Sensitivity Analyses of Clinical Trial Designs: Selecting Scenarios and Summarizing Operating Characteristics

Larry Han¹, Andrea Arfèº, and Lorenzo Trippa¹c

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; ºDepartment of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY; cDepartment of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA

ABSTRACT

The use of simulation-based sensitivity analyses is fundamental for evaluating and comparing candidate designs of future clinical trials. In this context, sensitivity analyses are especially useful to assess the dependence of important design operating characteristics with respect to various unknown parameters. Typical examples of operating characteristics include the likelihood of detecting treatment effects and the average study duration, which depend on parameters that are unknown until after the onset of the clinical study, such as the distributions of the primary outcomes and patient profiles. Two crucial components of sensitivity analyses are (i) the choice of a set of plausible simulation scenarios and (ii) the list of operating characteristics of interest. We propose a new approach for choosing the set of scenarios to be included in a sensitivity analysis. We maximize a utility criterion that formalizes whether a specific set of sensitivity scenarios is adequate to summarize how the operating characteristics of the trial design vary across plausible values of the unknown parameters. Then, we use optimization techniques to select the best set of simulation scenarios (according to the criteria specified by the investigator) to exemplify the operating characteristics of the trial design. We illustrate our proposal in three trial designs.

1. Introduction

Clinical trial designs are becoming increasingly complex to meet the multifaceted needs and goals of precision medicine. Examples of complex designs include adaptive seamless phase i/ii designs for evaluating, early in the treatment development process, the dosing, safety, and activity of new drugs (Hobbs et al. 2019). Also, adaptive randomized trials with frequent interim looks at the data can evaluate one or more therapies simultaneously while attempting to minimize trial duration and resources (Berry et al. 2010; Thorlund et al. 2018). Additional examples of complex designs have been implemented in biomarker-stratified trials to evaluate the efficacy of a therapy and possible variations of treatment effects across patient subgroups (Mehta, Liu, and Theuer 2019).

When planning a new trial, it is necessary to predict and evaluate several operating characteristics. Relevant operating characteristics can include the likelihood of selecting an effective dose with low toxicity in a phase i/ii study, the probability of detecting treatment effects in a randomized study, the expected trial duration, costs, and other metrics to evaluate designs that often enroll patients from different subgroups. Multiple operating characteristics typically need to be examined jointly in order to evaluate the relevant tradeoffs achieved by candidate designs, such as balancing the accuracy in estimating treatment effects and the expected study duration.

The obvious challenge for evaluating a candidate design is that the vector of operating characteristics of the study design is not known and it is difficult to estimate before the onset of the trial. Indeed, the operating characteristics are usually a function of a vector of unknown parameters that identify the distribution of all relevant variables to be captured during the trial. For example, unknown parameters can include the enrollment and drop-out rates, the magnitude of treatment effects, and the prevalence of predictive biomarkers in the trial population. Uncertainty on these parameters makes it nontrivial to evaluate whether a candidate design is appropriate for implementing the new study.

Sensitivity analyses are commonly used to account for uncertainty on unknown parameters and operating characteristics when evaluating a candidate design. They typically proceed in three steps. First, a set of plausible scenarios, that is, specific values of the vector of unknown parameters, is selected. Next, the corresponding operating characteristics are computed using trial simulations or analytic results. Finally, based on the computed operating characteristics and their variations across the set of scenarios, the investigators evaluate whether the candidate design is appropriate to achieve the aims of the study. Throughout the manuscript, we use the terms sensitivity analysis or simulation report to indicate a set of scenarios and the associated operating characteristics which are computed to illustrate...
how the operating characteristics vary across plausible values of unknown parameters.

Producing a simulation report to effectively evaluate a study design has been recommended as one of the key supporting documents for interacting with the FDA (Mayer et al. 2019; U.S. FDA 2020). However, it can be difficult to select the set of unknown parameters, especially if the dimension of the vector of unknown parameters is moderate to high (say $\geq 5$). For the investigators, it might be unclear if the selected scenarios are adequate to illustrate the variations of the operating characteristics across potential values of the unknown parameters. Similarly, for regulators, there may be skepticism as to whether the selected scenarios are chosen to highlight positive aspects of the trial design without pointing at its limitations and negative aspects (Razavi et al. 2021). Another subtle challenge is the choice of the number of scenarios. Indeed, a large number of scenarios (say 100) may simplify the task of representing how the operating characteristics vary across potential values of the unknown parameters, but a simulation report that contains too many scenarios makes it difficult to interpret and communicate the included results.

We propose a method to choose an optimal set of scenarios for a simulation report that will provide relevant operating characteristics. This decision is based on a utility criterion, which formalizes the ability of any set of scenarios to represent the map between the unknown parameters and the operating characteristics. In some cases, we will consider a restriction of the parameter space to focus only on regions of plausible values of the unknown parameters. The utility criterion assigns high (low) utility to a set of scenarios if the table of potential unknown parameters and operating characteristics is an accurate (inaccurate) summary of how the design’s operating characteristics vary across the considered parameter space. We call the set of scenarios that maximizes the utility criterion the Representative and Optimal Sensitivity Analysis (ROSA) scenarios. To select the ROSA scenarios, we introduce a computational procedure that leverages (i) flexible regression methods like neural networks (NNs) (Goodfellow, Bengio, and Courville 2016) and (ii) optimization algorithms like simulated annealing (Bélisle 1992). Our approach is applicable to any trial design, regardless of the number of unknown parameters and the number of operating characteristics.

In summary, we propose ROSA as a computational tool that allows one to examine any clinical trial design by selecting a parsimonious set of simulation scenarios with the goal of representing the variations of the operating characteristics across plausible values of unknown parameters. To illustrate this approach, we conduct sensitivity analyses for three trial designs. The first is a two-arm randomized design that aims to test and estimate the effects of an experimental treatment compared to the standard of care (SOC). The second is a multi-stage randomized trial that leverages an auxiliary/surrogate outcome $S$ measured shortly after randomization for interim decisions and a primary outcome $Y$ with a longer ascertainment time (Niewczas, Kunz, and König 2019). The third is a biomarker-adaptive enrichment design similar to the design of the TAPPAS trial (Mehta, Liu, and Theuer 2019), a randomized phase III trial comparing TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (Jenkins, Stone, and Jennison 2011; Jones et al. 2017). In the first design, we consider a single unknown parameter and a single operating characteristic, whereas for the latter two designs we consider multiple unknown parameters and multiple operating characteristics.

2. Selecting Sensitivity Scenarios

2.1. Notation and Problem Set-Up

We introduce our procedure to select $K$ sensitivity scenarios $\theta_1, \ldots, \theta_K \in \Theta$, where $\Theta$ is the set of potential values of the unknown parameters $\theta$. We assume that $\Theta$ is a bounded subset of $\mathbb{R}^d$ and use the notation $|| \cdot ||_2$ to indicate the Euclidean norm on $\mathbb{R}^d$. We will restrict $\Theta$ to a subset $\Theta'$ when there is sufficient prior information from completed studies or clinical experience. We identify ROSA scenarios $\theta^*_1, \ldots, \theta^*_K$ as the scenarios that maximize a utility criterion $U$

$$\theta^*_1, \ldots, \theta^*_K = \arg\max_{\theta_1, \ldots, \theta_K} U(\theta_1, \ldots, \theta_K),$$

where

$$U(\theta_1, \ldots, \theta_K) = -\max_{\theta' \in \Theta} \left\{ \min_{k=1, \ldots, K} D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)] \right\}. \quad (2)$$

We can symmetrically define the corresponding loss function $L = -U$ by inverting the sign in equation (2). Here, $D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ is a metric between the operating characteristics $\mathbf{f}(\theta') = (f_1(\theta'), \ldots, f_K(\theta'))$ and $\mathbf{f}(\theta_k) = (f_1(\theta_k), \ldots, f_K(\theta_k))$. We will consider metrics of the form

$$D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)] = \sum_{r=1}^{R} w_r || f_r(\theta') - f_r(\theta_k) ||_2,$$

where $w_1, \ldots, w_R$ are nonnegative weights that sum to one. The weights can be user-specified to calibrate the relative importance of different operating characteristics. Setting the weights to $1/R$ results in equal weighting for each operating characteristic.

We can now provide an explicit interpretation of the utility function $U$ in (2). Consider a set of scenarios $\{\theta_1, \ldots, \theta_K\}$ – the order of the entries is not relevant—and an arbitrary scenario $\theta'$ in $\Theta$. For $1 \leq k \leq K$, the metric $D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ is a summary of the differences between the operating characteristics at $\theta'$ and the same operating characteristics when we consider the $k$th scenario $\theta_k$. Therefore, $\min_{k=1,\ldots,K} D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ can be viewed as an approximation error between $\mathbf{f}(\theta')$ and a similar vector of operating characteristics selected among our $K$ options $\mathbf{f}(\theta_1), \ldots, \mathbf{f}(\theta_K)$. Expression (2) identifies through the maximization operator the worst-case (with highest approximation error) that we can obtain by varying $\theta'$ in $\Theta$. We maximize the utility function $U$ and use $\theta^*_1, \ldots, \theta^*_K$ to indicate the ROSA scenarios. Alternative utility criteria and loss functions are described later in the manuscript. A table of notation used throughout the article is provided in Table 1.
2.2. An Example with a Geometric Interpretation

To provide a geometric interpretation of the utility criterion $U$, we illustrate how one set of $K$ scenarios can be preferable to a different set of $K$ scenarios (Figure 1). Specifically, suppose we aim to design a single-arm trial with an interim analysis that allows for early-stopping for futility. The goal of the trial is to compare the response rate of an experimental drug $\theta_1$ with that of the SOC $\theta_0$ at the end of the study. However, because study patients only receive the experimental drug, the response...
rate under the SOC $\theta_0$ is estimated $\hat{(\theta_0)}$ before the onset of the study, for example using data from a previous trial. At the interim analysis, the trial may stop for futility if the preliminary evidence of positive treatment effects $\Delta_{interim}$ is insufficient to continue the study. During the final analysis, the null hypothesis $H_0 : \theta_1 \leq \theta_0$ (the experimental therapy is not superior to the historical control) is tested against the alternative hypothesis $H_1 : \theta_1 > \theta_0$ (the experimental therapy is superior to the historical control). In this design, $\theta = (\theta_0, \theta_1)$ are the unknown parameters, and $\Theta = [0, 1]^2$. Suppose that there are two operating characteristics of interest: (i) $f_1$, the probability of a positive result ($H_0$ is rejected) and (ii) $f_2$, the expected sample size. 

The left panel of Figure 1 is a representation of $\Theta$. We are interested in the two operating characteristics of the single-arm design. Two sets of $K = 6$ scenarios are proposed. The first set of scenarios $\theta_1^1, \ldots, \theta_1^6$ (blue points) is chosen by varying both unknown parameters at the same time, while the second set $\theta_2^1, \ldots, \theta_2^6$ (red points) is chosen by varying only $\theta_0$ while fixing the value of $\theta_1$. The two sets of scenarios, the corresponding operating characteristics, and associated loss $L = -U$ are represented in the right panel of Figure 1. The first set of scenarios (blue points) is preferred over the second set (red points) because it is more representative of the variation of the operating characteristics over $\Theta$. Geometrically, the loss $L(\theta_1^1, \ldots, \theta_2^6)$ associated with the blue points is identical to the minimum radius of the circles with centers $\hat{f}(\theta_1^1), \ldots, \hat{f}(\theta_1^6)$ (see Figure 1) necessary to cover the operating characteristics surface $f(\Theta)$.

### 2.3. Estimating the Operating Characteristics

We describe an algorithm to numerically approximate the operating characteristics $f(\theta)$ for every $\theta \in \Theta$. This is necessary to solve the optimization problem in (2). Indeed, in most cases the function $f(\theta)$ cannot be computed in closed form.

We briefly outline our four-step procedure. In the first step, we choose a large number $J$ (say $J = 1000$) of training scenarios $\theta_1^1, \ldots, \theta_1^J$. In the second step, we use Monte Carlo simulations to obtain estimates $\hat{f}(\theta_1^j), \ldots, \hat{f}(\theta_1^j)$ of $f(\theta_1^j), \ldots, f(\theta_1^j)$. In the third step, we train a flexible regression model—we use NNs in our implementation—based on the data points $(\theta_1^j, \hat{f}(\theta_1^j)), \ldots, (\theta_1^j, \hat{f}(\theta_1^j))$. The output of this step is a regression function $\hat{f}(\theta)$ that is easy to compute at any $\theta \in \Theta$ and that approximates $f(\theta)$. In the fourth step, we validate the regression model based on $J'$ (say $J' = 200$) independent simulations $(\theta_1^j, \hat{f}(\theta_1^j)), \ldots, (\theta_1^j, \hat{f}(\theta_1^j))$. Steps 1-3 of this procedure are summarized in Algorithm 1. Steps 4 is described in Algorithm 2.

In more detail, in step 1, to select the training scenarios $\theta_1^1, \ldots, \theta_1^J$, we randomly select $J$ scenarios in $\Theta$ using Latin hypercube sampling (LHS) (McKay, Beckman, and Conover 2000). LHS generates $J$ scenarios by first partitioning the $d$ unknown parameter dimensions into $J$ nonoverlapping intervals and selecting one value from each interval at random. The $J$ values obtained for the first unknown parameter $\theta_1$ are randomly paired with the $J$ values obtained for the second $\theta_2$, and so on, for all $d$ unknown parameters to form $J$-tuples, which constitute the training scenarios $\theta_1^1, \ldots, \theta_1^J$.

In step 2, we estimate the operating characteristics of the trial design. In the article, we focus on operating characteristics that can be defined as expected values, which are often of great interest, for example, bias, power, duration of the trial, etc. These operating characteristics can be written down as

$$f(\theta) = \mathbb{E}_\theta[\phi(Z, \theta)]$$

for some function $\phi$, where the random vector $Z$ represents the data generated during the trial—including the collection of treatment assignment indicators and realized patient outcomes—under scenario $\theta$. In practice, to estimate $f(\theta)$, we proceed as follows. First, for each of the training scenarios $\theta_1^i$, $1 \leq j \leq J$, we simulate $M$ (say $M = 200$) clinical trials following the trial design. Then, we use the $M$ scenario-specific simulated trials to compute the estimate

$$\tilde{f}(\theta_j^i) = M^{-1} \sum_{m=1}^M \phi(Z_{j,m}, \theta_j^i), \quad 1 \leq j \leq J,$$

where $Z_{j,m}$ is the $m$th trial dataset simulated under the $j$th training scenario $\theta_j^i$. Throughout the manuscript $\phi(Z, \theta)$ will take values in a compact set. For example, $\phi$ can be the indicator that captures if a null hypothesis of interest has been rejected at the end of the study, or the duration of the simulated trial. One possibility for handling unbounded operating characteristics (e.g., median squared error) is to apply monotone transformations, from the real line to the unit interval, that rescale the operating characteristics. In this case, the selection of representative scenarios would be influenced by the specific monotone map used to express the operating characteristic.

In step 3, we have only two inputs, the scenarios $\theta_1^i$ and the estimates $\tilde{f}(\theta_j^i), 1 \leq j \leq J$, to fit a function $\hat{f}(\theta)$. For example, one could use NNs, splines (Bookstein 1989), or Gaussian processes (Rasmussen 2003). We use NN regression functions in our applications because these are easy to compute using widely available software and have been demonstrated to have good...


Algorithm 2: Validating the approximation of the operating characteristics \( \hat{f} \)

1. **Input**: Approximation of the operating characteristics \( \hat{f} \).
2. Trial design, \( j' \).
3. Randomly select \( j' \) scenarios \( \theta_{1}^{j'}, \ldots, \theta_{j'}^{j'} \in \Theta \) independently from previous computations (Algorithm 1).
4. for \( j' = 1 \) to \( j' \) do
   5. Simulate \( M' \) trials \( Z_{j',m'} \).
   6. Compute \( \hat{f}(\theta_{j'}^{j'}) = M'^{-1} \sum_{m'=1}^{M'} \varphi(Z_{j',m'}, \theta_{j'}^{j'}) \).
   7. Compute \( \hat{f}(\theta_{j'}^{j'}) \).
8. **Output**: Set of differences \( \hat{f}(\theta_{j'}^{j'}) - \hat{f}(\theta_{j'}^{j'}) \) and scatterplots to jointly visualize the operating characteristic estimates \( \hat{f}(\theta_{j'}^{j'}) \) and the independent estimates \( \hat{f}(\theta_{j'}^{j'}) \), \( 1 \leq j' \leq J' \). Compute summaries of the differences (e.g., median, range, or other descriptive statistics).
9. **Interpretation**: Differences between \( \hat{f}(\theta_{j'}^{j'}) \) and the independent estimates \( \hat{f}(\theta_{j'}^{j'}) \), \( 1 \leq j' \leq J' \), consistently close to zero provide evidence that \( \hat{f} \) is an accurate approximation of \( f \).

To summarize, we can approximate the loss function \( \mathcal{L}(\theta_{1}, \ldots, \theta_{K}) \) over the entire parameter space \( \Theta \) by \( \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \) using a diffuse and finite subset \( \Theta^{F} \).

2.5. Optimization by Simulated Annealing

We now aim to approximately minimize the loss function \( \hat{\mathcal{L}} \). To illustrate the need for approximate solutions, consider the setting of a single unknown parameter \( d = 1 \), a finite \( \Theta \), and an easy-to-compute loss function \( \mathcal{L} \). Even in this simple setting, identifying \( \theta_{1}, \ldots, \theta_{K} \in \Theta \) can be challenging. For example, to select \( K = 10 \) representative scenarios \( \theta_{1}, \ldots, \theta_{K} \) from 1000 points \( \{\theta_{j}; 1 \leq j \leq 1000\} = \Theta \), the loss function \( \hat{\mathcal{L}} \) would need to be calculated for \( 2.63 \times 10^{13} \) different possible sets \( \{\theta_{1}, \ldots, \theta_{K}\} \). In what follows, we describe the use of simulated annealing (Algorithm 3), a simple strategy to reduce the outlined computational burden, regardless if \( \Theta \) is finite or not. Kirkpatrick, Gelatt, and Vecchi (1983), Bélisle (1992), and Spall (2005).

The simulated annealing algorithm proceeds as follows. First, initial scenarios \( \theta_{1}^{1}, \ldots, \theta_{K}^{1} \) are proposed, for example by sampling \( \theta_{1}^{1}, \ldots, \theta_{K}^{1} \) from a probability distribution with support \( \Theta \). Then, iteratively for \( 1 \leq i \leq I \), the current scenarios \( \theta_{1}^{i}, \ldots, \theta_{K}^{i} \) are perturbed by adding to them zero-mean noise variables \( z_{1}^{1}, \ldots, z_{K}^{1} \), thus, obtaining new proposed scenarios \( \theta_{1}^{i}, \ldots, \theta_{K}^{i} \) (this step is represented by the “Perturb” operator in Algorithm 3). At each iteration, the proposed scenarios \( \theta_{1}^{i}, \ldots, \theta_{K}^{i} \) can either be accepted (i.e., \( \theta_{1}^{1}, \ldots, \theta_{K}^{1} \rightarrow \theta_{1}^{i}, \ldots, \theta_{K}^{i} \)) or rejected (i.e., \( \theta_{1}^{i}, \ldots, \theta_{K}^{i} \rightarrow \theta_{1}^{1}, \ldots, \theta_{K}^{1} \)). The acceptance or rejection of the proposed scenarios is stochastic, with probability \( \rho_{i} \) (defined below), which is a function of \( \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \) and \( \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \).

The acceptance probability \( \rho_{i} \) is equal to 1 when \( \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) < \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \). That is, if the proposed scenarios decrease the current loss value, then the proposed scenarios are accepted. If instead \( \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \geq \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) \), then \( \rho_{i} \) is

\[
\rho_{i} = \exp \left( \frac{\hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K}) - \hat{\mathcal{L}}(\theta_{1}, \ldots, \theta_{K})}{T_{i}} \right),
\]

where \( T_{i} \), \( 0 \leq i \leq I \), is a decreasing sequence of positive real numbers often called the “cooling schedule” of the algorithm. A common cooling schedule is \( T_{i} = T_{0} \cdot r^{i-1} \), where \( T_{0} \) is a constant and \( r \in (0, 1) \) is a multiplicative contraction, but other forms are possible (Spall 2005). In our applications, we use a piecewise-constant cooling schedule (Husmann, Lange, and Spiegel 2017).

After simulating the outlined Markov Chain for a fixed number \( I \) of iterations, the final set of scenarios \( \{\theta_{1}^{1}, \ldots, \theta_{K}^{I}\} \) approximately minimizes the loss function \( \hat{\mathcal{L}} \) (Bélisle 1992).
In our ROSA implementation, we use multiple independent replicates of Algorithm 3, with different initial scenarios \( \theta^1_1, \ldots, \theta^k_K \), to investigate convergence of the random trajectory \([\theta^1_i, \ldots, \theta^k_i]; i \geq 1 \). Intuitively, we evaluate if the replicated trajectories, with different starting values, terminates with nearly identical final vectors \( \{\theta^1_{i+1}, \ldots, \theta^k_{i+1}\} \) and negligible differences in the \( \hat{L}(\theta^1_{i+1}, \ldots, \theta^k_{i+1}) \) values.

3. Applications: Sensitivity Analyses of Trial Designs

We illustrate the ROSA approach by performing sensitivity analyses for three designs of different complexity levels. In each example, we describe the design of the trial, the unknown parameters, and the operating characteristics of interest. By illustrating the ROSA methodology in three trial designs, we show its flexibility with potential applications to evaluate nearly any clinical trial design. Indeed, ROSA only requires the possibility of simulating the trials under potential unknown parameters \( \Theta \) and the definition of the operating characteristics of interest.

3.1. Application 1: Two-Arm RCT

In the first example, we will only consider a single unknown parameter (i.e., \( \theta \in \mathbb{R} \)) and a single operating characteristic \( f(\theta) \) that can be computed analytically. In this case, the optimal set of scenarios \( \{\theta^*_1, \ldots, \theta^*_k\} \) can be computed exactly, without resorting to approximation methods. This simple and stylized setting is useful to highlight the similarity of the approximations and selected scenarios computed by ROSA with their exact counterparts.

3.1.1. Trial Design

We consider the design of a two-arm randomized trial (1:1 randomization ratio) with a sample of \( n = 30 \) patients. For each \( i = 1, \ldots, n \), we let \( A_i = 0 \) or 1 if the \( i \)th study patient is assigned to the control or experimental arm. The outcomes of the \( n \) study patients are \( Y_1, \ldots, Y_n \), which we assume to be independent and normally distributed. If \( A_i = a \) then \( Y_i \) has mean \( \mu_a \) and standard deviation \( \sigma \) equal to 30. In the analysis of the study, a \( z \)-statistic will be used to test the null hypothesis \( H_0: \mu_1 - \mu_0 \leq 0 \) against the alternative \( H_1: \mu_1 - \mu_0 > 0 \) at 5% significance level.

3.1.2. Aim of the Sensitivity Analysis

The goal of the sensitivity analysis is to summarize the variation of the probability of rejecting \( H_0 \), a function \( f(\theta) \) of the unknown treatment effect \( \theta = \mu_1 - \mu_0 \in \Theta = \mathbb{R} \). For example, if we knew that \( \theta = 13.5 \), then \( f(\theta) = 0.80 \), but in general \( \theta \) is an unknown value. Suppose we aim to identify \( K = 3 \) scenarios \( \theta^*_1, \theta^*_2, \theta^*_3 \) that maximize the utility \( l(\cdot) \), that is,

\[
\theta^*_1, \theta^*_2, \theta^*_3 = \arg\max_{\theta_1, \theta_2, \theta_3 \in \Theta} l(\theta_1, \theta_2, \theta_3),
\]

where \( l(\theta_1, \theta_2, \theta_3) = -\max_{\theta \in \Theta} \min_{z \in \{-1, 0, 1\}} [f(\theta') - f(\theta)] \).

In this trial, we have a single unknown parameter \( \Theta = \mathbb{R} \), and the operating characteristic of interest is monotone, continuous, invertible, and ranges from 0 to 1. Therefore, it is straightforward to see that the optimal scenarios \( \theta^*_1, \theta^*_2, \theta^*_3 \) correspond to the operating characteristic values that evenly divide the interval \( (0, 1) \). To be precise, \( \{f(\theta^*_1), f(\theta^*_2), f(\theta^*_3)\} = \{1/6, 1/3, 1/6\} \); these are the three values of a regular grid on the interval \( (0, 1) \). Figure 2(a) illustrates the optimal set of scenarios when \( K = 3, 5, 10 \). Since \( f(\theta) \) can be calculated exactly, the optimal scenarios \( \theta^*_1, \theta^*_2, \theta^*_3 \) can be obtained by computing the inverse function \( f^{-1} \) at the values \( 1/6, 1/3, 5/6 \).

3.1.3. Implementing and Benchmarking ROSA

The exact computation of the optimal set of scenarios provides a solid benchmark for an initial evaluation of ROSA (Algorithms 1-3). We can compare the exact solution with the results from ROSA, which has the advantage of being applicable to other designs and operating characteristics that are not available in closed form.

We implement our ROSA approach to identify \( K = 3 \) scenarios. We randomly select \( J = 1000 \) scenarios \( \theta^1_1, \ldots, \theta^1_{1000} \) with independent samples from the Uniform\((-5, 25)\) distribution. Note that \( f(-5) \approx 0 \) and \( f(25) \approx 1 \). For each \( \theta^1_j, 1 \leq j \leq 1000 \), we simulate \( M = 200 \) trials to compute the estimate \( \hat{f}(\theta^1_j) = 200^{-1} \sum_{m=1}^{200} \psi(Z_{jm}, \theta^1_j) \), where \( \psi(Z_{jm}, \theta^1_j) \in \{0, 1\} \) either accepts or rejects the hypothesis \( H_0 : \theta^1_j \leq 0 \) for trial \( m \) and scenario \( j \). Then, we compute a continuous function \( \tilde{f}(\theta) \) using the independent estimates \( \tilde{f}(\theta^1_j) \) and a NN with 3 hidden layers (8, 64, and 64 neurons, respectively) and ReLU activation functions. Finally, to select three sensitivity scenarios, we use a simulated annealing algorithm based on an initial parameterization.
Figure 2. Sensitivity analysis of a RCT (operating characteristic: probability of rejecting $H_0$). Panel A: Exact solutions when $K = \{3, 5, 10\}$. Panel B: Comparison of $K = 3$ scenarios selected through exact calculation (red asterisks) and by 20 ROSA implementations with different initial proposals (blue points). Panel C: Graphical tool to choose the number $K$ of sensitivity scenarios.

$T_1 = 1000$ and final parameterization $T_{min} = 0.1$ (see, Algorithm 3). We repeat these three steps (selection of scenarios, use of the NN, and optimization with simulated annealing) 20 times, each time initializing $\theta_1, \theta_2, \theta_3$ with independent random draws from the Uniform$(-5, 25)$ distribution. The results of the exact approach (red asterisks) compared with ROSA (blue points) are shown in Figure 2(b). The scenarios $\theta_1^*, \theta_2^*, \theta_3^*$ selected by simulated annealing (blue dots) are close to the exact solution (red asterisks).

### 3.1.4. Choice of Number K of Scenarios

In practice, the decision regarding the number $K$ of scenarios to report is left to the analyst. This choice can be supported by a graph like Figure 2(c), which allows the investigator to determine the minimum number $K$ of scenarios needed to guarantee a loss $L(\theta_1^*, \ldots, \theta_K^*)$ no larger than a targeted threshold. For example, to guarantee a loss no larger than 0.050 in this example, we need to select at least 10 scenarios for the simulation report.

We ran ROSA with $K = 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30$, and compared the loss $L$ in the resulting set of scenarios with that of the exact solution. The difference in the loss $L$ of the exact and approximate optima was less than 1% across all $K$ values that we considered (Figure 2(c)). Table 2 indicates that the computation time of the simulated annealing algorithm scales well as $K$ increases and that, as expected, the loss $L$ decreases as $K$ increases. All analyses were run on a Windows laptop with an Intel(R) Core(TM)i7-7700HQ 2.80 GHz processor, 16GB RAM, and 6MB of cache memory.

### Table 2. ROSA computation time, ROSA loss $L$, minimum (exact) loss $L$, and relative difference in loss of ROSA scenarios compared to the exact solutions.

| Number K of Scenarios | Time (seconds) | ROSA Loss $L$ | Min. Loss $L$ | Rel. Diff. |
|-----------------------|---------------|---------------|---------------|------------|
| 5                     | 8.8           | 0.101         | 0.100         | 1.0%       |
| 6                     | 8.8           | 0.084         | 0.083         | 0.7%       |
| 7                     | 9.1           | 0.072         | 0.071         | 0.8%       |
| 8                     | 9.2           | 0.062         | 0.0625        | 0.7%       |
| 9                     | 9.1           | 0.056         | 0.056         | 0.6%       |
| 10                    | 9.1           | 0.050         | 0.050         | 0.2%       |
| 20                    | 10.1          | 0.025         | 0.025         | 0.5%       |
| 30                    | 10.2          | 0.017         | 0.0167        | 0.8%       |

3.2. Application 2: Interim Decisions Based on Auxiliary Outcomes

In the second example, we consider sensitivity analyses with multiple unknown parameters and two operating characteristics. We illustrate the use of our computational procedures, including the operating characteristics approximation procedure (Algorithm 1), the validation procedure (Algorithm 2), and the simulated annealing optimization procedure (Algorithm 3). We investigate whether it is appropriate to fix the value of some of the unknown parameters across all sensitivity scenarios. Identical values for a subset of the unknown parameters can simplify the interpretation of the sensitivity analysis but can...
also introduce severe limitations in faithfully representing how the operating characteristics vary across plausible values of the unknown parameters.

### 3.2.1. Trial Design

We consider a two-arm, two-stage randomized trial with a binary primary outcome \( Y \) and a binary auxiliary outcome \( S \) (Niewczasz, Kunz, and König 2019). The primary outcome \( Y \) is available \( T_Y \) months after randomization, while the auxiliary outcome \( S \) is available after \( T_S < T_Y \) months. For example, in glioblastoma trials, 12-month progression-free survival (PFS) and 24-month overall survival (OS) have been used as auxiliary and primary outcomes, respectively (Han et al. 2014). The approach that we illustrate is applicable for any value of \( T_Y \) and \( T_S < T_Y \).

We let \( N_a \) be the planned number of patients for arms \( a = 0, 1 \) (i.e., control and experimental arms) and indicate with \( p_a \) the response probability \( P(Y = 1 \mid A = a) \). Similarly, let \( n_a \) be the planned number of patients assigned to arm \( a \) before the interim analysis, and \( q_a \) indicate the response probability \( P(S = 1 \mid A = a) \). The difference \( \Delta = p_1 - p_0 \) is the treatment effect on \( Y \). The primary aim of the trial is to test \( H_0 : \Delta \leq 0 \) versus \( H_1 : \Delta > 0 \), at level \( \alpha \). The design of Niewczasz, Kunz, and König (2019), which calculates the conditional power (CP) using the standard normal distribution, here we set the cutoff point to be \( \alpha \). Specifically, the CP is calculated based on \( Z_S = \frac{\hat{p}_1 - \hat{p}_0}{(\hat{q}_1 - \hat{q}_0)\sqrt{\hat{q}(1-\hat{q})(n_1^{-1} + n_0^{-1})}} \), where \( \hat{p}_a \) is the estimate of \( p_a \) and \( \hat{q}_a \) is a weighted average of \( q_1 \) and \( q_0 \). We replicate the design of Niewczasz, Kunz, and König (2019), which calculates at the interim analysis the conditional power (CP) using the auxiliary outcome \( S \) to determine whether to stop the trial for futility or not. Specifically, the CP is calculated based on \( Z_S \) and the information fraction \( t_S = \frac{N_1^{-1} + N_0^{-1}}{n_1^{-1} + n_0^{-1}} \) as

\[
CP(t_S) = 1 - \Phi \left( \frac{Z_1 - \alpha - Z_S^{1/2}}{\sqrt{1 - t_S}} \right),
\]

where \( Z_1 - \alpha \) is the \( 1 - \alpha \) quantile of the standard normal distribution and \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution. Here, we set the cutoff point to be 0.5 so that the trial continues when \( CP(t_S) \geq 0.5 \).

### 3.2.2. Aim of the Sensitivity Analysis

The complexity of the simulation report increases with \( K \) (the number of scenarios), \( d \) (the number of entries of the unknown parameters \( \Theta \)), and \( R \) (the number of operating characteristics \( f(\Theta) \)). Here the full set of unknown parameters \( \Theta \subset \mathbb{R}^d \) include the enrollment rate \( e \in (0, \infty) \), the response rates \( p_a \in (0, 1) \) for \( Y \) in \( A = a \), the response rates \( q_a \in (0, 1) \) for \( S \) in \( A = a \), and the correlation between \( Y \) and \( S \) in \( A = a, \rho_a \in (-1, 1) \).

Controlling the complexity of the simulation report is important to ensure high interpretability of the report, which will be discussed by several stakeholders. There are a few potential strategies to reduce the complexity of the simulation report. First, it is often possible to consider only a subset of the parameter space \( \Theta' \subset \Theta \) based on prior knowledge of plausible values of the unknown parameters. For example, previous clinical studies can indicate a plausible range for the enrollment rate \( e \), the response rates \( p_0 \) under the SOC, and other parameters that are expected to have minimal variations across trials. In addition, we can also consider fixing multiple entries of the \( K \) vectors \( \theta_1, \ldots, \theta_K \) to some reference values. In this case the space from which we select scenarios \( \theta_1, \ldots, \theta_K \) is further reduced to \( \Theta'_{\text{ref}} \subset \Theta' \). For example, if the operating characteristics have low sensitivity with respect to the correlation parameters \( \rho_a \) or the enrollment rate \( e \) of the study, then we can fix these unknown parameters to common values (i.e., estimates) across all \( K \) scenarios.

ROSA allows us to evaluate whether it is appropriate to assign the same value to one or more unknown parameters (e.g., \( p_0 \) and \( p_1 \)) across all \( K \) scenarios. In other words, we evaluate a simulation report with all scenarios in a restricted subset \( \Theta'_{\text{ref}} \subset \Theta' \). A simulation report with scenarios in \( \Theta'_{\text{ref}} \) can potentially be easier to interpret compared to a report in which all \( d \) entries of \( \theta \) vary across scenarios by reducing the number of dimensions \( d \) of the unknown parameters and pointing to the most relevant unknown parameters when discussing the variations of the operating characteristics across \( \Theta' \). We can select scenarios from the restriction \( \Theta'_{\text{ref}} \subset \Theta' \) only if the capability of the simulation report of representing the operating characteristics variations across \( \Theta' \) is preserved. Our case study investigates this aspect. The operating characteristics of interest \( f \) in our case study are the probability of rejecting the null hypothesis of no treatment effect on \( Y \) at the end of the study and the average sample size.

### 3.2.3. Implementing and Benchmarking ROSA

Using our ROSA procedure, we randomly select \( J = 1000 \) training scenarios using LHS and conduct \( M = 500 \) Monte Carlo simulations for each of the \( J \) training scenarios to obtain estimates of the operating characteristics across \( \Theta' \). Here \( \Theta' \) is a product space with the enrollment rate \( e \in (0, 2, 1) \), the response rates \( p_0 \in (0.2, 0.4) \) for \( Y \) in \( A = a \), the response rates \( q_1 \in (0.2, 0.4) \) for \( S \) in \( A = a \), and the correlation between \( Y \) and \( S \) in \( A = a, \rho_a \in (0, 0.6) \). For \( \Theta'_{\text{ref}} \), we fix the enrollment rate \( e = 0.5 \) and the response rates \( p_0 = q_0 = 0.3 \) in the control groups.

We use a NN to obtain an interpolation of the operating characteristics. As described in Algorithm 2, to evaluate if the estimates of the operating characteristics are accurate, we compare them to independent Monte Carlo estimates of size \( M = 100,000 \) on a set of \( J' = 200 \) uniformly-distributed validation points spanning the plausible parameter space \( \Theta' \). The coefficients of determination \( R^2 \) in this comparison are above 0.96. This suggests that the NN accurately estimates the operating characteristics.

We compare two simulation reports, and our goal is to provide stakeholders the simplified version if it accurately describes the operating characteristics. The first one includes scenarios
from $\Theta' \subset \mathbb{R}^2$ restricted by prior knowledge from completed studies and clinical experience and the second includes scenarios from $\Theta'_{re} \subset \Theta'$ further restricted by fixing the value of some entries of $\theta$ as described above. We use simulated annealing to identify two sets of scenarios in $\Theta'$ and $\Theta'_{re}$, respectively. In both cases we minimize the same loss function $L$ defined over $K$-tuples of $\Theta'$ points. We also calculate the loss $L$ associated with these two optimal sets of scenarios from $\Theta'$ and $\Theta'_{re}$. In Figure 3, we illustrate the difference in loss $L$ between these two optimal sets; as expected, the loss $L$ decreases as $K$ increases. We observe in Figure 3 that for any value of $K$, the loss $L$ associated with the optimal set of scenarios restricted to $\Theta'_{re}$ is larger compared to the optimal scenarios in $\Theta'$. However, the difference is modest, and the gain in interpretability of a sensitivity analysis report with fewer unknown parameters may be worth the slightly larger loss. For example, if an investigator requires the loss to be under a threshold of $L = 0.2$, then it is sufficient to consider $K = 10$ scenarios, regardless of whether we consider scenarios selected from $\Theta'$ or $\Theta'_{re}$.

3.3. Application 3: Biomarker-Driven Adaptive Enrichment

In the third example, we discuss sensitivity analyses dedicated to an adaptive trial with subpopulations defined by biomarkers, considering multiple unknown parameters and multiple operating characteristics of interest. As a motivating example, in several oncology trials, a major decision is whether to restrict patient enrollment to a targeted subgroup of patients (e.g., biomarker-positive subgroup) or to enroll a broader patient population. Enrolling only a biomarker-positive subgroup may deny a substantial number of patients access to an effective therapy; whereas enrolling a larger population may compromise the power to detect positive treatment effects. Several trial designs discussed in the literature attempt to address the outlined problem through interim looks at the data.

3.3.1. Trial Design

We consider an adaptive two-stage enrichment trial design with one-to-one randomization (Jenkins, Stone, and Jennison 2011; Jones et al. 2017; Mehta, Liu, and Theuer 2019). The design is applicable in the setting where a biomarker-positive subgroup of patients is hypothesized to benefit more from the experimental treatment than the rest of the study population. The design includes a single interim analysis, and it uses progression-free survival (PFS) for interim decision-making, while overall survival (OS) is the endpoint for the final analysis, which occurs when a pre-specified number of events is reached. The interim analysis uses the estimated PFS hazard ratio (HR) to capture potential early signals of treatment effects. In the implementation of Jenkins, Stone, and Jennison (2011), which we replicate, the HR is estimated for both the overall population ($\hat{HR}_{HR}$) and the biomarker-positive subgroup ($\hat{HR}^+_{HR}$). An interim decision determines which group is enrolled and tested during the second stage of the trial:

A promising results in the biomarker-positive population. If the HR estimate $\hat{HR}_{HR} < 0.6$ but $\hat{HR}_{HR} \geq 0.8$, then the trial will continue enrolling only biomarker-positive patients and the final analysis will test $H^0_0$. Here $H^0_0$ is the null hypothesis of no differences in OS between treatment and control groups in the biomarker-positive population. The null hypothesis is rejected if $\omega_1 \Phi^{-1}(1 - p^+_1) + \omega_2 \Phi^{-1}(1 - p^+_2) < 1.96$, where $p^+_1$ ($p^+_2$) is a log-rank $p$-value computed using only OS data from patients randomized during the first (second) stage of the trial. The weights ($\omega_1, \omega_2$) and the standard normal cumulative distribution function $\Phi$ are used to summarize evidence of
treatment effects from the two stages of the trial. We refer to Jenkins, Stone, and Jennison (2011) for details on the choice of \((\omega_1, \omega_2)\) and other aspects of the final analysis.

B - Promising results in the overall population only. If \(\hat{\theta}_{HR}^+ \geq 0.6\) but \(\hat{\theta}_{HR} \leq 0.8\), then the trial will continue enrolling all patients and the final analysis will only test \(H_0^f\), the null hypothesis of no differences in OS in the overall population. In this case the null hypothesis is tested using stage-specific OS log-rank p-values \((p_1^f, p_2^f)\) and combining evidence from the two stages of the trial.

C - Unpromising results. If \(\hat{\theta}_{HR}^+ \geq 0.6\) and \(\hat{\theta}_{HR} \leq 0.8\), then the trial stops early for futility.

D - Promising early results for both populations. Lastly, if the estimated HR in the biomarker-positive subgroup \(\hat{\theta}_{HR}^+ \leq 0.6\) and the overall population \(\hat{\theta}_{HR} \leq 0.8\), then the trial will continue enrolling all patients and testing efficacy both in the overall population and in the biomarker-positive subgroup.

The potential conclusion at the final analysis are (i) to recommend the new treatment for biomarker-positive patients, (ii) recommend the new treatment for both biomarker-positive and biomarker-negative patients, or (iii) not recommend the experimental treatment for future patients.

### 3.3.2. Aims of the Sensitivity Analysis

We focus on the following three operating characteristics: (i) \(f_1\), the probability of enrolling only biomarker-positive patients in the second stage, (ii) \(f_2\), the probability of enrolling both biomarker-positive and biomarker-negative patients in the second stage, and (iii) \(f_3\), the probability of no evidence of positive treatment effects, which is equal to the probability of not rejecting the null hypotheses.

We choose plausible intervals for the unknown parameters based on prior literature. Specifically, the recruitment rate \(\theta_1 \in (0.5, 1)\) per week, the prevalence of the biomarker-positive subgroup \(\theta_2 \in (0.15, 0.25)\), the PFS HR comparing the treatment and control groups in the biomarker-positive subgroup \(\theta_3 \in (0.5, 1.2)\), the PFS HR comparing treatment and control in the biomarker-negative subgroup \(\theta_4 \in (0.6, 1.2)\), the OS HR comparing treatment and control in the biomarker-positive subgroup \(\theta_5 \in (0.7, 1.2)\), the OS HR comparing treatment and control groups in the biomarker-negative subgroup \(\theta_6 \in (0.8, 1.2)\), the correlation between OS and PFS in the biomarker-positive subgroup \(\theta_7 \in (0.3, 0.6)\), and the correlation between OS and PFS in the biomarker-negative subgroup \(\theta_8 \in (0.2, 0.7)\). Marginal exponential distributions and latent frailty terms were used for simulating correlated OS and PFS times (Michael and Schucany 2002). More flexible models such as the Weibull distribution can be considered.

### 3.3.3. Implementing and Benchmarking ROSA

For the outlined two-stage trial with biomarker populations, our ROSA pipeline can be used to compute multiple simulation reports, varying both the list of operating characteristics \(f\) and the definition of \(\Theta^f\). For example, one can fix the OS HRs in the biomarker-positive and negative populations to focus on the design sensitivity to other parameters, such as the PFS HRs. Similarly, the set of unknown parameters \(\Theta^f\) can be restricted to \(\theta\) values with positive effects only for the biomarker-positive population. Importantly, one set of training simulations can be re-used to compute multiple sensitivity tables where the definitions of \(f\) and \(\Theta^f\) vary.

We examine the difference in the marginal losses

\[
\mathcal{L}_r(\theta_1, \ldots, \theta_K) = \max_{\theta \in \Theta} \left\{ \min_{k=1, \ldots, K} \left\| f_r(\theta) - f_r(\theta_k) \right\|_2 , \right. \\
1 \leq r \leq R,
\]

when the set of scenarios are chosen by optimizing different loss functions. For example, let \(S^r\) be the set of scenarios that minimize the marginal loss \(\mathcal{L}_r\) in (4). Similarly, let \(S^r\) be the set of scenarios that minimize the joint loss \(\mathcal{L} = -\mathcal{L}_r(\theta)\) in (2). Then it is intuitive that \(\mathcal{L}_r(S^r) \leq \mathcal{L}_r(S)\), \(1 \leq r \leq R\). In different words, the marginal losses \(\mathcal{L}_r\) tend to be smaller when the set of scenarios is chosen to minimize \(\mathcal{L}_r\) compared to a set of scenarios that minimizes \(\mathcal{L}\) with the aim of representing multiple operating characteristics. If the discrepancy \(\mathcal{L}_r(S^r) - \mathcal{L}_r(S)\) is relatively small for all \(R\) total operating characteristics, then this indicates that it is reasonable to select a single set of scenarios \(S\) to illustrate how the \(R\) operating characteristics vary jointly across \(\Theta\). We describe the difference between the marginal losses \(\mathcal{L}_r, r = 1, 2, 3\), when scenarios \(\theta_1, \ldots, \theta_K\) in \(\Theta^f\) are chosen by optimizing \(\mathcal{L}_r\) in (4) – optimum: \(S^r = \theta_1^*, \ldots, \theta_K^*\), or by optimizing \(\mathcal{L}\) as in (2) – optimum: \(S = \theta_1^*, \ldots, \theta_K^*\). Recall that \(S\) is computed with the goal of illustrating how multiple operating characteristics vary across \(\Theta\). Our results indicate that for all three operating characteristics, \(\mathcal{L}_r(S) > \mathcal{L}_r(S^r), r = 1, 2, 3\); as expected, there is an increase of the marginal losses \(\mathcal{L}_r\) when the set of scenarios is selected to illustrate jointly the variations of multiple operating characteristics across \(\Theta\). However, this difference is small (\(< 10\%) for all \(K \in \{2, 5, 10, 15\}\). Furthermore, for each \(K \in \{2, 5, 10, 15\}\), the relative difference is similar across the three operating characteristics \(f_1, f_2, f_3\) (Figure 4). This result supports the use of identical weights and of a single sensitivity table, with the same set of scenarios \(S\) to illustrate jointly all three operating characteristics.

### 4. Discussion

The evaluation of complex designs such as dose-finding studies (Iasonos, Gönen, and Bosl 2015), factorial trials (Green, Liu, and O’Sullivan 2002), and response-adaptive trials (Pallmann et al. 2018) focuses on multiple operating characteristics, such as the level of toxicities, the probability of selecting the correct treatment arm, or frequentist operating characteristics, including power and false positive probabilities. During the design stage of a complex clinical trial, simulation reports are typically produced to discuss sample size, interim analyses, and other major decisions with various stakeholders. The simulation report consists of one or a few tables dedicated to showcasing how major operating characteristics \(f(\theta)\) vary across potential values of unknown parameters in \(\Theta\). In most cases, the analyst
focuses on subsets of plausible parameters \( \Theta' \subset \Theta \), for example, values concordant with previous studies, or subsets of potential \( \theta \) values of particular interest, for example with positive and clinically relevant treatment effects.

Simulations are fundamental in the design of complex trials since operating characteristics can rarely be obtained analytically and are crucial in the assessment of study designs for regulators, pharmaceutical companies and other stakeholders (U.S. FDA 2020). However, a limited number of scenarios or poorly chosen scenarios could be inadequate to highlight variations of the operating characteristics across plausible unknown parameters and can result in sub-optimal decisions. We propose ROSA as a useful tool that can support investigators at this design stage when selecting which and how many scenarios to include in these simulation reports.

We focus on choosing an informative number \( K \) of scenarios \( \theta_1, \ldots, \theta_K \) among the plausible unknown parameters to summarize the variations of key operating characteristics. Our approach minimizes an explicit loss function and uses established techniques for functional approximation (NNs) and numerical optimization (simulated annealing). We showcase our approach in three trials. Importantly, our approach is general and can be applied to nearly any clinical trial design. It only requires simulations to mimic the clinical trial under hypothetical scenarios.

Although our approach is general, we focused on loss functions \( \mathcal{L} \) of a specific form (2). It is possible to consider different loss functions. For example, one could consider the loss function
\[
\hat{\mathcal{L}}(\theta_1, \ldots, \theta_K) = E_{g^* \sim g(\cdot)} \left\{ \min_{k=1, \ldots, K} D[f(\theta')_k, f(\theta_k)] \right\},
\]
where \( g(\cdot) \) is a probability distribution on \( \Theta \) (e.g., a posterior distribution obtained from previous data). The distribution \( g \) could be used to incorporate prior information about the unknown parameters in the selection of sensitivity scenarios. Moreover, the metric \( D : \Theta^2 \rightarrow \mathbb{R} \) can be extended to capture both differences between operating characteristics at plausible values \( \theta, \theta' \in \Theta \) and other aspects, such as the difference between expected values of the outcomes \( Y \) at \( \theta \) and \( \theta' \).

One major challenge in the presentation of simulation reports is the need for simplicity and interpretability of the results. To this end, we considered fixing one or more unknown parameters to identical values across the \( K \) scenarios, which may be reasonable when there is priori knowledge of certain unknown parameters. There are other ways to simplify a simulation report, such as removing operating characteristics that do not vary across plausible unknown parameters, or reporting only the range of the operating characteristics across \( \Theta \) instead of presenting the operating characteristics for each representative scenario.

Variations of the ROSA approach may also consider optimization algorithms other than simulated annealing and regression methods alternative to NN for approximating the operating characteristics across \( \Theta \). The methodology that we proposed here can be used to handle other relevant problems, such as missing data. Indeed, during the design of the trial, there is often uncertainty on whether the analyses will involve missing data or not and the potential consequences of the missingness pattern. Probability models that include pretreatment variables, outcomes, and missing data patterns are useful to explore the robustness of the design. In this case, ROSA can support the selection of scenarios with different missing data patterns.

**Supplementary Materials**

The online supplement includes example code (ExampleCode.R) for implementing ROSA.

**Acknowledgments**

The authors thank Cyrus Mehta and Christina Howe for helpful conversations and feedback that greatly enhanced the article.

**Disclosure Statement**

The authors report there are no competing interests to declare.
**Funding**

LH was supported by the Clinical Orthopedic and Musculoskeletal Education and Training (COMET) Program, NIAMS grant T32 AR055885. LT was supported by the NIH grant R01LM013352.

**ORCID**

Larry Han ORCID id: http://orcid.org/0000-0002-0577-9661

**References**

Bélisle, C. J. (1992), "Convergence Theorems for a Class of Simulated Annealing Algorithms on \( \mathbb{R}^d \)," *Journal of Applied Probability*, 885–895. [77,80]

Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010), *Bayesian Adaptive Methods for Clinical Trials*, Boca Raton, FL: CRC Press. [76]

Bookstein, F. L. (1989), "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 567–585. [79]

Goodfellow, I., Bengio, Y., and Courville, A. (2016), *Deep Learning*, Cambridge, MA: MIT Press. [77,80]

Green, S., Liu, P.-Y., and O’Sullivan, J. (2002), "Factorial Design Considerations," *Journal of Clinical Oncology*, 20, 3424–3430. [85]

Han, K., Ren, M., Wick, W., Abrey, L., Das, A., Jin, J., and Reardon, D. A. (2014), "Progression-Free Survival as a Surrogate Endpoint for Overall Survival in Glioblastoma: A Literature-based Meta-Analysis from 91 Trials," *Neuro-Oncology*, 16, 696–706. [83]

Hobbs, B. P., Barata, P. C., Kanjaniyan, Y., Paller, C. J., Perlmutter, J., Pond, G. R., Prowell, T. M., Rubin, E. H., Seymour, L. K., Wages, N. A., et al. (2019), "Seamless Designs: Current Practice and Considerations for Early-Phase Drug Development in Oncology," *JNCI Journal of the National Cancer Institute*, 111, 118–128. [76]

Hornik, K. (1991), "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, 4, 251–257. [80]

Husmann, K., Lange, A., and Spiegel, E. (2017), "The r Package Optimization: Flexible Global Optimization with Simulated-Annealing," [80]

Iasonos, A., Gönen, M., and Bosl, G. J. (2015), "Scientific Review of Phase I Protocols with Novel Dose-Escalation Designs: How Much Information is Needed?" *Journal of Clinical Oncology*, 33, 2221–2225. [85]

Jenkins, M., Stone, A., and Jennison, C. (2011), "An Adaptive Seamless Phase II/III Design for Oncology Trials with Subpopulation Selection Using Correlated Survival Endpoints," *Pharmaceutical Statistics*, 10, 347–356. [77,84,85]

Jones, R. L., Attia, S., Mehta, C. R., Liu, L., Shankala, K. K., Robinson, S. L., Ravi, V., Penel, N., Stacchioti, S., Tap, W. D., et al. (2017), "Tappas: An Adaptive Enrichment Phase 3 Trial of TRC105 and Pazopanib Versus Pazopanib Alone in Patients with Advanced Angiosarcoma (AAS)," *Journal of Clinical Oncology*, 35, TPS11081. [77,84]

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680. [80]

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993), "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function," *Neural Networks*, 6, 861–867. [80]

Mayer, C., Perevorskaya, I., Leonov, S., Dragalin, V., Pritchett, Y., Bedding, A., Hartford, A., Fardipour, P., and Ciconnetti, G. (2019), Simulation Practices for Adaptive Trial Designs in Drug and Device Development," *Statistics in Biopharmaceutical Research*, 11, 325–335. [77]

McKay, M. D., Beckman, R. J., and Conover, W. J. (2000), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 42, 55–61. [79]

Mehta, C., Liu, L., and Theuer, C. (2019), "An Adaptive Population Enrichment Phase III Trial of TRC105 and Pazopanib Versus Pazopanib Alone in Patients with Advanced Angiosarcoma (Tappas Trial)," *Annals of Oncology*, 30, 103–108. [76,77,84]

Michael, J., and Schucany, W. (2002), "The Mixture Approach for Simulating New Families of Bivariate Distributions with Specified Correlations," *The American Statistician*, 56, 48–54. [85]

Niewczas, J., Kunz, C. U., and König, F. (2019), "Interim Analysis Incorporating Short- and Long-Term Binary Endpoints," *Biometrical Journal*, 61, 665–687. [77,83]

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., et al. (2018), "Adaptive Designs in Clinical Trials: Why Use Them, and How to Run and Report Them," *BMC Medicine*, 16, 1–15. [85]

Rasmussen, C. E. (2003), "Gaussian Processes in Machine Learning," in *Summer School on Machine Learning*, eds. O. Bousquet, U. Luxburg, and G. Rätsch, pp. 63–71, Berlin: Springer. [79]

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Piano, S. L., Iwanaga, T., Becker, W., et al. (2021), "The Future of Sensitivity Analysis: An Essential Discipline for Systems Modeling and Policy Support," *Environmental Modelling & Software*, 137, 104954. [77]

Spall, J. C. (2005), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Hoboken, NJ: Wiley. [80]

Thorlund, K., Haggstrom, J., Park, J. J., and Mills, E. J. (2018), "Key Design Considerations for Adaptive Clinical Trials: A Primer for Clinicians," *BMJ*, 360, k698. [76]

US Food and Drug Administration (2020), "Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products," updated December. [77,86]