are then fit to these simulated values.

The sampling distribution of π is then predicted over a test set of size 800 generated over Θ in the same fashion as the training set. Figure 2a,b shows the predicted probability of stopping early for efficacy (using a decision threshold of $U = 0.95$) derived from the distribution of π together with 95% credible intervals over a two-dimensional subspace of Θ, i.e., $p_1 \times \text{OR}$. The upper panel shows a slice of the subspace where the training set (denoted by square dots) is located, while the lower panel shows a slice of the subspace "in between" the training points. The 95% credible intervals represent two layers of uncertainty, i.e., the observation (Monte Carlo) error of the initial DOC evaluations and the uncertainty associated with the estimation/prediction step in obtaining the sampling distribution of the test statistic and the DOCs as its quantile.

As mentioned earlier, a strength of the proposed approach is that it obtains the sampling distribution of the test statistic rather than focusing on a single operating characteristic corresponding to a fixed decision criterion. This means that once the model is trained, one can readily explore a variety of decision thresholds U in (7) without any additional simulation runs.

To showcase this feature, Figure 3 shows the results in Figure 2a—excluding the training points—for the decision thresholds $U = 0.98$ and $U = 0.9$. These values correspond to more conservative and less conservative decision rules relative to $U = 0.95$, respectively. Using the more conservative decision criterion ($U = 0.98$) results in a smaller probability of stopping early for superiority for $\text{OR} < 1$ but controls the probability of a false-positive result at the interim analysis (Figure 3a, where $\text{OR} = 1$). The more permissive decision threshold ($U = 0.9$), however, increases the chance of stopping the trial early because of efficacy but increases the chance of a false-positive result to 12.5% (Figure 3b, where $\text{OR} = 1$).

Likewise, we may explore other operating characteristics such as the probability of concluding futility, according to (8) for various decision thresholds ℓ. Figure 4 shows the probabilities of concluding futility at the interim analysis for a restricted range of parameter values. This probability is negligible for larger effect sizes with the specified design settings.

The results in Figures 2–4 help in selecting a set of decision criteria for a given set of model parameters corresponding to baseline risks and effect assumptions. For example, if a base risk vector of $\mathbf{p} = (0.75, 0.22, 0.01, 0.02)$ and $\text{OR} = 0.7$ is assumed, then to allow a 65% (CI: 56%–75%) chance of stopping the trial early for superiority while controlling the interim false-positive rate at 2.3% (CI:1.3%–3.5%) requires a superiority decision threshold of $U = 0.98$. For the same set of assumptions for $\mathbf{p}$ but assuming no effect ($\text{OR} = 1$), a futility

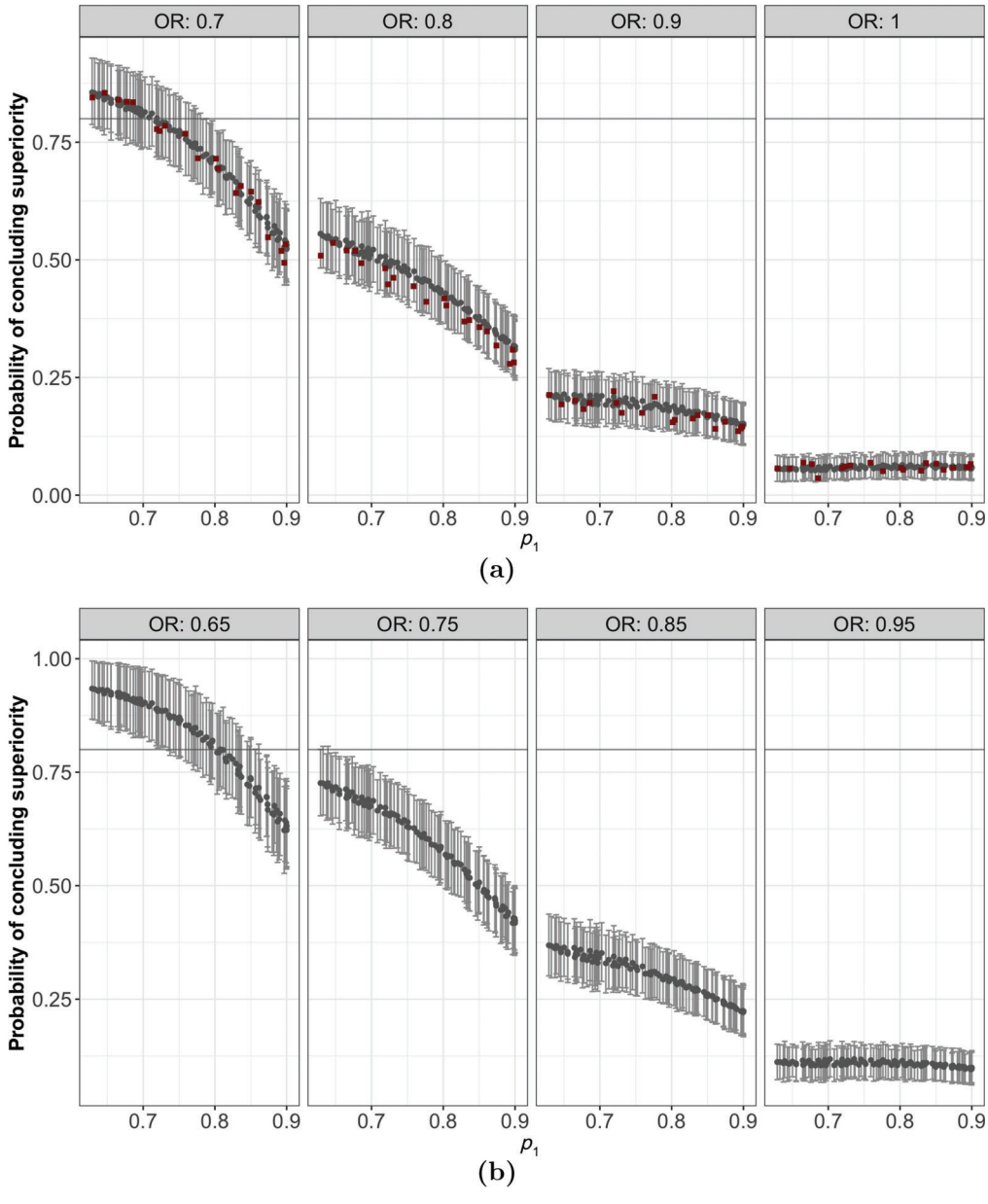

FIGURE 2: Point estimates for the interim probability of concluding superiority (grey round dots) with 95% credible intervals. Panel (a) shows a slice of the subspace where the training set (denoted by square dots) is located. Panel (b) shows a slice of the subspace "in between" the training points.

decision threshold of $\ell = 0.05$ allows about 5% (CI: 1.1%–16%) chance of stopping the trial early for futility at the interim analysis.

4.3. Computational Savings

We next present the computational savings in the application of the proposed approach in the previous subsection. All operations were performed in series. Clearly, further savings can be

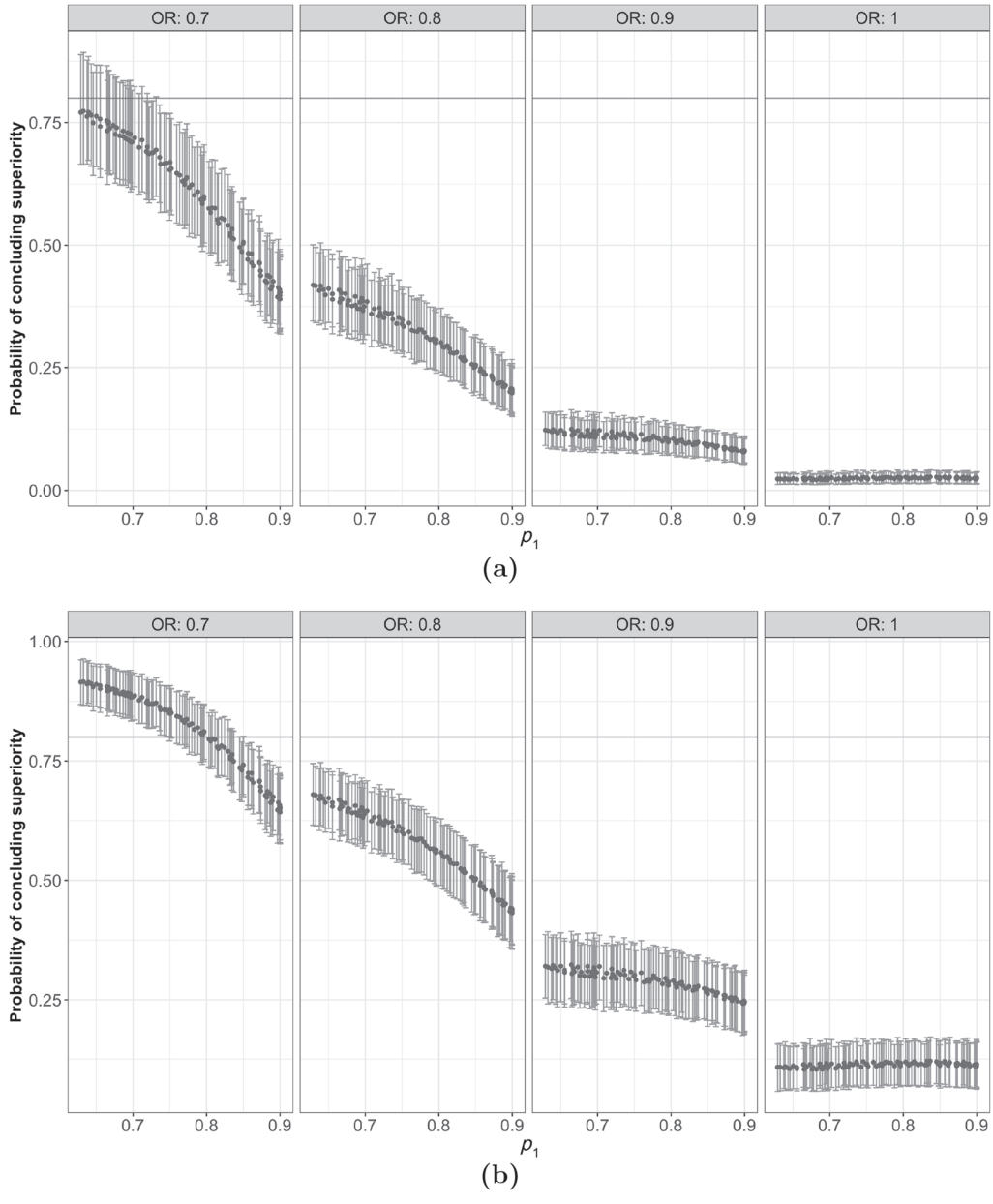

FIGURE 3: Point estimates and 95% credible intervals for the probability of concluding superiority with the decision thresholds (a) 0.98 and (b) 0.9.

attained through parallel computing. We argue, however, that significant computational resources are required to achieve the same level of savings through only parallel computing.

On a 1.4-GHz Quad-Core Intel Core i5 processor, one trial simulation with one chain and 1000 Hamiltonian Monte Carlo iterations took slightly over 2 s. Therefore, 1000 trial simulations for one set of model parameters would take about 30 min if run serially. For 80 points in the training set, this takes about 40 h, and for the 800 points in the test set 400 h. The proposed approach relies on simulation with 80 training points and therefore includes the 40 h

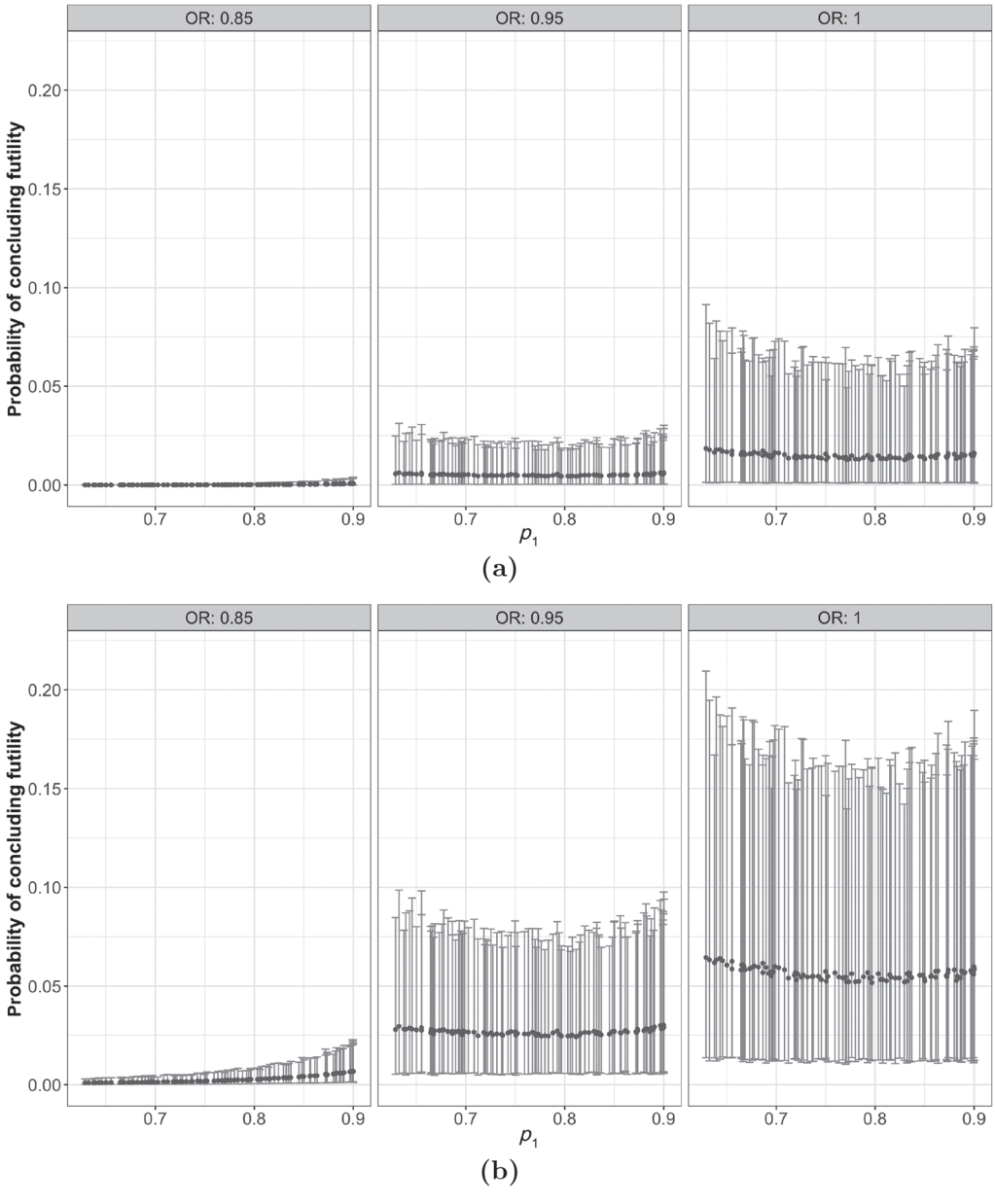

FIGURE 4: Point estimates and 95% credible intervals for the probability of concluding futility with the decision thresholds (a) 0.01 and (b) 0.05.

of computation time. However, it takes about 20 s to fit each GP model (two chains and 2000 iterations each), i.e., less than 1 min in total. Sampling from the posterior predictive distribution for the test set takes about 1 min. Therefore, the computation time for making predictions on the test set is reduced from 400 h to 2 min. To reduce the computation time via parallel computing alone to a comparable level would require significant computational resources (thousands of processing units). Of course, parallel computing together with the proposed methods can result in further reductions in computation times.

4.4. Assessment of Predictive Performance via Cross-validation

In this section, we provide an assessment of the accuracy and precision of the proposed approach in estimating DOCs relative to trial simulation. We use the training set θ_t (of size $n_t = 80$) used in the previous section to calculate the root mean squared error (RMSE) using leave-one-out cross-validation:

$$\text{RMSE} = \sqrt{\frac{1}{n_t K}\sum_{i=1}^{n_t}\sum_{k=1}^{K}(\phi_k(\theta_i) - \phi_t(\theta_i))^2}.$$

Here, $\phi_k(\theta_i) = P(\pi > 0.95 | a_k(\theta_i), b_k(\theta_i))$ is the estimate of the interim probability of superiority obtained as an upper tail probability of a beta distribution with the parameters given by the kth posterior samples $a_k(\theta_i)$ and $b_k(\theta_i)$ drawn from GP posteriors trained over the $n_t - 1$ points in θ_t with θ_i excluded, and $\phi_t(\theta_i)$ is the "true" interim power obtained via trial simulation at θ_i.

The RMSE, as defined above, is averaged both over the posterior distribution and the parameter space and is evaluated at 0.036. The estimation error—including both bias and variance—varies across the parameter space. Therefore, assessing pointwise accuracy and precision is important. Figure 5 shows the cross-validated bias and posterior standard error for the 80 points in the training set.

While estimation bias is small in magnitude overall (Figure 5a), there appears to be a systematically positive bias in the OR ≈ 0.8 region of the parameter space. This is the region where the sampling distribution changes and, as discussed in Section 5, the beta distribution with stationary GP prior distributions is not able to adequately capture this change. The same phenomenon contributes to the large posterior variance in these regions (Figure 5b). However, the estimation error remains small and the 95% credible intervals provide 100% coverage as indicated in Figure S.1 of the Supplementary Material.

For a more extensive assessment of the proposed approach, which takes into account the sensitivity of the predictions to the design of the training set, see the simulation study provided in the Supplementary Material. The simulation study is designed within the simple framework of a binary outcome with a beta–binomial model to be able to take advantage of the conjugate modelling framework and the consequent analytic posterior distribution. Because of the analytic posterior distribution, the DOCs can be estimated for a large number of points across the parameter space and for multiple training sets arising from the random training set design. This allows a thorough exploration of the accuracy and precision of DOCs estimates. The results of the simulation study are consistent with the conclusions drawn from the cross-validated performance measures above.

5. DISCUSSION

In this article, we proposed a set of methods for estimating the sampling distribution of a Bayesian probability statement—used for decision making in Bayesian adaptive trials—over a model parameter space. Our goal was to estimate a variety of DOCs and to assess their sensitivity to trial assumptions and design configurations in an efficient manner. We took advantage of the spatial correlation throughout the model parameter space to interpolate the parameters of the sampling distribution. We modelled the parameters of the sampling distribution as independent GPs trained over a set of estimates obtained by simulating the trial design over an initial set of model parameter values.

The main advantage of the proposed approach is that it enables exploration of a variety of operating characteristics as well as adequate uncertainty quantification. The methods presented in this article add efficiency to the overall process of collaborative trial design (possibly involving several iterations resulting from proposed changes and extensive explorations) in two ways:

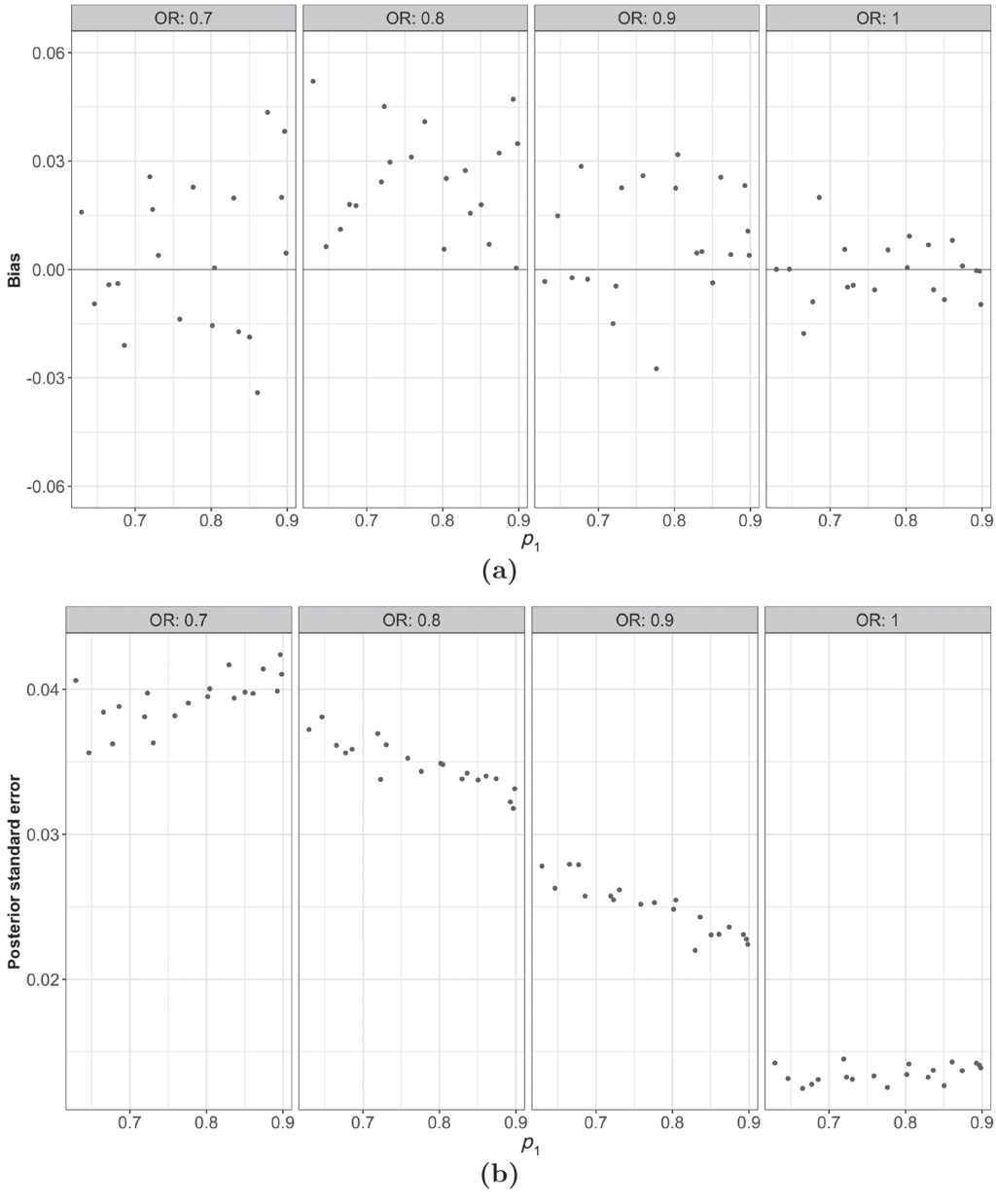

FIGURE 5: Cross-validated (a) bias and (b) posterior standard error in estimates of the probability of stopping early for 80 points across the parameter space.

first, by reducing the number of simulations required at each iteration of the process following a major change to the design and, second, by eliminating the need for additional simulations if the changes are only through model parameters or decision thresholds. This feature was showcased in Section 4, where the probabilities of stopping early for superiority or futility were estimated for a variety of decision thresholds with no additional simulation runs.

We applied the proposed methods to a hypothetical clinical trial design where an ordinal-scale disease progression endpoint was used with the PO model. The vector of base risks associated

 DOI: 10.1002/cjs.11699

with the levels of the ordinal outcome was defined as a simplex contained within the unit hypercube. This resulted in unique challenges in the design of the initial set of simulations used for training the GPs and in the exploration of DOCs estimates across the parameter space. We described a specialized design algorithm for this problem.

This article sets the foundation for the development of methods that facilitate the adoption of Bayesian measures for decision making in clinical trials. However, the proposed approach goes beyond trials that employ Bayesian decision rules and may be used generally where the sampling distribution of a test statistic is not available in an analytic form. For example, Barnett et al. (2021) recently proposed a novel statistical test in clinical trials with response adaptive randomization whose sampling distribution of the test statistic needs to be estimated by simulation: this approach can benefit from the methods proposed in this article. The proposed methodology is also applicable where the sampling distribution of the test statistic is available only asymptotically (which is the case for most frequentist tests) or only for large samples at interim analyses (Hadad et al., 2021).

We acknowledge that this work is a starting point and has limitations. There remains much room for further developments. We lay out some directions for future work below.

The parameters of the beta distribution must be positive. Therefore, specifying a GP prior that assigns nonzero probabilities to negative values is problematic. We addressed this issue using rejection sampling from the posterior predictive distribution. For the application in this article, the rejection rates remained zero or very small throughout the parameter space (see Section C in the Supplementary Material).

In cases where rejection sampling is inefficient, however, constraints need to be incorporated into the model. A variety of methods have been proposed for fitting GP models to observations from constrained functions (Riihimäki & Vehtari, 2010; Lin & Dunson, 2014; Golchi et al., 2015; Wang & Berger, 2016). As a common approach, the warped GP (Snelson, Rasmussen & Ghahramani, 2004) maps observations onto the real line via a monotonic function. All existing methods create a non-Gaussian process prior over the original function for which analytic predictions conditional on GP parameters cannot be obtained. Given the practical efficiency of a simple rejection-sampling scheme for the application in this article and the simplicity of unconstrained GP models, we did not find the additional complexities in incorporating positivity constraints to be justifiable. Doing so may be necessary in different modelling settings, however.

Another challenging feature of the problem is the quickly changing form of the sampling distribution of the test statistic in certain parts of the parameter space. The GP is not the most appropriate model for the parameters of this distribution, as this assumes stationarity throughout the input space. This can lead to increased bias in certain parts of the model parameter space. A variety of methods have been proposed to model nonstationary response surfaces (Schmidt & O'Hagan, 2003; Gramacy & Lee, 2008; Gramacy & Apley, 2013; Heinonen et al., 2016). In most cases, however, the solution comes at the cost of an analytic form for the predictions or uncertainty estimates and an increased computational burden. In the examples explored in the present article, the resulting bias was small enough, and so we consider the fit of a conventional GP to be satisfactory.

Finally, the parametric form assumed for the Bayesian test statistic, e.g., the beta distribution alone, might not capture the true sampling distribution throughout the model parameter space. A nonparametric modelling approach is therefore a potential future direction for this work.

CODE

The code for the simulation study and the implementation of the methods for the ordinal-scale endpoint and the PO model used in this article are provided in a public repository on GitHub, at https://github.com/sgolchi/DOCsest.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the two principal investigators of the PROTECT study, Dr. Francine M. Ducharme and Dr. Cecile Tremblay, the review committee, and Dr. Alexandra Schmidt for their invaluable comments, which resulted in significant improvements to this article. This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., & Esserman, L. (2009). I-SPY2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology and Therapeutics*, 86, 97–100.

Barnett, H., Villar, S., Geys, H., & Jaki, T. (2021). A novel statistical test for treatment differences in clinical trials using a response-adaptive forward-looking Gittins Index Rule. *Biometrics*, 1–12.

Berry, D. (1989). Monitoring accumulating data in a clinical trial. *Biometrics*, 45, 1197–1211.

Berry, D. & Eick, S. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, 14, 231–246.

Berry, S. M., Carlin, B. P., Lee, J. J., & Müller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*, CRC Press, New York, NY.

Biswas, S., Liu, D. D., Lee, J. J., & Berry, D. A. (2009). Bayesian clinical trials at the University of Texas M.D. Anderson Cancer Center. *Clinical Trials*, 6, 205–216.

Burnett, T., Mozgunov, P., Pallmann, P., Villar, S. S., Wheeler, G. M., & Jaki, T. (2020). Adding flexibility to clinical trial designs: An example-based guide to the practical use of adaptive designs. *BMC Medical Research Methodology*, 18(1), 352.

Buzdar, A. U., Ibrahim, N. K., Francis, D., Booser, D. J., Thomas, E. S., Rivera, E., Theriault, R. L. et al. (2005). Significantly higher pathological complete remission rate following neoadjuvant therapy with trastuzumab, paclitaxel and epirubicin-containing chemotherapy: Results of a randomized trial in HER-2- positive operable breast cancer. *Journal of Clinical Oncology*, 23, 3676–3685.

Carlin, B., Kadane, J., & Gelfand, A. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, 54, 964–975.

Draguljić, D., Santner, T. J., & Dean, A. M. (2012). Non-collapsing space-filling designs for bounded non-rectangular regions. *Technometrics*, 54, 169–178.

FDA (2019). *Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry*, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics

Golchi, S., Bingham, D. R., Chipman, H., & Campbell, D. A. (2015). Monotone emulation of computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1), 370–392.

Golchi, S. & Campbell, D. A. (2016). Sequentially constrained Monte Carlo. *Computational Statistics and Data Analysis*, 97, 98–113.

Gomes, C., Claeys-Bruno, M., & Sergent, M. (2018). Space-filling designs for mixtures. *Chemometrics and Intelligent Laboratory Systems*, 174, 111–127.

Gramacy, R. B. & Apley, D. W. (2013). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2), 561–578.

Gramacy, R. B. & Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483), 1119–1130.

Hadad, V. O. P., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2125–2135.

Harrell, F. and Lindsell, C. (2020). Statistical design and analysis plan for sequential parallel-group RCT for COVID-19. https://hbiostat.org/proj/covid19/bayesplan.html

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., & Lähdesmäki, H. (2016). Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*, Cadiz, Spain.

James, N. (2020). *Bayesian Cumulative Probability Models—bayesCPM*, GitHub repository. https://github.com/ntjames/bayesCPM/tree/master/pkg

Jasra, A., Stephens, D. A., & Doucet, A. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38, 1–22.

Jin, R., Chen, W., & Sudjianto, A. (2005). An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134, 268–287.

Johnson, M. E., Moore, L. M., & Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26, 131–148.

Joseph, V. R. (2016). Space-filling designs for computer experiments: A review. *Quality Engineering*, 28, 28–35.

Joseph, V. R. & Gul, E. (2015). Maximum projection designs for computer experiments. *Biometrika*, 102(2), 371–380.

Lekivetz, R. & Jones, B. (2014). Fast flexible space-filling designs for nonrectangular regions. *Quality and Reliability Engineering*, 31, 829–837.

Lin, D. K. J., Sharpe, C., & Winker, P. (2010). Optimized U-type designs on flexible regions. *Computational Statistics and Data Analysis*, 54(6), 1505–1515.

Lin, L. & Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101, 303–317.

Mak, S. & Joseph, V. R. (2018). Minimax and minimax projection designs using clustering. *Journal of Computational and Graphical Statistics*, 27(1), 166–178.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239–245.

Morris, M. D. & Mitchell, T. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43, 381–402.

Müller, P., Berry, D., Grieve, A., & Krams, M. (2006). A Bayesian decision-theoretic dose finding trial. *Decision Analysis*, 3, 197–207.

Murray, T. A., Yuan, Y., Thall, P. F., Elizondo, J. H., & Hofstetter, W. L. (2018). A utility-based design for randomized comparative trials with ordinal outcomes and prognostic subgroups. *Biometrics*, 74, 1095–1103.

O'Hagan, A. (1978). Curve fitting and optimal design for predictions. *Journal of the Royal Statistical Society: Series B*, 40, 1–42.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J. et al. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medical Research Methodology*, 16(1), 29.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, UK.

Riihimäki, J. & Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 645–652.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments (with discussion). *Statistical Science*, 4, 409–435.

Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*, Springer, New York, NY.

Schmidt, A. & O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society: Series B*, 65, 743–758.

Snelson, E., Rasmussen, C. E., & Ghahramani, Z. (2004). Warped Gaussian processes. In Thrun, S., Saul, L. K., Schølkopf, B. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 16, MIT Press, Cambridge, MA, USA.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88, 1392–1397.

Thall, P., Cook, J., & Estey, E. (2006). Adaptive dose selection using efficacy-toxicity trade-offs: Illustrations and practical considerations. *Journal of Biopharmaceutical Statistics*, 16, 623–638.

Wang, X. & Berger, J. O. (2016). Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4, 1–25.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Morris, M. D., & Schonlau, M. (1996). Response to James M. Lucas. *Technometrics*, 38, 199–203.

WHO. (2020). A minimal common outcome measure set for COVID-19 clinical research. *The Lancet*, 20(8), E192–E197.

Ye, K. Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *Journal of the American Statistical Association*, 93, 1430–1439.

Zhou, X., Liu, S., Kim, E., Herbst, R., & Lee, J. (2008). Bayesian adaptive design for targeted therapy development in lung cancer—A step toward personalized medicine. *Clinical Trials*, 5, 181–193.

Received 5 May 2021
Accepted 4 January 2022