

## RESEARCH ARTICLE

# Improving adaptive seamless designs through Bayesian optimization

Jakob Richter<sup>1</sup>  | Tim Friede<sup>2,3</sup>  | Jörg Rahnenführer<sup>1</sup><sup>1</sup>Fakultät Statistik, Technische Universität Dortmund, Dortmund, Germany<sup>2</sup>Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Göttingen, Germany<sup>3</sup>Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK), Standort Göttingen, Göttingen, Germany**Correspondence**Jakob Richter, Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany.  
Email: richter@statistik.tu-dortmund.de**Funding information**

Deutsche Forschungsgemeinschaft, Grant/Award Number: SFB 876



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**Abstract**

We propose to use Bayesian optimization (BO) to improve the efficiency of the design selection process in clinical trials. BO is a method to optimize expensive black-box functions, by using a regression as a surrogate to guide the search. In clinical trials, planning test procedures and sample sizes is a crucial task. A common goal is to maximize the test power, given a set of treatments, corresponding effect sizes, and a total number of samples. From a wide range of possible designs, we aim to select the best one in a short time to allow quick decisions. The standard approach to simulate the power for each single design can become too time consuming. When the number of possible designs becomes very large, either large computational resources are required or an exhaustive exploration of all possible designs takes too long. Here, we propose to use BO to quickly find a clinical trial design with high power from a large number of candidate designs. We demonstrate the effectiveness of our approach by optimizing the power of adaptive seamless designs for different sets of treatment effect sizes. Comparing BO with an exhaustive evaluation of all candidate designs shows that BO finds competitive designs in a fraction of the time.

**KEYWORDS**

adaptive seamless designs, Bayesian optimization, clinical trials, treatment selection

## 1 | INTRODUCTION

Clinical research and in particular drug development is typically structured into phases, for example, phases I–IV in drug development. Similar approaches exist for complex interventions, such as the medical research council (MRC) framework for nonpharmacological interventions (Campbell et al., 2000). The development phases include elements of learning and confirmation. For instance, Sheiner (1997) describe the drug development process as two learning and confirming cycles: The first cycle includes learning about the tolerated dose (Phase I) and confirming the efficacy of the selected dose in a selected group of patients (Phase IIa), whereas the second cycle consists of learning about the optimal use in representative patients (Phase IIb) and confirming an acceptable benefit or risk ratio (Phase III).

Traditionally, separate studies are performed for learning and confirming. However, the landscape in clinical research is changing. Boundaries between the development phases are increasingly dissolving in multiple aspects. Master protocols

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

are a new concept where multiple subgroups and substudies are integrated (Bogin, 2020). Here, simultaneously, more than one treatment and more than one disease type are investigated within the same overall trial structure. Adaptive designs are established, but are being further developed. In general, study designs become more flexible, and more principled approaches to decision-making using statistical modeling and integrated analyses across studies become popular. A related principle on the analysis side is evidence synthesis, which is a way of combining information from several studies that have examined the same question to come to an overall understanding.

The increasing complexity of both the designs and the analyses of clinical trials has the consequence that the planning of such studies often relies also on computer simulations, especially if no closed form solutions for calculations are available. This applies even to usually fairly straightforward tasks as sample size calculations. Various frameworks and guidance proposals have been developed over the past years on how to plan, execute, and interpret the results of such simulation studies. Method comparison studies are used to give clinicians and biostatisticians guidance when an approach improves currently used methods (Hanneman, 2008). Depending on the setting and the purpose of the simulation study, suitable metrics must be chosen to compare alternative designs or analysis strategies. Especially relevant are neutral comparison studies, which are characterized by Boulesteix et al. (2013) as follows: The primary goal of the respective article is not to introduce a new promising method, the authors are reasonably neutral, and the evaluation metrics, methods, and datasets are chosen in a rational way. A comprehensive framework for such an objective assessment of competing strategies for clinical trial designs is CSE (clinical scenario evaluation) (Benda et al., 2010; Friede et al., 2010). CSE was particularly developed to support the overall process of performing simulations of clinical trials, especially for complex designs and analysis strategies. The problem is decomposed into data models, analysis models, and evaluation models that specify the evaluation metrics.

Here, we propose to use a formal approach to optimization of the clinical trial design, in the situation of a very large number of available options for the parameters specifying the design. This set of candidate designs can be regarded as a multidimensional search space, including potentially many real-valued and categorical parameters. The vast space of possible designs makes the optimization computationally demanding. In the following, we demonstrate that the so-called *Bayesian optimization* (BO) framework (Jones, 2001) can be successfully applied to select a clinical trial design. As evaluation metric, we consider the specific case that the statistical power of the clinical trial design should be maximized, for given effect sizes and a fixed total sample size. Further, it is crucial that such a design is found in a short time.

BO, often also called model-based optimization, has been successfully applied to optimize expensive black boxes (i.e., functions that take long to evaluate and are not given in closed form) in many scenarios. Popular applications include hyperparameter tuning of machine learning methods (Snoek et al., 2012), general algorithm configuration (Hutter et al., 2011), and many more. Various implementations of BO methods are widely used, such as sequential model-based algorithm configuration (SMAC) (Hutter et al., 2011), Spearmint (Snoek et al., 2012), and BoTorch (Balandat et al., 2020). For this work, we use the implementation in the R-package `mLrMBO` (Bischl et al., 2017). This implementation has been successfully used to optimize hyperparameters of machine learning methods on various tasks (Bischl et al., 2017; Wozniak et al., 2018), and in the biomedical context it has been applied to optimize model weights (Browaeys et al., 2020; Richter et al., 2019). Also, BO is well suited for multiobjective optimization. Different adaptations of BO exist (Horn et al., 2015) that return a set of nondominated points. Many of them are implemented in `mLrMBO` as well. However, in this work, we restrict ourselves to single-objective optimization. Independently of this work, for instance, Wilson et al. (2021) successfully applied multiobjective BO to minimize the number of participants and the number of clusters in a cluster randomized controlled trial under the restriction that the power is above a given threshold.

The remainder of the paper is structured as follows. In the next section, we present as motivating example an application of adaptive seamless designs for chronic obstructive pulmonary disease (COPD). In Sections 3 and 4, we explain the background on adaptive seamless designs (asd) and BO, respectively. In Section 5, we present a simulation study that is closely related to the motivating example and in which the suitability of BO for optimizing trial designs is demonstrated. In the last Section 6, we discuss promising extensions as well as limitations of the approach.

## 2 | MOTIVATING EXAMPLE

The general idea of this paper is to apply BO for finding a trial design that is optimal with respect to a certain metric. As an example, we consider the maximization of the statistical power of a trial design, given a set of parameters to be chosen for the trial design. We compare the computation time as well as the maximal obtained power between the results obtained with BO and with an exhaustive grid search over the space of adjustable parameters. The presented optimization approach

can also be applied for optimizing other metrics than the power, as long as the metric itself is numeric and the parameters of the trial design are numeric and include none or just few categorical choices.

One specific trial design that falls into this category is *adaptive seamless designs* (Barnes et al., 2010) as implemented in the *asd* package (Parsons et al., 2012) and described in Section 3. In Friede et al. (2020), the authors present an application of *adaptive seamless designs* for chronic obstructive pulmonary disease (COPD). We use this COPD study to apply BO for finding a trial design with maximal statistical power, given the parameters of the respective design options.

For the COPD trial, an interim treatment selection was performed. Indacaterol is a drug administered to COPD patients. The patients were randomized to four doses of indacaterol (75, 150, 300, and 600  $\mu\text{g}$ ), to active controls, and to a placebo control. Here, as in Friede et al. (2020), we ignore the active controls. The primary outcome of the trial was the percentage of days of poor control over 26 weeks of the COPD patients. Since the recruitment took only a short time, another outcome was required for treatment selection, in this case forced expiratory volume in 1 s (FEV1) at 15 days. In the original study, the difference in FEV1 compared to placebo at 15 days for the different doses and corresponding confidence intervals was estimated. From these, standardized effect sizes were estimated as approximately 0.68, 0.82, 0.95, and 0.91 for the four indacaterol doses 75, 150, 300, and 600  $\mu\text{g}$ , respectively. For the final outcome days of poor control, also, from real data the standardized effect sizes were estimated. These were approximately 0.13, 0.17, 0.23, and 0.20, for the four doses. The approximate sample sizes per arm were 100 patients in stage 1 and 300 patients in stage 2.

In the analysis of Friede et al. (2020), the aim was to select two doses of indacaterol at the interim analysis that are passed to stage 2, resulting in a total sample size of  $5 \cdot 100 + 3 \cdot 300 = 1400$  patients. In the corresponding simulations, a positive correlation between early and final outcomes of 0.4 was assumed. Further, different settings with continuous early and continuous or binary final outcomes and different treatment selection strategies were compared.

### 3 | ADAPTIVE SEAMLESS DESIGNS

Traditionally, drug development is organized in four phases with individual clinical trials for the separate phases. Seamless designs combine elements from different development phases in a single trial, thereby offering the promise of speeding the development process. For instance, seamless phase II/III designs combine the learning about the dose–response relationship or subgroup heterogeneity typical for phase II with the confirmation of treatment effects in phase III. If the data of the phase II part are used for confirmatory purposes, then this needs to be accounted for in the testing strategy to maintain the family-wise type I error rate in the strong sense. These types of designs are called adaptive seamless designs. A recent overview can be found in Friede et al. (2020).

We start by providing some more details on the study design considered (Sections 3.1 and 3.2). Then, in Section 3.3 we briefly introduce a framework for optimizing such designs, namely clinical scenario evaluation (CSE).

#### 3.1 | Two-stage design with treatment selection

We consider a particular type of adaptive seamless designs, namely two-stage designs with treatment (or more specifically dose) selection. The trial starts in stage 1 with  $K$  experimental treatments or doses of one experimental treatment which are compared to a common control. Interest is in assessing the treatment effects  $\beta_k$  with  $k = 1, \dots, K$  comparing experimental treatment  $k$  versus control and testing the null hypotheses  $H_{0,k} : \beta_k = 0$  against the alternative  $H_{1,k} : \beta_k > 0$  (larger  $\beta_k$  being better). For the purpose of treatment selection, test statistics  $Z_{1,k}$  based on data from the first design stage are considered, with larger  $Z_{1,k}$  indicating more beneficial treatments. These can use data on the primary outcome used for testing or any other outcome that is available at the time of the interim analysis. In practical applications, early outcomes are often used to inform interim decisions, since the primary endpoint might take a longer time to observe. In the interim analysis, treatments are selected for continuation into the second stage; the set of selected treatments is denoted by  $S$ . The selection rules used here are described in Section 3.2. In the final analysis, the closed test principle is applied to account for the multiple hypotheses. The intersection hypotheses are tested by combining stagewise  $p$ -values for the respective intersection hypotheses using a prespecified combination function. Here we use the inverse normal combination function (Lehmacher & Wassmer, 1999), which is given by

$$w_1 \Phi^{-1}(1 - p_{\mathcal{K},1}) + w_2 \Phi^{-1}(1 - p_{\mathcal{K},2}) \quad (1)$$

where  $\Phi^{-1}(\cdot)$  denotes the quantile function of the standard normal distribution,  $w_1$  and  $w_2$  weights with  $w_1^2 + w_2^2 = 1$ , and  $p_{\mathcal{K},1}$  and  $p_{\mathcal{K},2}$  stagewise  $p$ -values testing intersection hypothesis  $H_{\mathcal{K}}$ . The  $p$ -values for stages 1 and 2 are calculated on those patients recruited in stages 1 and 2, respectively. Note that one might not be able to calculate  $p_{\mathcal{K},1}$  at the time of the interim analysis since some patients might still be under follow-up for the final endpoint.

### 3.2 | Selection rules

In practice, the totality of the data would be considered by a data monitoring committee (DMC) to make a recommendation regarding the treatment selection. However, it is advisable for the sponsor to define a selection rule, although in comparison fairly simplistic, to provide some guidance for the DMC regarding the sponsor's preferences. Furthermore, formal selection rules are indispensable to conduct simulation studies, for example, to explore sample size and power. In the following, we give some examples of selection rules which have previously been considered and which we will feature below.

The treatment selection might consider statistics  $Z_{1,k}$  for the final (primary) outcome or statistics  $Z_{1,k}^{(E)}$  calculated based on some early outcome. We present the selection rules using  $Z_{1,k}^{(E)}$ , but of course these would be replaced by  $Z_{1,k}$  if the final outcome was used for treatment selection. Again, larger statistics are considered better.

#### *$\kappa$ -best rule*

Let  $Z_{1,(1)}^{(E)}, \dots, Z_{1,(K)}^{(E)}$  denote the ordered statistics with  $Z_{1,(1)}^{(E)}$  being the largest and therefore best test statistic. The  $\kappa$ -best rule then selects the treatments associated with the  $\kappa$  largest values of the test statistic, that is,  $Z_{1,(1)}^{(E)}, \dots, Z_{1,(\kappa)}^{(E)}$ .

#### *Epsilon rule*

To our knowledge, this rule was first described by Kelly et al. (2005) and has since then featured in a number of simulation studies including Friede and Stallard (2008) and Friede et al. (2020). All treatments  $k$  with  $Z_{1,k}^{(E)} \geq \max_i Z_{1,i}^{(E)} - \epsilon$  with  $\epsilon \geq 0$  are selected. For  $\epsilon = 0$ , this rule is equivalent to the  $\kappa$ -best rule with  $\kappa = 1$ .

#### *Threshold rule*

With this rule, all treatments  $k$  are selected with  $Z_{1,k}^{(E)} \geq \tau$ , where  $\tau$  is a given threshold.

We note that with the  $\kappa$ -best rule a fixed number of treatments is carried forward into the next stage, whereas these are variable with the epsilon rule and the threshold rule.

### 3.3 | Clinical scenario evaluation for ASD

The clinical scenario evaluation framework was first proposed by Benda et al. (2010) and subsequently further developed by Friede et al. (2010). Figure 1 provides a summary of this framework. A clinical scenario evaluation is defined by the following three elements: (i) disease-specific features, (ii) design options, and (iii) design performance measures. The disease-specific features describe distributional assumptions on the early and final outcomes including variances, correlations, and treatment effect sizes. Here, some can be estimated from data of previous clinical trials, and some are simply unknown. Considering adaptive seamless designs, the design options include the time point of the interim analysis and the treatment selection rule, which can in principle both be chosen without any restrictions. In contrast, some other design options such as the number of treatments and the total sample size are constrained by the environment and infrastructure available. The performance measures (sometimes also referred to as metrics) in the context with adaptive seamless designs are, for example, power, sample size distributions, or duration of the study.

The evaluation of some performance measures requires Monte Carlo simulations, since no closed-form solutions are available. The R-package `asd` by Parsons et al. (2012) can be used for such simulations. As it generates test statistics, following multivariate normal distributions, it is applicable to a wide range of outcomes, at least approximately, and generally more efficient in comparison to simulating individual participant data. The simulation model was described in more detail by Friede et al. (2020). It features selection at interim based on early as well as final outcomes.

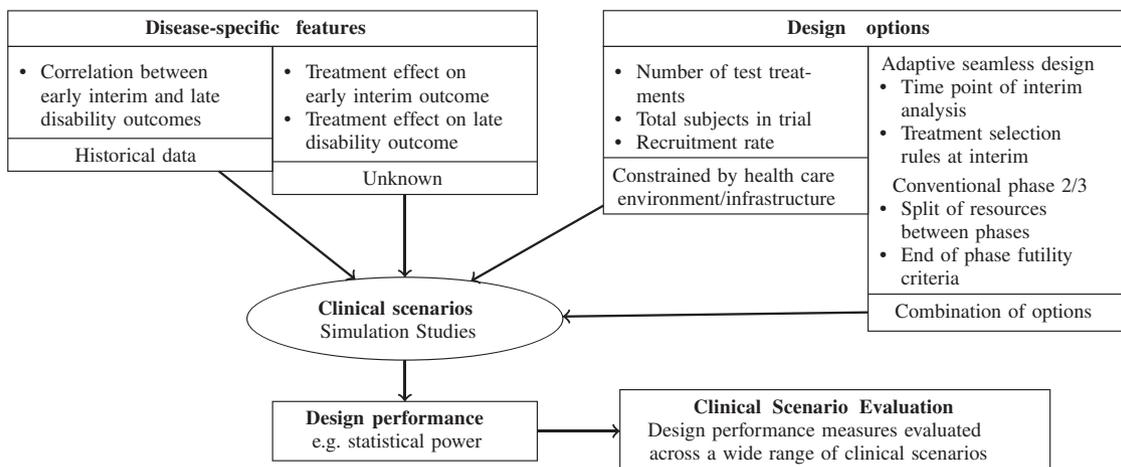


FIGURE 1 Application of the refined clinical scenario evaluation framework to adaptive seamless designs in progressive multiple sclerosis (the figure is adapted from Figure 2 in Friede et al., 2010)

## 4 | BAYESIAN OPTIMIZATION

BO (Jones, 2001), also known as model-based optimization (MBO), is a state-of-the-art technique for expensive black-box optimization problems (Shahriari et al., 2016). In comparison to other black-box optimization methods, like genetic algorithms or simulated annealing, BO is especially suitable when evaluating a configuration (e.g., running a simulation with certain parameters, here denoted by  $\theta$ ) is very time consuming. In this situation, it becomes infeasible to evaluate the black box for thousands of configurations.

### 4.1 | General principle

BO solves an optimization problem within a bounded search space  $\Theta$ :

$$\theta^* := \arg \max_{\theta \in \Theta} f(\theta), \quad (2)$$

where  $f(\theta)$  denotes the evaluation of the black box with the input configuration  $\theta$ . To reduce the number of evaluations on  $f$  the key idea of BO is to only evaluate values of  $\theta$  that are estimated to lead to a high value of  $f(\theta)$ . The estimate is generated by a so-called *surrogate model*. Typically, this is a regression model that predicts the outcome of  $f$  based on previous evaluations of  $f$ . First, an initial design of already evaluated configurations is needed. Then, iteratively, the BO algorithm fits the surrogate on the previous evaluations and proposes a new configuration  $\theta$  and evaluates it on  $f$ . The steps are repeated until a budget is exhausted.

The proposal of a new configuration  $\theta$  is obtained by maximizing a so-called acquisition function. The acquisition function guides the search to promising new regions in the search space  $\Theta$  by using the mean prediction  $\hat{\mu}(\theta)$  as well as the uncertainty quantification  $s^2(\theta)$  of the surrogate. It balances between exploration of not yet evaluated regions in  $\Theta$  and exploitation. The acquisition function achieves exploration by assigning high values to areas where the surrogate predicts a high uncertainty. Exploitation is achieved by assigning high values to areas of  $\theta$  where the surrogate predicts a high outcome of  $f$ . For deterministic functions, the expected improvement (Jones et al., 1998) is arguably the most popular acquisition function.

The expected improvement as well as other acquisition functions are derived under the assumption that at each point  $\theta \in \Theta$  the posterior of the function outcome follows a normal distribution. The standard Gaussian process regression, that is often used as surrogate, models the posterior as a Gaussian with the parameters  $\hat{\mu}$  and  $s^2(\theta)$  under the assumption that  $f$  is a realization of a Gaussian process. In practice, this assumption often cannot be verified. However, benchmarks show that the approach nevertheless works reliably for many practical problems (Bischi et al., 2017; Snoek et al., 2012).

## 4.2 | Acquisition function for noisy outcomes

Since this work focuses on optimization of a stochastic simulation, we assume that our nondeterministic optimization problem can be formulated as follows:

$$\theta^* := \arg \max_{\theta \in \Theta} f(\theta) + \varepsilon, \text{ with } \varepsilon \sim N(0, \sigma^2). \tag{3}$$

In this setting, the expected improvement is not feasible, as shown in Huang et al. (2006). They propose to use the augmented expected improvement (AEI) instead. Therefore, in a first step, we calculate the *effective best solution*:

$$\theta^{**} := \arg \max_{\theta \in \mathcal{D}} \hat{\mu}(\theta) - c \cdot s(\theta), \tag{4}$$

with  $c$  as a tuning parameter that is usually set to 1 and  $\mathcal{D}$  as the design that contains all previously evaluated values of  $\theta$ . The effective best point is the pessimistic estimate of the best observed outcome so far. In the final step, we calculate the augmented expected improvement (AEI) (Huang et al., 2006) as follows:

$$\text{AEI}(\theta) = (\hat{\mu}(\theta) - \hat{\mu}(\theta^{**})) \cdot \Phi\left(\frac{\hat{\mu}(\theta) - \hat{\mu}(\theta^{**})}{s(\theta)}\right) + s(\theta) \cdot \phi\left(\frac{\hat{\mu}(\theta) - \hat{\mu}(\theta^{**})}{s(\theta)}\right) \cdot \underbrace{\left(1 - \frac{\sigma}{\sqrt{\sigma^2 + s^2(\theta)}}\right)}_{\text{Correction}}, \tag{5}$$

where  $\Phi$  and  $\phi$  are the distribution and density function of the standard normal distribution, and  $\sigma^2$  denotes the random error (nugget effect) in the Kriging model. The formula is derived under the assumption of a Gaussian posterior at each point  $\theta$ . The intuition in formula (5) is that the first summand favors exploitation while the second summand favors exploration. The correction term is necessary to avoid exploration in areas where the model uncertainty  $s^2(\theta)$  equals the random error  $\sigma^2$ , because in this case further evaluations will not decrease the model uncertainty.

## 4.3 | Selection of the best point from noisy outcomes

If we chose the configuration  $\theta$  from all previously evaluated configurations according to the best outcome of  $f$  as the optimization result, we would be overly optimistic. It is likely that a single best outcome can be partially attributed to the random error  $\sigma^2$ . Instead, we are interested in the optimum of the true posterior mean. Therefore, we employ the surrogate estimate of the posterior mean for each evaluated configuration to cancel out the noise. The configuration for which the surrogate estimates the best outcome is then returned as the optimization result  $\hat{\theta}^*$ :

$$\hat{\theta}^* := \arg \max_{\theta \in \mathcal{D}} \hat{\mu}(\theta). \tag{6}$$

Using the stochastic outcome  $f(\hat{\theta}^*)$  observed during the optimization process would still potentially lead to an overly optimistic result. Therefore, an independent calculation of  $f(\hat{\theta}^*)$  should be conducted to obtain a fair estimate of the true value of the best outcome.

## 5 | SIMULATION STUDY

The goal of this simulation study is to investigate the potential benefit of BO for efficiently finding a clinical trial design that maximizes a given metric. The analysis is motivated by the clinical example with COPD patients introduced in Section 2, with treatment selection in the first of two stages. In the first stage, five arms (defined by the control and four different administered doses) are considered, and in the second stage only a selected subset of arms is kept for further evaluation. The metric to be optimized is the statistical power, where a positive outcome means a trial arm with true effect is detected as the significant effect. Depending on the application the power can be calculated as the proportion of correctly rejecting any, all, or a subset of the elementary hypotheses (Senn & Bretz, 2007).

TABLE 1 Effect sizes used for simulation in the first (early) and second (final) stage

Effect set	Stage	Treatment				
		0	1	2	3	4
paper	Early	0	0.680	0.82	0.950	0.91
	Final	0	0.130	0.17	0.230	0.20
linear	Early	0	0.200	0.40	0.600	0.80
	Final	0	0.050	0.10	0.150	0.20
sigmoid	Early	0	0.100	0.20	0.700	0.80
	Final	0	0.025	0.05	0.175	0.20
paper2	Early	0	0.680	0.82	0.950	0.91
	Final	0	0.260	0.34	0.460	0.40

TABLE 2 Overview of selection strategies to determine which trial arms are kept for the second stage

Notation	Procedure of selection strategy
$\kappa$ -best	Select $\kappa = 1, 2, 3$ or 4 (all) arms, respectively, according to maximal test statistic
eps	Select all arms with test statistic less than $\epsilon$ (epsilon) below maximal test statistic
thresh	Select all arms with test statistic larger than the value $\tau$ (threshold)

TABLE 3 Parameters for the selection strategies representing the search space of trial designs

Parameter	Range
Selection strategy	{1-best, 2-best, 3-best, all, eps, thresh}
$r$ (ratio)	(0,1)
$\epsilon$ (only for eps)	[0,4]
$\tau$ (only for thresh)	[0,10]

## 5.1 | Clinical trial design and parameters

For the true effect sizes in the simulation, we consider four *effect sets* whose values are inspired and partially based on the corresponding numbers of the motivating example (see Table 1). For the effect set *paper*, we use exactly the same numbers as in Friede et al. (2020). For the effect set *paper2*, the effect sizes for the second stage are doubled. As a further case, we consider a *linear* effect set with effect sizes linear increasing from 20% to 80% for the first stage and effect sizes divided by 4 for the second stage as well as a *sigmoid* relationship, again with four times larger effect sizes for the first stage. These sets reflect different realistic situations. Sometimes, there are delayed treatment effects (e.g., in the immunotherapy clinical trial setting) where the early effects are smaller than the delayed final effects. To include this scenario, we added a further effect set, where the numbers between the early and the final stage of the scenario *paper* are flipped (see the Online Appendix).

The strategies for treatment selection for the second stage in the two-phase trial design and their corresponding parameters that are used in the simulation study are given in Table 2 and Table 3, respectively. Choices for the boundaries of  $\epsilon$  and  $\tau$  are obtained from preliminary studies, and based on the fact that these are thresholds to the absolute amount of normally distributed random variables. Also, Figure S4 in the Online Appendix shows that the optimum is within the given boundaries. The strategies follow the concept of adaptive seamless designs discussed in Section 3. The  $\kappa$ -best strategies select the 1, 2, 3, or 4 best strategies according to their respective test statistics in the first stage (with *all* representing all  $k = 4$  treatments besides the control). The strategies *eps* and *thresh* represent strategies with a flexible number of treatments taken to the second stage. In Table 3, the ranges for these two parameters and for the proportion of the total samples (ratio  $r$ ) allocated to the second stage are given.

## 5.2 | Calculation of the statistical power

The simulation of the statistical power of the trial design can be formulated as a black-box function as follows:

$$y = f(n_{\text{stage1}}, n_{\text{stage2}}, \text{selection strategy}, \epsilon, \tau, \text{effect set}), \tag{7}$$

whereas  $n_{\text{stage1}}$  and  $n_{\text{stage2}}$  be the numbers of patients (samples) allocated to each treatment in stage 1 and stage 2 of the trial, respectively and *effect set* remains fixed as we are interested in the optimal trial design for a given *effect set*. Our main optimization goal is to maximize the statistical power  $y$  depending on the trial design.

However, we are interested in keeping the total sample size  $n_{\text{total}}$  constant in the overall trial for a fair comparison of the power of different designs. Therefore, we introduce the ratio of patients (samples) per treatment in stage 1, compared to patients per treatment in both stages

$$r = \frac{n_{\text{stage1}}}{n_{\text{stage1}} + n_{\text{stage2}}} \tag{8}$$

as a new parameter that will replace  $n_{\text{stage1}}$  and  $n_{\text{stage2}}$  in (7).

In the following, we will explain how we derive the values of  $n_{\text{stage1}}$  and  $n_{\text{stage2}}$  in dependence of  $r$  and  $n_{\text{total}}$  so that we can maximize

$$y = \tilde{f}(r, \text{selection strategy}, \epsilon, \tau, \text{effect set}, n_{\text{total}}) \tag{9}$$

instead of (7), whereas  $n_{\text{total}}$  and *effect set* remain fixed.

Let  $k_1$  and  $k_2$  denote the number of treatments in stage 1 and stage 2, respectively. Within one stage, the same number of patients is allocated to all different treatments. Then  $n_{\text{total}}$  is given by

$$n_{\text{total}} = k_1 \cdot n_{\text{stage1}} + k_2 \cdot n_{\text{stage2}}. \tag{10}$$

In our scenario (as described in Section 5.1), we have  $k_1 = 5$  and  $k_2 \in \{2, \dots, 5\}$ , where  $k_2$  depends on the result of the treatment selection strategy in stage 1. The control treatment is included in both stages.

Given the number of treatments  $k_1$  and  $k_2$  in the two stages and the total sample size  $n_{\text{total}}$ , optimizing the power with respect to  $n_{\text{stage1}}$  and  $n_{\text{stage2}}$  is equivalent to optimizing with respect to  $r$ , since

$$\begin{aligned} n_{\text{stage1}} &= n_{\text{stage1}} \cdot \frac{n_{\text{total}}}{k_1 \cdot n_{\text{stage1}} + k_2 \cdot n_{\text{stage2}}} = n_{\text{total}} \cdot \frac{n_{\text{stage1}}}{k_1 \cdot n_{\text{stage1}} + k_2 \cdot n_{\text{stage2}}} \\ &= n_{\text{total}} \cdot \frac{r}{k_1 \cdot r + k_2 \cdot (1 - r)} \text{ and} \\ n_{\text{stage2}} &= n_{\text{total}} \cdot \frac{1 - r}{k_1 \cdot r + k_2 \cdot (1 - r)}. \end{aligned} \tag{11}$$

These formulas give the numbers of patients (samples) per treatment arm in the two stages, if  $k_2$  is known. This is only the case for selection strategies 1, 2, and 3.

For the selection strategies *eps* and *thresh*,  $k_2$  depends on the outcome of stage 1. In these cases, for given values of  $n_{\text{stage1}}$  and  $\epsilon$  or  $\tau$ , respectively, we can obtain an estimate  $\hat{k}_2$  of  $k_2$  with a so-called calibration step. For a set of values for  $n_{\text{stage1}}$  in the interval  $[\lceil 0.01 \cdot \frac{n_{\text{total}}}{k_1} \rceil, \lceil \frac{n_{\text{total}}}{k_1} \rceil]$ , we simulate stage 1 of the trial multiple times. We calculate  $n_{\text{stage2}} = (1 - r)/r \cdot n_{\text{stage1}}$  from formula (8) and  $\hat{k}_2$  as the average of the resulting values of  $k_2$  for all simulations with a fixed value for  $n_{\text{stage1}}$ . Then we select the value of  $n_{\text{stage1}}$  such that the corresponding total sample size is closest to  $n_{\text{total}}$ . More precisely, we minimize

$$h(n_{\text{stage1}}) := \left( \left( k_1 \cdot n_{\text{stage1}} + \hat{k}_2 \cdot \frac{1 - r}{r} \cdot n_{\text{stage1}} \right) - n_{\text{total}} \right)^2 \tag{12}$$

with respect to  $n_{\text{stage1}}$ , where the estimate  $\hat{k}_2$  is obtained as described above dependent on  $n_{\text{stage1}}$ . The calibration step is necessary to guarantee that the total sample size  $n_{\text{total}}$  is (on average) the same, also for the variable selection strategies *eps* and *thresh*.

The above procedure gives us the values for  $n_{\text{stage1}}$  and  $n_{\text{stage2}}$ , allowing us in the following to find the optimal clinical trial design for a given *effect set* and a fixed total sample size, by optimizing the parameter values for  $r$ , *Selection strategy*,  $\epsilon$  and  $\tau$  through maximizing  $\hat{f}$  in (9).

### 5.3 | Optimization strategies

In the simulation study, we compare the ability of two optimization approaches for finding an optimal parameter configuration, on the one hand an exhaustive grid search on the parameter space and on the other hand BO. For different scenarios, we compare the statistical power of the best found clinical trial design. We also compare the runtime needed to determine the respective optimization results. Next we introduce the two optimization strategies.

#### *Grid and Grid small*

We conduct an *exhaustive grid search* with a resolution of  $l$  points per real-valued dimension and one point for each categorical dimension. Referring to our application, for each selection strategy we evaluate a numerical grid with a resolution of  $l$ . All six selection strategies have the numeric parameter  $r$ . The strategies *epsilon* and *thresh* both have an additional second parameter,  $\epsilon$  or  $\tau$ , respectively. For  $l = 25$ , we call this approach *Grid*, with a total of  $4 \cdot 25 + 2 \cdot 25 \cdot 25 = 1350$  design configurations that are evaluated for one scenario. We chose  $l = 25$  generously to obtain an estimate of the optimum that should be close to the true optimum. Previous studies used half the resolution to assess the behavior of similar problems (Friede et al., 2020). We define *Grid Small* to be a subset of this grid with  $l = 7$  points per dimension, resulting in  $4 \cdot 7 + 2 \cdot 7 \cdot 7 = 126$ . This results in a number of evaluations that can be easily carried out in practice.

Both grids are evaluated with 20 stochastic replications to estimate the variance of the estimated power.

The best point per evaluated grid is determined by selecting the configuration with the best outcome, that is the concrete trial design with the largest power. The corresponding value is generally too optimistic, as the random noise on the estimated value can contribute to the quality of the outcome. To account for this optimism and report unbiased performance estimates, for each of the 20 replicate simulations, we identify the best configuration and consider the corresponding 19 outcomes of the other 19 replicate simulations, similar to a cross-validation framework in machine learning applications.

#### *BO (model-based optimization) and BO Grid*

For applying BO, we first define how the surrogate model that predicts the outcome of  $f$  for unknown values of  $\theta$  (parameter configurations) is generated. As regression model for the surrogate we choose Kriging (also called Gaussian process regression) and use the implementation in the R-package `DiceKriging` (Roustant et al., 2012), because it is known to work well for fairly low dimensional search spaces. We configure Kriging to apply the Matérn  $\frac{5}{2}$  kernel with an estimated *nugget effect* to account for the noisy response of  $f$  and without scaling the input variables to  $[0,1]$ . Note, that by doing so, we implicitly assume that  $f$  is a realization of a Gaussian process and that each single outcome is a realization of a Gaussian. However, as stated in Section 4, we can expect the method to work well even if the assumption is slightly violated.

Kriging expects a numerical input, but our search space contains categorical parameters and is even hierarchical. First, to transform the categorical parameters of the search space (see Table 3) into numerical values we use one-hot encoding. Furthermore, our search space is hierarchical because the parameters  $\epsilon$  and  $\tau$  are only active for the respective selection strategies *eps* and *thresh*. The hierarchical structure introduces inactive values for the configurations ( $\theta$ ), that is if the *selection strategy* is set to *eps* then the parameter  $\tau$  of *thresh* is inactive. If a value is inactive, we set it to a value two times as high as the maximum in its active range (e.g., if  $\tau$  is inactive it will be set to 20). This ensures that within the original range the Kriging model is not or only minimally affected by the inactive values. Note that this trick only works because we did not scale the input variables to  $[0,1]$  and because the estimated covariance of the Matérn  $\frac{5}{2}$  kernel is approximately zero for distances greater than 4. With the above steps, we ensure that the input for the Kriging is purely numeric.

As an acquisition function, we choose the AEI as explained in Section 4. We start BO with an initial design of 16 randomly sampled points, following an established rule-of-thumb of four points per dimension, which is also well in line with the recommendation of using around 10% of the budget for the initial design (Bossek et al., 2020). We allow 100 further iterations, summing up to a total budget of 116 evaluated configurations of the simulator  $\hat{f}$ .

TABLE 4 Parameter choices for the R function `treatsel.sim`

Meaning of parameter	Parameter	Value
Number of simulation iterations	<code>nsim</code>	1000
Correlation between early and final outcomes	<code>corr</code>	0.4
One-sided significance level	<code>level</code>	0.025
Vector of treatment numbers for determining power	<code>p<sub>test</sub></code>	(3,4)

At the end of the optimization procedure, the best configuration is chosen according to the best prediction obtained on the set of evaluated points as explained in Section 4.3. Again, overoptimism must be avoided. Thus, the best selected parameter configuration is then independently evaluated 20 times again.

The whole BO procedure is repeated 20 times to estimate the variance of the optimal solution found.

Additionally, we investigate how BO performs when it is limited to return configurations that are within the grid introduced previously. Therefore, we introduce *BO Grid*, which executes the optimization exactly as BO, but the best configuration is mapped to the closest value on the grid. In other words, for *BO Grid* we conduct the independent evaluation using the configuration within the grid that is closest to the optimal configuration that is proposed by BO. The comparison of the results of *BO Grid* and *Grid* will allow two possible conclusions: First, if *BO Grid* is worse than BO, we can assume that the resolution of *Grid* is not high enough to find an optimal configuration. Second, if *BO Grid* is better than *Grid*, it shows that *Grid* did not select the optimal configuration, due to noise. Summing up, there are two factors that can prevent *Grid* from finding the optimal configuration: lack of resolution and noise of the outcome.

### 5.4 | Implementation

The algorithms used in this study are implemented in R. For the BO, the R-package `m1rMBO` (Bischi et al., 2017) is used. For the simulation of clinical trials, the function `treatsel.sim` of the R-package `asd` (Parsons et al., 2012) is used, with parameter values given in Table 4. The vector of treatment numbers `ptest` determines the hypotheses for which rejections are counted towards the power. Here, the rejection of one or both of the hypotheses that treatment 3 or 4 has no effect against the control determines the power, as in Friede et al. (2020).

Note that the parameter `nsim` denotes the internal simulation iterations of  $f$ . The 20 additional stochastic repetitions we conduct for the *grid search* and BO are additional replicates.

### 5.5 | Main results for comparison of algorithms

In this section, we present an overview of the results for the comparison of the four optimization strategies: exhaustive *Grid* search, grid search with a reduced resolution (*Grid Small*), BO, and *BO Grid*. We focus our comparison on independent, replicated evaluations of the configurations proposed by each optimizer. The evaluations provide unbiased estimates of the performance ( $y_{\text{valid}}$ ) of all optimizers.

For this study we consider 12 scenarios. The scenarios are generated by combining the four different effect sets (see Table 1) with the three different numbers for  $n_{\text{total}}$ . Each optimizer is applied to each scenario with 20 stochastic repetitions.

Figure 2 shows the estimated power values  $y_{\text{valid}}$  for the best solutions in each scenario. The plots show that BO generally gives results that are comparable to those obtained with the (exhaustive) *Grid* method across all scenarios. Only in some cases, *Grid* has a slight advantage (*effect: linear*,  $n_{\text{total}} = 2000$ ; *effect: sigmoid*,  $n_{\text{total}} = 1000$ ; *effect: sigmoid*,  $n_{\text{total}} = 2000$ ). Below we will see that these scenarios have in common that the optimal configuration is close to the border of the search space. In other cases, BO yields slightly better results (*effect: paper*,  $n_{\text{total}} = 2000$ ; *effect: paper2*,  $n_{\text{total}} = 500$ ; *effect: sigmoid*,  $n_{\text{total}} = 500$ ). Across all scenarios, BO and *Grid* perform similarly; although due to its evaluation budget, BO is allowed only 116 evaluations as opposed to the 1350 evaluations for *Grid*, when using a resolution of 25 points per dimension. The difference in the number of evaluations is clearly reflected in the average runtime of the optimization methods given in Table 5. Runtimes were measured on a single core of a compute node with 2 Intel Xeon E5-2697v2 (2.70 GHz) with 12 Cores each and 512 GB of RAM in a nonexclusive usage of the node. BO is approximately 20 times faster.

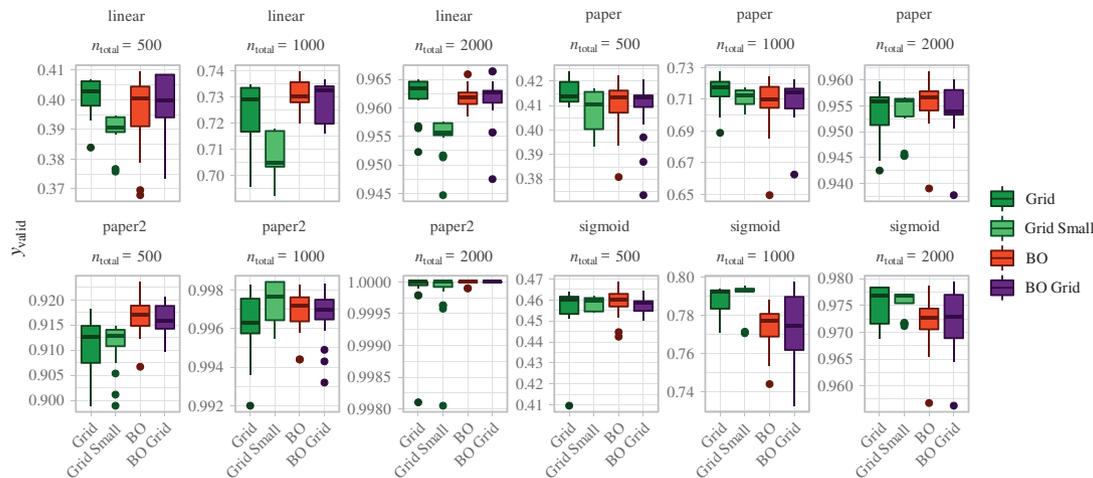


FIGURE 2 Box plots of performance (power) values on independent replicates of the best solutions for the four optimizers, the different effect sets, and different numbers of  $n_{\text{total}}$  (500, 1000, 2000)

TABLE 5 Average runtime in hours, for evaluating one grid and one optimization run of MBO

	linear			paper			paper2			sigmoid			evals
	500	1000	2000	500	1000	2000	500	1000	2000	500	1000	2000	
Grid	51.2	55.3	56.3	56.7	58.7	63.0	55.1	59.2	61.6	50.7	51.3	56.8	1350
Grid Small	3.8	4.1	4.1	4.1	4.2	4.6	4.1	4.3	4.4	3.7	3.8	4.1	126
BO	3.5	3.3	3.7	3.3	3.6	3.7	3.5	3.3	3.2	3.3	2.9	3.6	116

For *BO Grid*, only solutions on the grid that was evaluated by *Grid* are considered in the evaluation. Comparing the results of *BO Grid* to those of *BO*, we mainly see similar performances for both. This indicates that the resolution of *Grid* is fine enough to achieve similar performance to *BO*.

Comparing *BO Grid* to *Grid*, we observe that for the scenarios *effect: paper*,  $n_{\text{total}} = 500$  and *effect: paper*,  $n_{\text{total}} = 1000$ , *BO Grid* achieves slightly better performance than *Grid*. This suggests that the *Grid* optimization selects the best outcome “optimistically” from all observed outcomes and does not take into account the stochasticity of the problem, that is in this case the noise on the predicted power. Accordingly, we assume that for the mentioned scenarios, the noise was so high, that *Grid* was misguided by too optimistic outcomes when selecting the final configuration as optimization result. In our evaluation, the independent validation  $y_{\text{valid}}$  represents the unbiased performance of this final configuration, and the overoptimism becomes apparent. In such scenarios *BO* performs better, since the final configuration is determined based on the mean prediction of the surrogate of the BO, as explained in Section 4.3. Further below, we will study on a single scenario if a reduced noise due to a higher number of simulations  $n_{\text{sim}}$  decreases the effect of the overoptimistic selection of *Grid*.

In cases in which *BO* was not able to find a better solution than *Grid*, also *BO Grid* cannot yield better results as it just selects the configuration on the Grid that is closest to the configuration found by *BO*. Therefore, any increase in performance of *BO Grid* compared to *BO* is purely due to chance.

To compare *BO* with a grid search that takes approximately the same time (see Table 5) and number of evaluations, we consider *Grid Small*. Here, a resolution of seven points per dimension leads to 126 evaluations, close to 116 allowed evaluations of *BO*, with only a small advantage for *Grid Small*. Despite the 10 more evaluations allowed, in the majority of cases *BO* is superior to *Grid Small* (see Figure 2). The advantage is particularly apparent for all scenarios with effect set *effect: linear* and for scenario *effect: paper2*,  $n_{\text{total}} = 500$ . Only for the scenarios (*effect: sigmoid*,  $n_{\text{total}} = 1000$ ) and (*effect: sigmoid*,  $n_{\text{total}} = 2000$ ), *Grid Small* has a notable advantage over *BO* and *Grid*. In the second case, the absolute differences are negligible and all methods lead to a high power. Also, for this case we will later show that the optimal configuration is close to the border of the search space.

As stated above, the high resolution of the exhaustive *Grid* probably leads to the selection of slightly suboptimal configurations close to the optimum due to stochasticity. These points are not included in the more coarse grid used by *Grid*

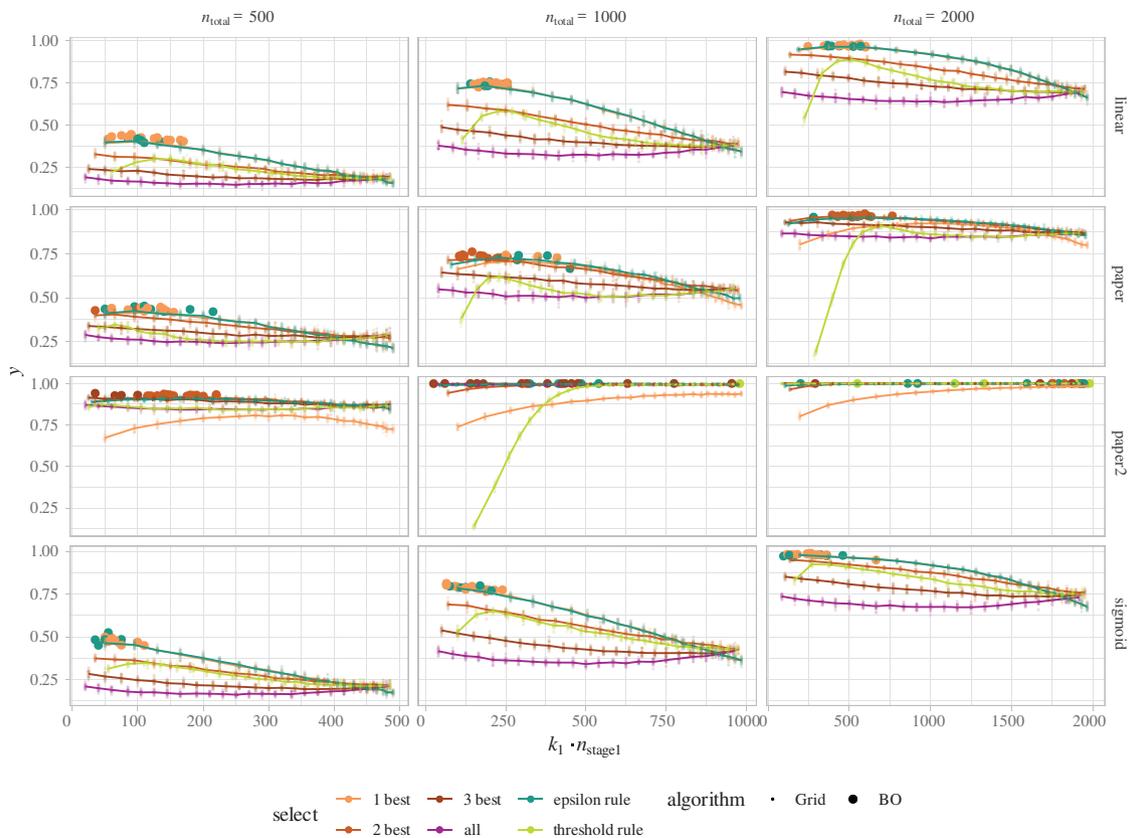


FIGURE 3 All evaluated configurations and their outcomes on the *Grid*. For the epsilon and threshold rule, only the curve for the  $\epsilon$  and  $\tau$  value with highest power ( $y$ ) is displayed. For each configuration in the grid, 20 results are obtained. A line connects the mean outcomes for each configuration. In addition, the best configurations identified from *BO* are displayed

*Small*. In other words, if a configuration close to the optimum is included in the grid used by *Grid Small*, then it is more likely to select it with less overoptimism, as the number of candidates in the region is smaller.

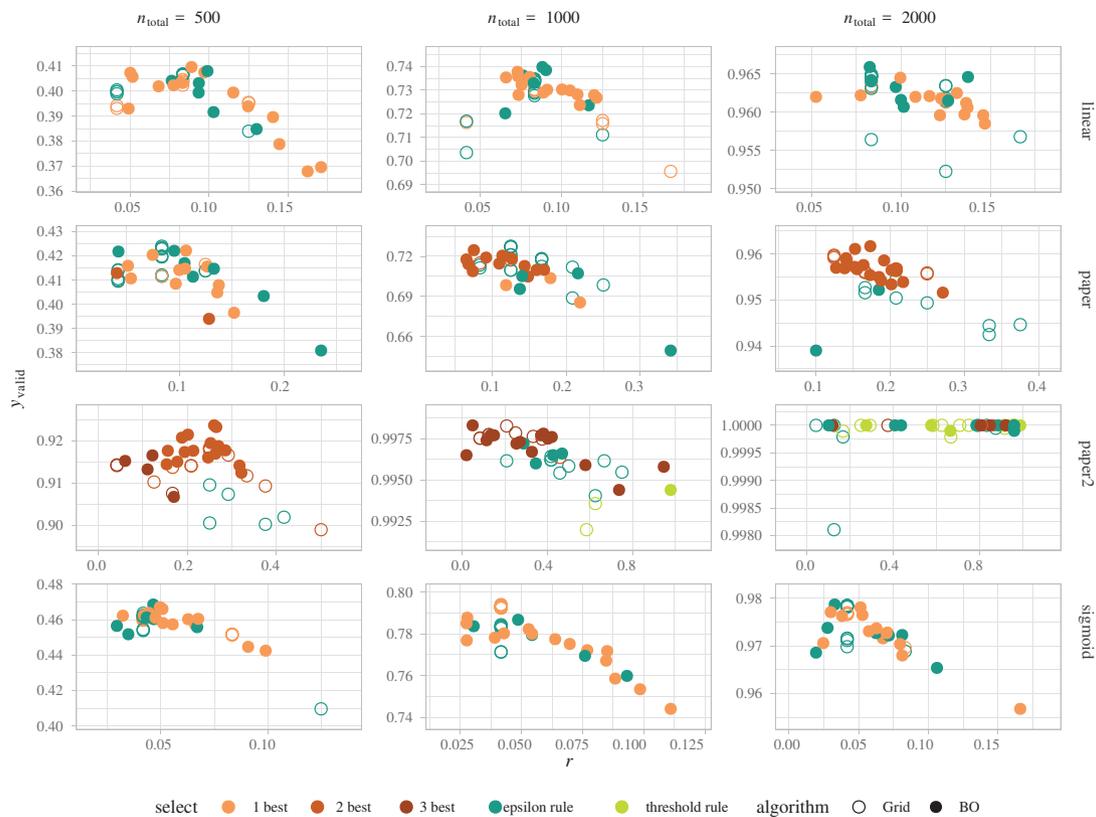
Note that between the scenarios the absolute differences of the optimization results are of different practical relevance. For example, for scenario *effect: sigmoid*,  $n_{total} = 2000$  *Grid* and *Grid Small* are superior to *BO*, but the difference is smaller than 0.5%.

### 5.6 | Detailed analysis of identified configurations

We now investigate the results in more detail. We analyze the differences depending on the *selection strategy* and the number of samples at stage 1, as determined by the *stage ratio* ( $r$ ). Figure 3 shows the observed outcomes for all evaluations on the grid and the final configurations found by *BO*. For the selection strategies *epsilon rule* and *threshold rule* only curves representing the  $\epsilon$  and  $\tau$  values with the best outcome, respectively, are shown. Note that for the effect sets *linear* and *sigmoid* the curves for selection strategy *1 best* and *eps* are nearly identical. As intuitively anticipated, in almost all scenarios the worst results were obtained if all available samples are used in stage 1 and none in stage 2 (i.e.,  $r = 1$ ).

In all scenarios, *BO* identified the peak of all combined curves. Within the 20 stochastic repetitions per scenario, *BO* found different configurations, but all had similar high power. The fact that no unique best setting was found across the repetitions can be attributed to the relatively high noise of the simulation and the relatively flat peaks, that is, different configurations yield similarly good outcomes. In cases in which multiple optimization runs result in different configurations, it might be advisable to let a human expert choose the final configuration, also depending on the medical background.

The curves also show the potential of optimization for the different scenarios. If the total sample size is high ( $n_{total} = 1000$ ), then nearly all design configurations yield high statistical power and optimization is less important. If the total sample size is too small ( $n_{total} = 500$ ), then no configuration yields a practically usable design. In this situation, a fast optimization helps to determine if a desired total sample size can actually yield sufficiently powerful designs. Of special



**FIGURE 4** Best configurations found with *Grid* and *BO* runs, that is corresponding  $y_{\text{valid}}$  values for the ratio  $r$  for the selected configurations. For the epsilon rule,  $\epsilon$  and for the threshold rule  $\tau$  are not displayed

interest are the scenarios where the correct choice of the design configuration can make a difference between a design with acceptable power and a design with too low power, for example, for the scenarios (*effect: linear*,  $n_{\text{total}} = 1000$ ; *effect: linear*,  $n_{\text{total}} = 2000$ , *effect: paper*,  $n_{\text{total}} = 1000$ ; *effect: sigmoid*,  $n_{\text{total}} = 1000$ ). Note, that the selection strategies *thresh* and *all* never selected powerful designs.

Figure 4 visualizes in detail which configurations were selected by *BO* and *Grid* and the corresponding values for  $y_{\text{valid}}$  (result of the independent validation). The optimal values for  $\epsilon$  and  $\theta$  are not shown. For some scenarios, most runs of *Grid* returned the same optimal configurations (e.g., for *effect: sigmoid*,  $n_{\text{total}} = 1000$ ). In such cases, the signal-to-noise ratio was sufficiently high to reliably find the best configuration. In some scenarios, *Grid* was superior to *BO*, as could be seen in the previously shown box plots in Figure 2. Furthermore, for all scenarios with *effect: sigmoid* and the scenarios *effect: linear*,  $n_{\text{total}} = 500$  and *effect: paper*,  $n_{\text{total}} = 500$ , small values of  $r$  lead to an optimal outcome. Here, *BO* was not able to find these values close to the border of the search space. A common remedy for this problem is to log-transform the values of  $r$ .

In other cases, *Grid* found different configurations in different runs, while *BO* lead to more stable results, also with higher power values (e.g., *effect: linear*,  $n_{\text{total}} = 1000$ ; *effect: paper2*,  $n_{\text{total}} = 500$ ; *effect: sigmoid*,  $n_{\text{total}} = 500$ ). Here we can assume, that *BO* successfully accounted for the noise. A special case is scenario *effect: paper2*,  $n_{\text{total}} = 2000$ : Here nearly all found configurations yield perfect outcomes.

## 5.7 | Results for more simulation rounds

To investigate the effect of the function noise variance ( $\sigma^2$ ) on the optimization result, we increase the simulation iterations from  $n_{\text{sim}} = 1000$  to  $n_{\text{sim}} = 5000$ , restricting to the scenario *effect: paper*,  $n_{\text{total}} = 2000$ . Additionally, we run the algorithms 100 times instead of the previously used 20, in order to obtain more precise estimates for the power values in the independent validation.

FIGURE 5 Optimization progress for *BO*: Best outcome found so far for each iteration, in comparison to the best outcome found in the grid search, for effect set *paper* and  $n_{total} = 2000$

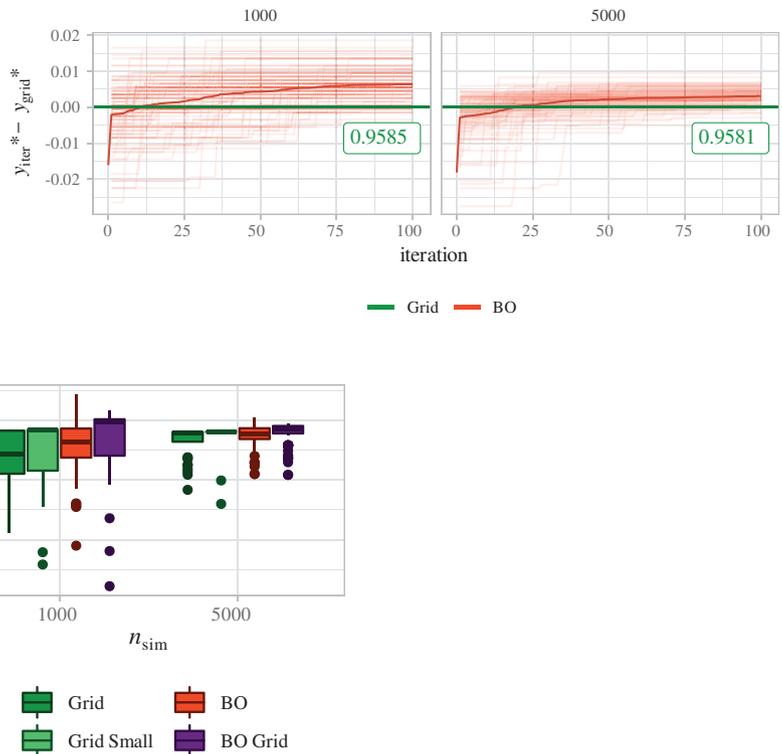


FIGURE 6 Validated performance of all optimization strategies on effect set *paper* and  $n_{total} = 2000$  with 100 stochastic repetitions and different values for  $n_{sim}$

We expect that increasing  $n_{sim}$  decreases the variance of the power values of the identified design configurations. For the exhaustive *Grid*, we expect that the decreased noise leads to fewer selections of overly optimistic outcomes. *BO* should profit as well from a less noisy objective function since the predictions of the surrogate function will be more precise.

Figure 5 shows that indeed the lower noise level leads to more stable simulation runs, meaning that single optimization runs lead to results closer to the mean result. Both curves indicate that more iterations of *BO* would not yield substantially better outcomes as no substantial improvements are observed after 80 iterations. It has to be noted that this plot reports the best observed outcome so far and not the one that is chosen as final result. Despite the differences in the optimization paths, the median validation error reported in Figure 6 is nearly identical for *BO* for the low and the high value of  $n_{sim}$ . Therefore, we can assume that the “bias correction” of *BO* as explained in Section 4.3 seems to be effective. For all methods, the IQR displayed in the box plots is much smaller for  $n_{sim} = 5000$ , indicating that the solutions that are found vary less between the 100 stochastic replications. Also, the absolute differences between the methods are even smaller now. This demonstrates that for the selected scenario all methods can find the optimal configuration if the results are not noisy. However, an increased value of  $n_{sim}$  also increases the runtime linearly, which conflicts with the goal to find the optimal configuration quickly. In summary, *BO* can find configurations with a small number of required simulations (higher noise) that are nearly as good as the optimal solution which is identified with a large number of simulations.

## 6 | DISCUSSION

*BO* is a state-of-the-art optimizer that can be used for parameter optimization in a wide range of applications. Here, we propose, for the first time, to optimize the power of adaptive seamless designs using *BO*. In clinical trials, determining the test methods and sample sizes is a key step in the planning process. For our optimization, we chose to maximize the test power, given a set of treatments with corresponding effect sizes and a fixed number of samples. If determining the test power involves intensive simulations, trying out different treatment design configurations can become a laborious task. Either large computational resources are required for an exhaustive investigation of all the different configurations, or a tedious manual trial-and-error process is required.

In this paper, we analyzed if *BO* can quickly find an adaptive seamless design with high power, more efficiently than an exhaustive grid search. For various sets of treatment effect sizes and total sample sizes, *BO* was able to find competitive

designs in a fraction of the time. In contrast to grid search, BO was able to account for stochasticity in the metric that should be optimized. For our optimization, we chose wide intervals for the parameters of the selection rules (i.e., epsilon and threshold rule) to show that BO is generally applicable. In practical applications, however, these values might sometimes be based on clinical rather than statistical considerations or the range of these values might be restricted by such considerations. Here we consider an optimization problem maximizing the power given a total sample size. If, for example, the task was to select all treatments with clinically relevant effect sizes, then clearly the selection parameter needs to be informed by clinical reasoning. In general, the application of BO to clinical trial design selection in general seems promising. Additional restrictions on the search space decrease its size and therefore promise even more reliable results. The same concept could be applied for subgroup selection in adaptive enrichment designs, which make use of interim data in selecting the target population for the remainder of the trial (Burnett & Jennison, 2020).

Only in cases in which the optimal values of one parameter are close to the lower border, we sometimes observed poor performance of BO. Here, a possible improvement can be to apply a log transformation to such parameters. We chose to maximize the power for a fixed number of samples. Other aspects like study duration or the cost of a trial in general could be of interest either as an additional constraint or an additional outcome. Also the number of samples can be included as an additional outcome instead of being fixed. Multiple outcomes that are subject to optimization give a (possibly restricted) multiobjective optimization problem. The result of which is a set of points (a Pareto front) that maximizes the statistical power and, for example, minimizes the total sample size at the same time. The user could then see for which sample size the power is above a certain threshold (e.g., 95%). Another way to speed up the optimization would be to iteratively increase the number of simulation rounds of the trial design evaluation. If few simulations already indicate a bad outcome for a configuration then additional simulation rounds are not necessary. Such a procedure can be guided by statistical testing and is already common in machine learning under the terms *racing* or *multifidelity optimization*.

In this work, we assumed the effect sizes to be fixed. In Stallard et al. (2009), the authors propose a Bayesian approach, where the effect sizes follow a prior, to calculate the expected trial performance. This requires a more expensive simulation making the optimization of a clinical trial design according to such metric an excellent candidate for BO.

## ACKNOWLEDGMENT

This work was partly supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, A3.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## OPEN RESEARCH BADGES



This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Jakob Richter  <https://orcid.org/0000-0003-4481-5554>

Tim Friede  <https://orcid.org/0000-0001-5347-7441>

## REFERENCES

- Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G. & Bakshy, E. (2020). Botorch: Programmable Bayesian optimization in PyTorch, *Advances in Neural Information Processing Systems* 33.
- Barnes, P. J., Pocock, S. J., Magnussen, H., Iqbal, A., Kramer, B., Higgins, M., & Lawrence, D. (2010). Integrating inhaled corticosteroid dose selection in a clinical study in COPD using an adaptive seamless design. *Pulmonary Pharmacology & Therapeutics*, 23, 165–171.
- Benda, N., Branson, M., Maurer, W., & Friede, T. (2010). Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal*, 44, 299–315.
- Bischi, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). Mlrnbo: A modular framework for model-based optimization of expensive black-box functions. [arXiv:1703.03373 [stat.ML]].
- Bogin, V. (2020). Master protocols: New directions in drug discovery. *Contemporary Clinical Trials Communications*, 18, 1–5.

- Bossek, J., Doerr, C., & Kerschke, P. (2020). Initial design strategies and their effects on sequential model-based optimization: An exploratory case study based on bbob. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference, GECCO '20* pp. 778–786). Association for Computing Machinery.
- Boulesteix, A.-L., Lauer, S., & Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLOS ONE*, *8*, 1–11.
- Browaeys, R., Saelens, W., & Saeyns, Y. (2020). Nichenet: Modeling intercellular communication by linking ligands to target genes. *Nature Methods*, *17*, 159–162.
- Burnett, T., & Jennison, C. (2020). Adaptive enrichment trials: What are the benefits? *Statistics in Medicine*, *40*(3), 690–711.
- Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ*, *321*, 694–696.
- Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N., Valdés-Márquez, E., & Chataway, J. (2010). Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Information Journal*, *44*, 713–718.
- Friede, T., & Stallard, N. (2008). A comparison of methods for adaptive treatment selection. *Biometrical Journal*, *50*, 767–781.
- Friede, T., Stallard, N., & Parsons, N. (2020). Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in r. *Biometrical Journal*, *62*, 1264–1283.
- Hanneman, S. (2008). Design, analysis, and interpretation of method-comparison studies. *AACN Advanced Critical Care*, *19*, 223–234.
- Horn, D., Wagner, T., Biermann, D., Weihs, C., & Bischl, B. (2015). Model-based multiobjective optimization: Taxonomy, multi-point proposal, toolbox and benchmark. In *International Conference on Evolutionary Multi-Criterion Optimization* (pp. 64–78). Springer.
- Huang, D., Allen, T. T., Notz, W. I., & Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential Kriging meta-models. *Journal of Global Optimization*, *34*, 441–466.
- Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In Coello, C. A. C. (Ed.), *Learning and intelligent optimization* (pp. 507–523). Lecture Notes in Computer Science, Vol. 6683. Springer, Berlin Heidelberg.
- Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, *21*, 345–383.
- Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*, 455–492.
- Kelly, P. J., Stallard, N., & Todd, S. (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, *15*, 641–658.
- Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, *55*, 1286–1290.
- Parsons, N., Friede, T., Todd, S., Marquez, E. V., Chataway, J., Nicholas, R., & Stallard, N. (2012). An r package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics & Data Analysis*, *56*, 1150–1160.
- Richter, J., Madjar, K., & Rahnenführer, J. (2019). Model-based optimization of subgroup weights for survival analysis. *Bioinformatics*, *35*, i484–i491.
- Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, diceOptim: Two r packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. *Journal of Statistical Software, Articles*, *51*, 1–55.
- Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, *6*, 161–170.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, *104*, 148–175.
- Sheiner, L. B. (1997). Learning versus confirming in clinical drug development. *Clinical Pharmacology & Therapeutics*, *61*, 275–291.
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q. (Eds.), *Advances in neural information processing systems 25* (pp. 2951–2959). Curran Associates.
- Stallard, N., Posch, M., Friede, T., Koenig, F., & Brannath, W. (2009). Optimal choice of the number of treatments to be included in a clinical trial. *Statistics in Medicine*, *28*, 1321–1338.
- Wilson, D. T., Hooper, R., Brown, J., Farrin, A. J., & Walwyn, R. E. (2021). Efficient and flexible simulation-based sample size determination for clinical trials with multiple design parameters. *Statistical Methods in Medical Research*, *30*, 799–815.
- Wozniak, J. M., Jain, R., Balaprakash, P., Ozik, J., Collier, N. T., Bauer, J., Xia, F., Brettin, T., Stevens, R., Mohd-Yusof, J., Cardona, C. G., Essen, B. V., & Baughman, M. (2018). Candle/supervisor: A workflow framework for machine learning applied to cancer research. *BMC Bioinformatics*, *19*, 491.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Richter, J., Friede, T., & Rahnenführer, J. (2022). Improving adaptive seamless designs through Bayesian optimization. *Biometrical Journal*, *64*, 948–963. <https://doi.org/10.1002/bimj.202000389>