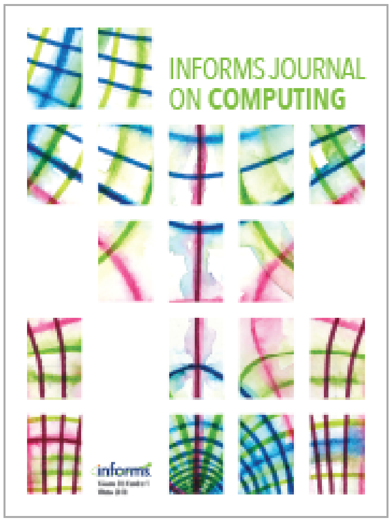


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Min-Max Optimal Design of Two-Armed Trials with Side Information

Qiong Zhang, Amin Khademi, Yongjia Song

informs.
http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Vol. 34, No. 1, January–February 2022, pp. 165–182
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Min-Max Optimal Design of Two-Armed Trials with Side Information

Qiong Zhang,[a] **Amin Khademi,**[b] **Yongjia Song**[b]

[a] School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina 29631; [b] Department of Industrial Engineering, Clemson University, Clemson, South Carolina 29631

Contact: qiongz@clemson.edu, https://orcid.org/0000-0003-1995-2127 (QZ); khademi@clemson.edu, https://orcid.org/0000-0002-5281-8715 (AK); yongjis@clemson.edu, https://orcid.org/0000-0001-6839-522X (YS)

Received: October 2, 2019
Revised: June 17, 2020; November 16, 2020; December 31, 2020
Accepted: January 13, 2021
Published Online in Articles in Advance: June 7, 2021

https://doi.org/10.1287/ijoc.2021.1068

Abstract. In this work, we study the optimal design of two-armed clinical trials to maximize the accuracy of parameter estimation in a statistical model, where the interaction between patient covariates and treatment are explicitly incorporated to enable precision medication decisions. Such a modeling extension leads to significant complexities for the produced optimization problems because they include optimization over design and covariates concurrently. We take a min-max optimization model and minimize (over design) the maximum (over population) variance of the estimated interaction effect between treatment and patient covariates. This results in a min-max bilevel mixed integer nonlinear programming problem, which is notably challenging to solve. To address this challenge, we introduce a surrogate optimization model by approximating the objective function, for which we propose two solution approaches. The first approach provides an exact solution based on reformulation and decomposition techniques. In the second approach, we provide a lower bound for the inner optimization problem and solve the outer optimization problem over the lower bound. We test our proposed algorithms with synthetic and real-world data sets and compare them with standard (re)randomization methods. Our numerical analysis suggests that the proposed approaches provide higher-quality solutions in terms of the variance of estimators and probability of correct selection. We also show the value of covariate information in precision medicine clinical trials by comparing our proposed approaches to an alternative optimal design approach that does not consider the interaction terms between covariates and treatment.

Summary of Contribution: Precision medicine is the future of healthcare where treatment is prescribed based on each patient information. Designing precision medicine clinical trials, which are the cornerstone of precision medicine, is extremely challenging because sample size is limited and patient information may be multidimensional. This work proposes a novel approach to optimally estimate the treatment effect for each patient type in a two-armed clinical trial by reducing the largest variance of personalized treatment effect. We use several statistical and optimization techniques to produce efficient solution methodologies. Results have the potential to save countless lives by transforming the design and implementation of future clinical trials to ensure the right treatments for the right patients. Doing so will reduce patient risks and reduce costs in the healthcare system.

History: Accepted by J. Paul Brooks, Area Editor for Applications in Biology, Medicine, & Healthcare.
Funding: This research is supported by the National Science Foundation [Grant 1651912].

Keywords: precision medicine • clinical trials • experimental design • minimax optimization

1. Introduction

The average cost of bringing a new treatment to market has surpassed $2.6 billion, and expensive clinical trials are the major driver of such a high cost (Tufts 2014). In particular, the total cost of clinical trials can reach $300–$600 million for large global trials; the costs usually increase with each phase of the trial (Giffin et al. 2010). Clinical trial costs depend on a variety of factors, such as the number of participants, number and locations of research facilities, complexity of the trial protocol, and reimbursement provided to investigators. In particular, the top three cost drivers of clinical trial expenditures are clinical procedure, administrative staff, and site monitoring costs (Sertkaya et al. 2016). Several different communities, including statistics/biostatistics, public health sciences, economics, machine learning, and operations research, have studied different aspects of this complex procedure.

Specifically, we aim to incorporate patients' covariate information into the optimal design. This is motivated by the recent significant interest in *precision medicine* (sometimes it is also referred to as *personalized medicine*, and we use both terms interchangeably throughout the paper). Precision medicine seeks to maximize the quality of healthcare by providing individual-level healthcare for each patient and has recently gained prominence as the future of healthcare (Kosorok and Laber 2019). In fact, there is significant evidence that ignoring patient individualized information in prescribing medicine can impact the efficacy of treatment and potentially be harmful. For example, Schork (2015) provided surprising statistics that the top 10 highest-grossing drugs help between 1 in 25 and 1 in 4 of the patients. This number for statins can be as low as 1 in 50. Motivated by such evidence, governments and healthcare institutions have emphasized precision medicine and allocated significant resources for research and development in this area (Hayden 2015). The majority of the literature investigates the optimal decision making of personalized treatment based on statistical analysis, see, for example, Shi et al. (2018).

A key step in personalized medicine is the ability to design clinical trials that focus on an individual, not average, response to treatment (Schork 2015). The key difference between precision medicine and population-level treatment in terms of their statistical analysis lies in the difference between their respective statistical models. This means that experiment designs originally developed for population-level treatment analysis may be inappropriate for precision medicine: it may deteriorate the accuracy and efficiency of their model estimation. It is critically important to investigate optimal design specifically for precision medicine, that is, how to collect experimental data aiming to optimize the effectiveness of the subsequent statistical analysis for precision medication in the present of patient covariates. To the best of our knowledge, this setting has not been addressed in the literature. To fill this gap, this paper extends a conventional approach for optimal design of clinical trials by incorporating patients' personalized covariate information, focusing on two-armed clinical trials. Formally, given a set of patients with covariate information, we study how to allocate them to two different treatments in order to maximize the worst-case accuracy (over covariates) of statistical inferences about the treatment efficacy. Note that there are many other types of designs for clinical trials in the literature, such as response adaptive and Bayesian; for a survey of different types of designs for clinical trials, see Berry (2006), Press (2009), Kotas and Ghate (2018), and references therein.

In this study, we consider the optimal design of clinical trials with two treatment options. This includes Phase III clinical trials, where a novel treatment is usually tested against a standard treatment or a placebo. Phase III clinical trials are the most expensive ones, and improving the accuracy of statistical inferences about the treatment efficacy can significantly improve the quality of the procedure (Giffin et al. 2010). Our proposed design also applies to a class of Phase II clinical trials where two dosage levels are involved and the decision maker is interested in the dose with the highest response. Phase II clinical trials usually involve finding the minimum effective dose, maximum tolerable dose, and 95% effective dose (Berry et al. 2002). Our proposed design is not directly applicable to Phase I clinical trials, in which the safe dose range is usually found by dose escalation principles such as $3+3$ design (Le Tourneau et al. 2009).

The theory of optimal experiment design started with the early development by Fisher (1936). Classical optimal designs focus on reducing the variabilities of parameter estimation in a statistical model. Different types of optimal designs are often led by optimizing different utility functions of the variance-covariance matrix of the estimated parameters (Wu and Hamada 2011, Morris et al. 2015). For example, the D-optimal design corresponds to an optimal solution of minimizing the determinant of the generalized variance matrix of the parameter estimates for the underlying statistical model. As a result of the complex objective function employed in an optimal design problem, the corresponding optimization problem is usually challenging to solve. Off-the-shelf optimization solvers are usually incapable of providing exact optimal solutions of these optimization problems; see, for example, Singh and Xie (2020).

Specialized solution methodologies must be developed to address this computational challenge. For instance, Bertsimas et al. (2015) have proposed a design problem that minimizes the maximum discrepancy in mean and variance among different treatment groups. Their proposed designs yield a significant improvement over (re)randomized designs in terms of statistical inference in the population level. Our paper incorporates precision medication in statistical modeling and develops optimal designs to improve the accuracy of this task. Thus, the resulting structure of our optimization formulations is significantly different from the one proposed in Bertsimas et al. (2015). Specifically, in the context of two-armed clinical trials, their problem reduces to a single-level mixed integer linear program, which can be adequately handled by an off-the-shelf optimization solver. In contrast, our problem corresponds to a min-max bilevel mixed integer nonlinear program, for which we propose specialized algorithms. One other related work is by Bhat et al. (2020), who studied optimal design of experiments with covariates for A-B testing both in offline and online settings. In their offline setting, similar to our work, they studied an optimal design of experiments with a linear response model. The major difference is that we incorporate the interaction effects between patient information and treatment allocation, while they do not. This simplification allows them to use a tractable approximation for the optimal design problem that minimizes the variance of the estimator. This is in stark contrast with our min-max bilevel mixed integer nonlinear programming formulations, which are computationally difficult to solve. In particular, the inclusion of the interaction terms between patient covariates and treatments is highly valuable in that it delivers different treatment effects for different covariates representing different individuals, which is the goal of precision medicine. We demonstrate the value of incorporating this interaction in our numerical experiments.

1.1. Main Contributions

We summarize our main contributions as follows:

- First, we formulate the optimal design problem for a two-armed clinical trial by incorporating the interaction between treatment allocation and patient covariate information. In the literature of two-armed clinical trials, previous attempts do not consider the interaction between treatment and patient covariates in the context of optimal design in the presence of patient covariates. In contrast, our model explicitly incorporates covariate information in treatment effects, for which the optimization of statistical accuracy is formulated as a min-max bilevel optimization problem: the decision maker seeks to minimize the worst case (over covariates) variance of the estimates of individualized treatment effect.
- Second, we propose a surrogate model by approximating the variance of the estimator. The core of the approximation is adopting an asymptotic balance design in the Taylor expansion of the objective function. Despite this approximation, the optimization problem is still a min-max bilevel mixed integer nonlinear program, for which we propose two solution approaches to solve it. The first approach solves the surrogate model exactly and is based on a reformulation of the min-max problem using decomposition techniques. The second approach provides a lower bound for the inner maximization problem and the outer minimization is carried out over this lower bound. The appealing feature of the lower bounding approach is that it yields a single-level optimization problem, which scales well with the size of the problem including large clinical trials with hundreds of patient covariates.
- Finally, we apply our algorithms on several sets of synthetic data and a case study with real data for patient covariates from a large clinical trial for Warfarin, a popular anticoagulant medication used to treat blood clots. Our numerical results show that the proposed algorithms outperform the standard randomization and rerandomization methods that are widely used in the literature in all tested settings. We also compare our proposed design with that of Bhat et al. (2020) as a benchmark to shed light on the value of incorporating patient covariates into treatment effect. In particular, our results show that the proposed lower bounding approach performs robustly in terms of the corresponding objective values of both the surrogate model and the original model. This observation suggests that the lower bounding approach can be a fast and reliable option for optimal design of precision clinical trials.

1.2 Paper Organization

Section 2 motivates and formulates the problem. Section 3 provides a surrogate model and introduces two solution approaches for the proposed optimal design problem. Section 4 summarizes our numerical study, and Section 5 gives a case study of the proposed approaches. Section 6 concludes the paper.

2. A Min-Max Optimal Design Problem for Precision Medicine Clinical Trials

This section develops an optimal design objective that is oriented toward the goal of improving the accuracy of precision medicine decisions. Following classical assumptions in the literature (Qian and Murphy 2011, Atkinson 2015, Laber et al. 2016), we consider a linear model to describe the treatment-response relationship in the presence of patient covariate information. In particular, let $x \in \{-1, 1\}$ denote the two treatment levels, $\mathbf{z} = (1, z_1, \ldots, z_{p-1})^\top \in \mathcal{Z} \subset \mathbb{R}^p$ with $p > 1$ denote the noncontrollable patient covariates, and $y \in \mathbb{R}$ be a numerical response. The treatment-response relationship is then given by

$$y = \mathbf{z}^\top \boldsymbol{\alpha} + x\mathbf{z}^\top \boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{p-1})^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ are the linear coefficients and ε models randomness in response and follows a normal distribution $N(0, \sigma^2)$. The purpose of personalized medicine is to recommend patient-specific treatment. To that end, the decision maker seeks to find the best (in terms of maximal response) treatment for each patient given its covariate $\mathbf{z} \in \mathcal{Z}$, which is defined by

$$x^*(\mathbf{z}) := \operatorname*{argmax}_{x \in \{-1,+1\}} \{\mathbf{z}^\top \boldsymbol{\alpha} + x\mathbf{z}^\top \boldsymbol{\beta}\} = \operatorname*{argmax}_{x \in \{-1,+1\}} \{x\mathbf{z}^\top \boldsymbol{\beta}\}. \tag{2}$$

Apparently, the optimal decision in (2) can be viewed as a function of the individual's covariate information $\mathbf{z}$. We observe that the objective in (2) can be expressed by

$$x\mathbf{z}^\top \boldsymbol{\beta} = x\beta_0 + \sum_{i=1}^{p-1} xz_i\beta_i.$$

Thus, the difference between the two treatment decisions for each individual depends on the significance of the coefficient parameters associated with covariates $xz_1, \ldots, xz_{p-1}$, that is, the interaction between treatment x and the patient covariates. If coefficients $\beta_1, \ldots, \beta_{p-1}$ are zeros in the model, we see that the personalized optimal decision $x^*(\mathbf{z})$ is reduced to the population level optimal decision

$$x^* = \operatorname*{argmax}_{x \in \{-1,+1\}} x\beta_0, \tag{3}$$

as studied in Bhat et al. (2020). Compared with the work of Bhat et al. (2020), Model (1) significantly improves the relevance to precision medicine. On the other hand, it also notably increases the complexity of the statistical analysis and computation, as the resulting estimators are multidimensional and include patient covariates. We next elaborate on these challenges.

2.1. Optimal Design with Covariates: A Min-Max Bilevel Optimization Problem

Suppose that n patients are recruited for the clinical trial; the covariate information of each patient i is given by $\mathbf{h}_i \in \mathcal{Z}$; and all patient covariate information is represented by an $n \times p$ matrix $H = (\mathbf{h}_1,\dots,\mathbf{h}_n)^\top$, where $^\top$ denotes matrix transpose. Let $x_i \in \{-1, 1\}$ denote the treatment prescribed to patient i and let $\mathbf{x} = (x_1,\dots,x_n)^\top$ be the treatment allocation of n patients. After the trial is finished and all the responses of patients are collected, the estimated coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (1) can be expressed by $\hat{\boldsymbol{\alpha}}(\mathbf{x},H)$ and $\hat{\boldsymbol{\beta}}(\mathbf{x},H)$, which are functions of the design $\mathbf{x}$ and patient covariate information H. Based on estimates of the model parameters, the decision maker is able to infer the best treatment for each patient type with covariate $\mathbf{z} \in \mathcal{Z}$. Let $\hat{x}(\mathbf{z};H,\mathbf{x})$ denote the suggested treatment to patients with covariates $\mathbf{z}$ where the trial has patient information H and allocation prescribed is $\mathbf{x}$. A natural choice for $\hat{x}(\mathbf{z};H,\mathbf{x})$ is then given by

$$\begin{aligned}\hat{x}(\mathbf{z};H,\mathbf{x}) &:= \operatorname*{argmax}_{x\in\{-1,+1\}} \mathbf{z}^\top \hat{\boldsymbol{\alpha}}(\mathbf{x},H) + x\mathbf{z}^\top \hat{\boldsymbol{\beta}}(\mathbf{x},H)\\ &= \operatorname*{argmax}_{x\in\{-1,+1\}} x\mathbf{z}^\top \hat{\boldsymbol{\beta}}(\mathbf{x},H). \qquad (4)\end{aligned}$$

Recall that the treatment effect in Model (1) is identified by $\mathbf{z}^\top\boldsymbol{\beta}$, which can be estimated by $\mathbf{z}^\top\hat{\boldsymbol{\beta}}(\mathbf{x},H)$. In order to have a higher precision in statistical inference, it is natural to minimize the variance of $\mathbf{z}^\top\hat{\boldsymbol{\beta}}(\mathbf{x},H)$ for each individual value of $\mathbf{z}$. According to our assumptions, ε in (1) follows a normal distribution $\varepsilon \sim N(0,\sigma^2)$; thus $\mathbf{z}^\top\hat{\boldsymbol{\beta}}(\mathbf{x},H)$ also follows a normal distribution with mean $\mathbf{z}^\top\boldsymbol{\beta}$ and variance $\mathbf{z}^\top\Sigma_\beta(\mathbf{x},H)\mathbf{z}$, where $\Sigma_\beta(\mathbf{x},H)$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}(\mathbf{x},H)$. In other words, the quality of estimates depends on the allocation $\mathbf{x}$ and can be a subject for optimization. From an optimization perspective, we aim to minimize the worst-case (maximum) variance of the estimated interaction effect $\mathbf{z}^\top\hat{\boldsymbol{\beta}}(\mathbf{x},H)$ among all patient covariates $\mathbf{z} \in \mathcal{Z}$, which yields the following optimization problem:

$$\min_{\mathbf{x}\in\{-1,1\}^n} \max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^\top \Sigma_\beta(\mathbf{x},H)\mathbf{z}. \qquad (5)$$

Notice that, if the interaction between treatment and covariates is not included, the treatment and covariates follow an *additive structure* in the model. The accuracy of population-level optimal decision in (3) is determined by the accuracy of $\hat{\beta}_0(\mathbf{x},H)$, which is the estimator of the global treatment effect. Therefore, as studied in Bhat et al. (2020), the optimal design in this case corresponds to minimizing the variance of $\hat{\beta}_0(\mathbf{x},H)$, which can be written as

$$\operatorname{var}\left[\hat{\beta}_0(\mathbf{x},H)\right] \propto \frac{1}{\mathbf{x}^\top\left[I - H(H^\top H)^{-1}H^\top\right]\mathbf{x}}.$$

This is equivalent to solving the following convex quadratic 0-1 integer program:

$$\min_{\mathbf{x}\in\{-1,1\}^n} \mathbf{x}^\top H(H^\top H)^{-1}H^\top\mathbf{x}, \qquad (6)$$

which can be handled, for example, by a modern commercial solver such as Gurobi.

2.2. Challenges for Solving the Min-Max Bilevel Optimization (5)

We next characterize the variance-covariance matrix $\Sigma_\beta(\mathbf{x},H)$ in the objective function of (5) and point out the challenges in solving this optimal design problem. Notice that the dimension of covariates in (1) is $2p$ by including the main effect $\mathbf{z}$ and the interaction $x\mathbf{z}$. After stacking all the covariates from all n patients, we denote the $n \times 2p$ covariates matrix by

$$X = \left[H \ D_{\mathbf{x}}H\right],$$

where $D_{\mathbf{x}} = \operatorname{diag}(x_1,\dots,x_n)$ is a diagonal $n \times n$ matrix and H and $D_{\mathbf{x}}H$ are the matrices of patient covariates and the matrices that characterize interactions between treatment allocation and patient covariates, respectively. Thus, the variance matrix of the estimated parameters $(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ in (1) can be expressed by

$$\sigma^2(X^\top X)^{-1} = \sigma^2\begin{bmatrix} H^\top H & H^\top D_{\mathbf{x}}H \\ H^\top D_{\mathbf{x}}H & H^\top H\end{bmatrix}^{-1}.$$

By taking the inversion of the above block matrix, the variance of the estimator for $\boldsymbol{\beta}$ corresponds to the second diagonal block entry, which is given by

$$\Sigma_\beta(\mathbf{x},H) = \sigma^2\left(H^\top H - H^\top D_{\mathbf{x}}H(H^\top H)^{-1}H^\top D_{\mathbf{x}}H\right)^{-1}. \qquad (7)$$

Plugging Equation (7) into optimization Problem (5) results in a min-max bilevel nonconvex mixed integer nonlinear program, which is notoriously difficult to solve even when the inner problem is a mixed integer program (DeNegre 2011, Tang et al. 2016). Furthermore, the covariate matrix in (7) makes the optimization problem challenging to handle directly because of the matrix inverse. The next section introduces our proposed approaches to address the computational challenges in solving (5).

3. Solution Methodology

Our proposed solution approaches are based on an approximation to optimization Problem (5) using a surrogate objective function. We describe this surrogate objective function in Section 3.1. The construction of the surrogate model is based on a natural asymptotic result on the number of patients allocated to each treatment; that is, as the number of patients increases, the optimal design converges to a balanced design, which is the gold standard in the literature; see, for example, Kallus (2018). After applying this surrogate function, the problem remains a min-max bilevel nonconvex mixed integer nonlinear program; we subsequently describe two solution approaches. In Section 3.2, we provide an exact algorithm to solve this surrogate model based on decomposition and reformulation methods. In Section 3.3, we derive a lower bound for the inner optimization problem, which allows the min-max bilevel optimization problem to be further approximated by a single-level optimization problem by replacing the inner optimization problem with this lower bound.

3.1. A Surrogate Optimization Model

In this section, we approximate $\Sigma_\beta(\mathbf{x}, H)$ based on some natural assumptions in our problem context and then use this approximation to construct a surrogate optimization problem for (5). Let n_+ and n_- denote the number of patients that are allocated to treatments 1 and -1, respectively. In addition, let $\mathbb{P}_\mathbf{z}$ denote the probability measure defined over the covariate space and $\mathbb{E}_\mathbf{z}$ denote the expectation with respect to the said probability measure. The next lemma is crucial in our construction, which ensures that under an asymptotically balanced design, the following matrix behaves asymptotically as $n^{-1}I_p$, where I_p is a $p \times p$ identity matrix.

Lemma 1. *Let* $\mathbb{E}_\mathbf{z}(z_l z_k) = \gamma_{lk}$ *and assume that* $|\gamma_{ll}| \geq \gamma > 0,\ \forall l \in \{1, 2, \ldots, p\}$ *and* $n^{-1}(n_+ - n_-) = O(n^{-1})$*, then*

$$(H^\top H)^{-1} H^\top D_\mathbf{x} H = O_p(n^{-1}) I_p.$$

Proof. By law of large numbers, we have

$$n^{-1} \sum_{i=1}^{n} z_{il} z_{ik} = \gamma_{lk} + o_p(1).$$

Let Γ be a $(p+1) \times (p+1)$ matrix with ij-th entry γ_{ij}. As a result,

$$H^\top H / n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^\top = \Gamma + o_p(1).$$

$$\begin{aligned} H^\top D_\mathbf{x} H / n &= \frac{1}{n} \sum_{i=1}^{n} x_i \mathbf{z}_i \mathbf{z}_i^\top = \frac{1}{n} \sum_{\{i: x_i = 1\}} \mathbf{z}_i \mathbf{z}_i^\top - \frac{1}{n} \sum_{\{i: x_i = -1\}} \mathbf{z}_i \mathbf{z}_i^\top \\ &= \frac{n_+}{n} n_+^{-1} \sum_{\{i: x_i = 1\}} \mathbf{z}_i \mathbf{z}_i^\top - \frac{n_-}{n} n_-^{-1} \sum_{\{i: x_i = -1\}} \mathbf{z}_i \mathbf{z}_i^\top \\ &= \left(\frac{n_+}{n} - \frac{n_-}{n} \right) \left(\Gamma + o_p(1) \right) \end{aligned}$$

Consider that $n^{-1}(n_+ - n_-) = O(n^{-1})$, we have

$$(H^\top H)^{-1} H^\top D_\mathbf{x} H = O_p(n^{-1}) I. \quad \square$$

The assumption $n^{-1}(n_+ - n_-) = O(n^{-1})$ formalizes an asymptotically balanced design concept, which is considered a gold standard in the optimal design of experiments. The covariate matrix in (7) is difficult to manage because it does not give an explicit formula on $\mathbf{x}$, as a result of the matrix inverse. To simplify this expression, we apply the Taylor expansion of (7) and use results provided in Lemma 1.

Proposition 1. *Under the assumptions of* Lemma 1*, we have*

$$\begin{aligned} \Sigma_\beta(\mathbf{x}, H) = \sigma^2 \Big(I + (H^\top H)^{-1} H^\top D_\mathbf{x} H \, (H^\top H)^{-1} \\ \times H^\top D_\mathbf{x} H \Big) (H^\top H)^{-1} + O\,(n^{-4}) A\,(\mathbf{x}, H), \end{aligned} \tag{8}$$

where $A(\mathbf{x}, H)$ *is the remainder matrix for coefficients with* n^{-4} *and polynomials of higher degrees in the Taylor's expansion.*

Proof. Recall that

$$\begin{aligned} \underset{\beta}{\Sigma}(\mathbf{x}, H) &= \sigma^2 (H^\top H - H^\top D_\mathbf{x} H \, (H^\top H)^{-1} H^\top D_\mathbf{x} H)^{-1} \\ &= \sigma^2 ((H^\top H)(I - (H^\top H)^{-1} H^\top D_\mathbf{x} H \; (H^\top H)^{-1} \times H^\top D_\mathbf{x} H))^{-1} \\ &= \sigma^2 (I - (H^\top H)^{-1} H^\top D_\mathbf{x} H \; (H^\top H)^{-1} H^\top D_\mathbf{x} H)^{-1} (H^\top H)^{-1} \end{aligned}$$

According to the Taylor expansion of matrix inversion, if a matrix A has a spectral radius less than one

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

According to Lemma 1, we have that this condition holds when n is large enough. Therefore,

$$\begin{aligned} &\left(I - (H^\top H)^{-1} H^\top D_\mathbf{x} H \, (H^\top H)^{-1} H^\top D_\mathbf{x} H \right)^{-1} \\ &\quad = I + (H^\top H)^{-1} H^\top D_\mathbf{x} H \; (H^\top H)^{-1} H^\top D_\mathbf{x} H \\ &\quad + \sum_{k=2}^{\infty} \left((H^\top H)^{-1} H^\top D_\mathbf{x} H \, (H^\top H)^{-1} H^\top D_\mathbf{x} H \right)^k \end{aligned}$$

According to Lemma 1, the above higher order term becomes

$$\sum_{k=2}^{\infty} \left((H^\top H)^{-1} H^\top D_\mathbf{x} H (H^\top H)^{-1} H^\top D_\mathbf{x} H \right)^k = O(n^{-4}) A(\mathbf{x}, H),$$

where $A(\mathbf{x}, H)$ represents the reminder matrix. Then the conclusion holds. □

Proposition 1 paves the way to construct the surrogate model. In fact, it is natural to replace $\Sigma_\beta(\mathbf{x}, H)$ by the first term of (8). To streamline notation, let

$$\Psi(\mathbf{x}, H) = (H^\top H)^{-1} H^\top \, D_\mathbf{x} H \; (H^\top H)^{-1} H^\top D_\mathbf{x} H \; (H^\top H)^{-1}. \tag{9}$$

Then, a surrogate model for optimization Problem (5) can be formulated as

$$\min_{\mathbf{x}} \max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^{\top}\left((H^{\top}H)^{-1} + \Psi(\mathbf{x},H)\right)\mathbf{z}, \tag{10a}$$

$$\text{s.t.} \quad -1 \le \sum_{i=1}^{n} x_i \le 1, \tag{10b}$$

$$\mathbf{x} \in \{-1,+1\}^{n}. \tag{10c}$$

We provide numerical evidence regarding the quality of the surrogate model in Section 4.1.

Note that compared with the original Problem (5), we add an additional Constraint (10b) to the outer optimization problem to ensure a balanced design. Particularly, if n is an even number, this constraint becomes the exact balancing constraint, that is, $\sum_{i=1}^{n} x_i = 0$. If n is an odd number, exact balancing over $\mathbf{x}$ is impossible; thus we relax the constraint to be $-1 \le \sum_{i=1}^{n} x_i \le 1$. Before proceeding, the following lemma shows that the objective function of (10a)–(10c) can be rewritten as a quadratic function of $\mathbf{x}$ given a fixed $\mathbf{z}$.

Lemma 2. *The following equality holds*

$$\mathbf{z}^{\top}\Psi(\mathbf{x},H)\mathbf{z} = \mathbf{x}^{\top}\Upsilon(\mathbf{z},H)\mathbf{x}, \tag{11}$$

where

$$\Upsilon(\mathbf{z},H)=\left(H(H^{\top}H)^{-1}H^{\top}\right)\circ\left(H(H^{\top}H)^{-1}\mathbf{z}(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}\right),$$

and $\circ$ *denotes the Hadamard product (i.e., matrix elementary-wise product). In addition, matrix* $\Upsilon(\mathbf{z},H)$ *is positive semidefinite (PSD) for any* $\mathbf{z}$.

Proof. Let $\mathrm{tr}(\cdot)$ denote the trace of a matrix. The result follows by

$$\begin{aligned}
&(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}D_{\mathbf{x}}H(H^{\top}H)^{-1}H^{\top}D_{\mathbf{x}}H(H^{\top}H)^{-1}\mathbf{z}\\
&\quad= \mathrm{tr}\left\{(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}D_{\mathbf{x}}H(H^{\top}H)^{-1}H^{\top}D_{\mathbf{x}}H(H^{\top}H)^{-1}\mathbf{z}\right\}\\
&\quad= \mathrm{tr}\left\{D_{\mathbf{x}}H(H^{\top}H)^{-1}H^{\top}D_{\mathbf{x}}H(H^{\top}H)^{-1}\mathbf{z}(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}\right\}\\
&\quad= \mathbf{x}^{\top}\left\{H(H^{\top}H)^{-1}H^{\top}\circ H(H^{\top}H)^{-1}\mathbf{z}(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}\right\}\mathbf{x}. \;\square
\end{aligned}$$

Because both $H(H^{\top}H)^{-1}H^{\top}$ and $H(H^{\top}H)^{-1}\mathbf{z}(\mathbf{z})^{\top}(H^{\top}H)^{-1}H^{\top}$ are PSD, their Hadamard product $\Upsilon(\mathbf{z},H)$ is also PSD.

Although problem (10) is still a min-max bilevel mixed integer nonlinear program, its objective function is much easier to handle: given a fixed $\mathbf{x}$, the objective function is a convex quadratic function of $\mathbf{z}$ (see (10a)), and given a fixed $\mathbf{z}$, the objective function is a convex quadratic function of $\mathbf{x}$ (see (11)). We next develop exact and approximate approaches to solve it.

3.2. An Exact Algorithm for Solving the Surrogate Formulation

In this section, we develop an exact algorithm to solve formulation (10) by applying a cutting plane procedure on a reformulation of formulation (10). This reformulation is motivated by the fact that the objective function of problem (10) is a convex quadratic function of $\mathbf{x}$ given a fixed $\mathbf{z}$, and a convex quadratic function of $\mathbf{z}$ given a fixed $\mathbf{x}$. Before proceeding, we make an assumption on the patient covariate space to facilitate the derivation of the algorithm. In particular, we assume that $z_i \in \{-1,1\}$ for all $i = 1,2,\ldots,p-1$, that is, each covariate can be represented by a binary variable. This assumption results in $\mathcal{Z} = 1 \times \{-1,1\}^{p-1}$, where the first "1" indicates that the first covariate is set to be one (as the intercept). Note, however, that this assumption is made without loss of generality, because it is well known that a linear model with categorical covariates can be transformed into an equivalent linear model in which all the covariates are binary (Rencher and Schaalje 2008). Also, in the context of clinical trials, the covariates are usually categorical, such as age group, gender, and health category, which are discrete and bounded. Specifically for precision medicine, the covariates of interest include genomic biomarkers, which are typically discrete (Majewski and Bernards 2011). Therefore, one can easily construct a linear model with binary covariates. We denote $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{2^{p-1}}\}$.

First, the surrogate model (10) can be reformulated as

$$\begin{aligned}
&\min\ \theta\\
&\text{s.t. } \theta \ge \mathbf{z}^{\top}(H^{\top}H)^{-1}\mathbf{z}+\mathbf{x}^{\top}\Upsilon(\mathbf{z},H)\mathbf{x}, \quad \forall \mathbf{z} \in \mathcal{Z},\\
&\qquad -1 \le \sum_{i=1}^{n} x_i \le 1,\\
&\qquad \mathbf{x} \in \{-1,1\}^{n}.
\end{aligned} \tag{12}$$

The above formulation is a convex integer quadratic program, and we propose a cutting-plane-based exact solution approach. In particular, let $\mathcal{Z}_m \subset \mathcal{Z}$ and define the following so-called master problem that gives a relaxation of (12):

$$\begin{aligned}
&\min\ \theta\\
&\text{s.t. } \theta \ge \mathbf{z}^{\top}(H^{\top}H)^{-1}\mathbf{z} + \mathbf{x}^{\top}\Upsilon(\mathbf{z},\mathrm{H})\mathbf{x}, \quad \forall \mathbf{z} \in \mathcal{Z}_{\mathrm{m}},\\
&\quad -1 \le \sum_{i=1}^{n} x_i \le 1,\\
&\quad \mathbf{x} \in \{-1,1\}^{n}.
\end{aligned} \tag{13}$$

Let $\mathbf{x}_m$ and θ_m be an optimal solution to the Master Problem (13). If $\mathbf{x}_m$ and θ_m satisfy all of the constraints in Formulation (12), then, $\mathbf{x}_m$ and θ_m are optimal to (12). Otherwise, we should add elements in $\mathcal{Z} \setminus \mathcal{Z}_m$ to $\mathcal{Z}_m$ and resolve the Master Problem (13) to obtain a tighter relaxation. To find these elements (if any), we solve the following subproblem:

$$\delta_m = \max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^\top\left((H^\top H)^{-1} + \Psi(\mathbf{x}_m, H)\right)\mathbf{z}. \tag{14}$$

If $\theta_m \geq \delta_m$, the optimal solution is $\mathbf{x}_m$; otherwise, we add an optimal solution of (14) to $\mathcal{Z}_m$ and continue the procedure by resolving the Master Problem (13). Because $\mathcal{Z}$ is finite, this procedure converges to an optimal solution of (12) in a finite number of steps. In our implementation of the algorithm though, we use the stopping criteria of $\theta_m \geq \delta_m - \epsilon$ for some prespecified threshold $\epsilon > 0$. Observe that Subproblem (14) is a nonconvex quadratic integer program, which is difficult to solve in general. However, recall that each $z_i \in \{-1, 1\}$, then one can apply McCormick reformulation of bilinear and quadratic terms and the resulting reformulation will be a mixed integer linear program, which can be solved by an off-the-shelf optimization solver.

3.3. A Lower Bounding Approximation to the Inner Maximization Problem of (10)

In this section, we propose a lower bound for the inner maximization problem of (10) and we propose a heuristic approach for solving the surrogate model (10) by solving a single-level optimization problem, which is obtained by replacing the inner maximization problem with this lower bound. To that end, we assume that $\mathcal{Z} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$, that is, the collection of covariate vectors of patients coincide with the space of all possible covariates of interest. This assumption, though, is not restrictive because when decision makers set up a clinical trial to investigate the efficacy of a drug for a specific set of covariates, they recruit patients with these covariates. More importantly, although the validity of the lower bounding technique relies on this assumption, the solutions produced by following this heuristic approach can be used even for settings where the assumption is not satisfied. Our numerical results show that the solutions provided by this approach are competitive with those derived by the exact algorithm presented in Section 3.2 for the surrogate model, and it outperforms the exact algorithm in terms of the original objective (5) when the number of covariates is large; see Section 4.1.

Proposition 2. *For any given* $\mathbf{x}$*, the inner maximization problem of* (10) *is lower bounded by*

$$\frac{p}{n} + \frac{1}{n}\mathbf{x}^\top\left[\left(H(H^\top H)^{-1}H^\top\right)\circ\left(H(H^\top H)^{-1}H^\top\right)\right]\mathbf{x},$$

where $\circ$ *denotes the Hadamard product as in* (11).

Proof. By letting $\mathcal{Z} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$, we observe that $\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i^\top = H^\top H$. Let $\operatorname{tr}(\cdot)$ denote the trace of a matrix, then for a given $\mathbf{x}$, the inner problem in (10),

$$\begin{aligned}
&\max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^\top(H^\top H)^{-1}\mathbf{z} + \mathbf{z}^\top\Psi(\mathbf{x}, H)\mathbf{z}\\
&= \max_{\mathbf{z}\in\mathcal{Z}} \quad \operatorname{tr}\left((H^\top H)^{-1}\mathbf{z}\mathbf{z}^\top\right) + \mathbf{x}^\top\Upsilon(\mathbf{z}, H)\mathbf{x}\\
&\geq n^{-1}\sum_{i=1}^n\left(\operatorname{tr}\left((H^\top H)^{-1}\mathbf{h}_i\mathbf{h}_i^\top\right) + \mathbf{x}^\top\Upsilon(\mathbf{h}_i, H)\mathbf{x}\right)\\
&= n^{-1}\operatorname{tr}\left((H^\top H)^{-1}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i^\top\right)\\
&\quad + n^{-1}\mathbf{x}^\top\Bigg[\left(H(H^\top H)^{-1}H^\top\right)\\
&\quad\quad \circ\left(H(H^\top H)^{-1}\sum_{i=1}^n \mathbf{h}_i(\mathbf{h}_i)^\top(H^\top H)^{-1}H^\top\right)\Bigg]\mathbf{x}\\
&= \frac{p}{n} + \frac{1}{n}\mathbf{x}^\top\left[\left(H(H^\top H)^{-1}H^\top\right)\circ\left(H(H^\top H)^{-1}H^\top\right)\right]\mathbf{x}. \quad \square
\end{aligned}$$

Therefore, we settle to optimize the above lower bound, which depends on $\mathbf{x}$ only, instead of the inner maximization problem of (10). This results in the following single-level optimization problem of $\mathbf{x}$:

$$\begin{aligned}
&\min \quad \mathbf{x}^\top\left[\left(H(H^\top H)^{-1}H^\top\right)\circ\left(H(H^\top H)^{-1}H^\top\right)\right]\mathbf{x},\\
&\text{s.t.} \quad -1 \leq \sum_{i=1}^n x_i \leq 1,\\
&\qquad \mathbf{x} \in \{-1, +1\}^n.
\end{aligned} \tag{15}$$

Optimization Problem (15) is a convex quadratic integer program (matrix $(H(H^\top H)^{-1}H^\top)\circ(H(H^\top H)^{-1}H^\top)$ in the objective of (15) is PSD), which can be handled by certain off-the-shelf optimization solvers such as Gurobi.

4. Numerical Results

In this section, we present numerical results of our proposed algorithms on synthetic data sets. In order to streamline the exposition, we consider the following labels for our proposed algorithms and benchmark algorithms:

- EXACT is the cutting-plane-based approach for solving the surrogate model (10) described in Section 3.2. We use a time limit of 300 seconds for each iteration of this approach.
- LB_APPROX is the lower bound approximation for the surrogate model (10) described in Section 3.3.
- RAND is a standard (re)randomization technique used in the design of experiments. In the randomized allocation approach, we randomly allocate $n/2$ treatments of option -1 and $n/2$ treatments of option 1 to n patients (assuming n is an even number).
- ADDITIVE is the optimal design from solving the optimization problem in (6), which is based on the model with an additive structure between treatment effect and patient covariates, that is, the interaction between treatment and covariates is not included.

Next we present numerical results to evaluate the value of optimal design and quality of the surrogate model in Section 4.1, and then we present the value of including the interaction between treatment and patient covariates for precision medicine clinical trials in Section 4.2. The source code and data for experiment results shown in this section can be found at https://github.com/qiongzhangclemson/optimaldesign.

4.1. Value of Optimal Design and Quality of Surrogate Model

This section evaluates the accuracy of the surrogate model and the performance of the proposed optimization approaches with respect to random designs. Specifically, we use the objective value of the original problem in (5) (we refer to it as the "Original Objective Value") and the objective value of the surrogate model in (10) (we refer to it as the "Surrogate Objective Value") as the measurements. Because the optimal design "ADDITIVE" is constructed under a different objective (as expressed in (6)), we do not include it into the comparison in this subsection. The comparison via the surrogate objective demonstrates the value of optimal design compared with random design, and the comparison via the original objective shows the quality of the surrogate model.

For any given allocation $\mathbf{x}$, the optimal value of the inner problem of the original Problem (5) is given by

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^\top \textstyle\sum_\beta(\mathbf{x},H)\mathbf{z},$$

and the optimal value of the inner problem of the surrogate model (10) is given by

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \mathbf{z}^\top(H^\top H)^{-1}\mathbf{z}+\mathbf{z}^\top\Psi(\mathbf{x},H)\mathbf{z}.$$

For RAND, we generate 100 random allocations. The 1%, 5%, and 50% quantiles of the values from the 100 random allocations are denoted by "RAND(1%)," "RAND(5%)," and "RAND(50%)," which are compared with the optimal objective values (including the "Original Objective Value" and the "Surrogate Objective Value") obtained by EXACT and LB_APPROX. The purpose of including multiple quantiles of objective values from the random design is to show the spread of the objective value over the random designs, highlighting the value of optimal design.

We generate the synthetic data sets with random covariates matrix H in (2). Recall that H is an $n\times p$ matrix. The first column of H is loaded by ones, and we randomly generate the entries of the remaining $p-1$ columns with -1 or 1 of equal probability.

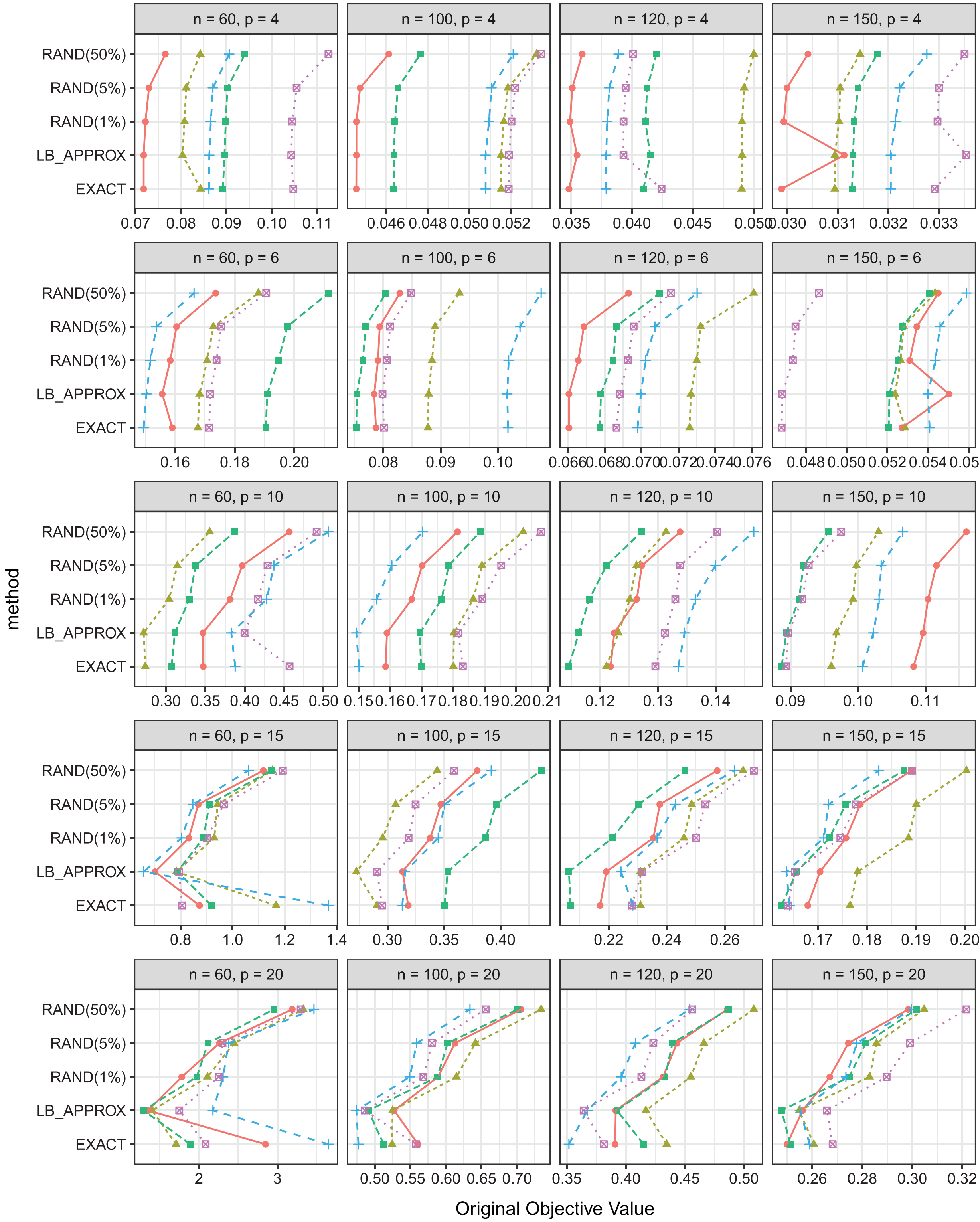
We first consider the performances of different approaches with small n and p values: $n\in\{60,100,120,150\}$, and $p\in\{4,10,15,20\}$. In order to provide a variety of estimates for the objective functions with respect to the random matrix H, we consider five randomly generated H matrices for each n and p combination. The objective values of the original optimization Problem (5) are depicted in Figure 1 for different algorithms, and the objective values of the surrogate model are depicted in Figure 2. In these two figures, each color represents results from one realization of the covariates matrix H (recall that we consider five realizations for each combination). From Figures 1 and 2, we observe that both EXACT and LB_APPROX provide more competitive results compared with the random allocation. We also observe that although EXACT produces the smallest objective value for the surrogate model in most instances, LB_APPROX is more robust in producing smaller objective values for both the original and the surrogate model. A few more observations from these two figures are in order:

- There are a few cases, especially when n is relatively small compared with p (e.g., $n=60$ and $p=15$ or 20), that the solution from EXACT is inferior compared with RAND with respect to the original objective value. The reason is that, in these cases, the surrogate model is a poor approximation to the original problem.
- When both n and p are large (e.g., $n=120$ or 150 and $p=20$), LB_APPROX may outperform EXACT in terms of both original and surrogate model objective values. The reason is that EXACT is time consuming to solve for these instances; for example, we observe that each iteration of the cutting plane algorithm often exceeds the given time limit of 300 seconds. Thus, the reported solution from EXACT could be a suboptimal solution of the surrogate model on these instances.

We now consider large-scale instances with $n=300$. When n is greater than 150, and p is greater than 30, the EXACT algorithm becomes computationally intractable. Therefore, we only compare the performance of LB_APPROX with that of RAND for $p=50$ and 100. The objective values of the original problem and the surrogate model are depicted in Figures 3 and 4, respectively. As can be seen from Figure 3, for $p=4$ and $p=10$, EXACT and LB_APPROX outperform randomized algorithms in terms of both the original and surrogate objective values. For $p=30$, EXACT performs poorly with respect to both objectives because only a suboptimal solution of a low quality is available when the time limit is reached. On the other hand, LB_APPROX outperforms randomized algorithms and produces robust performance in all ranges of p with respect to both original and surrogate models.

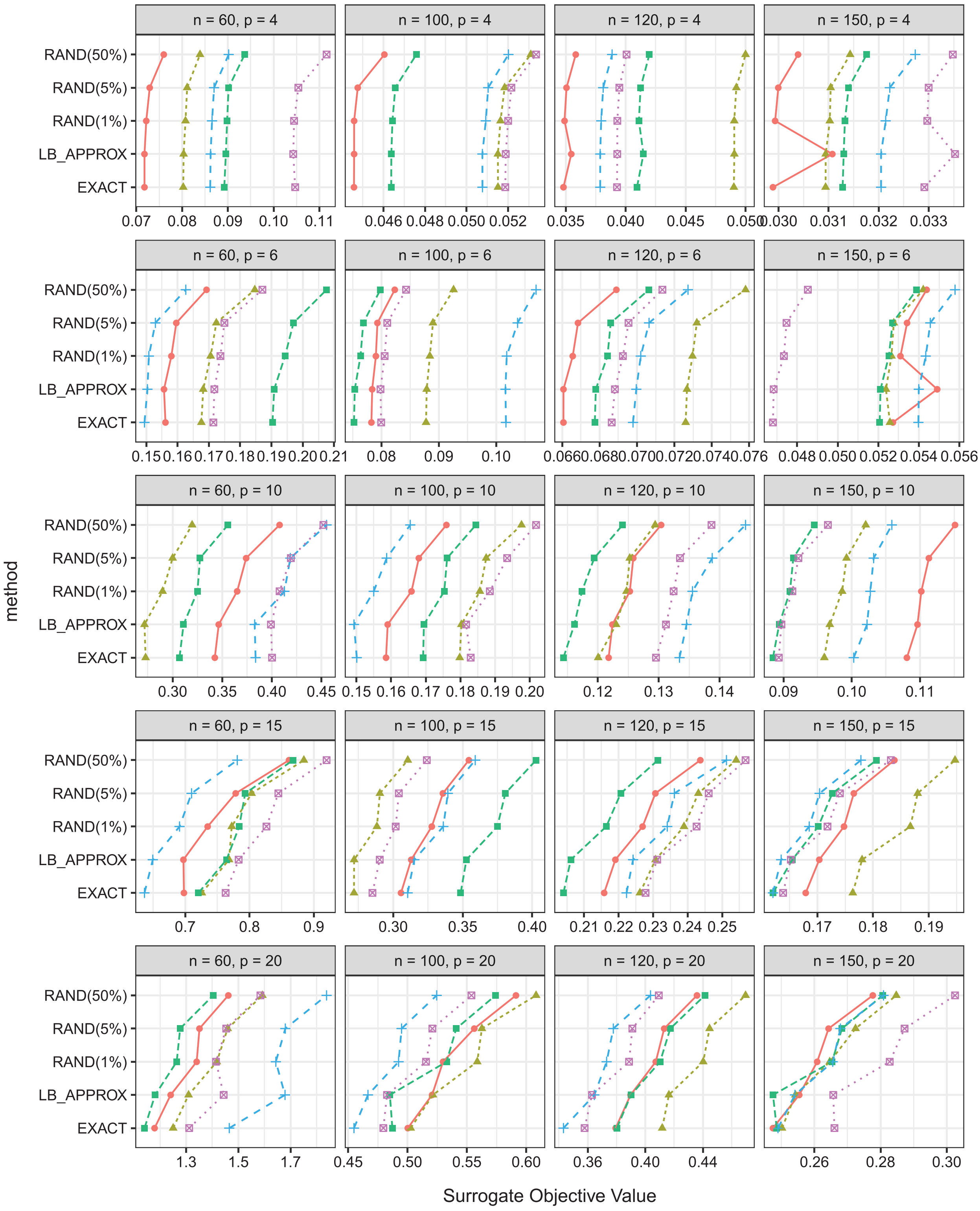
Next, we discuss the quality of our surrogate objective function compared with the original objective function. Figures 5 and 6 show the results of the instances as in Figures 1 and 3, respectively. In each subfigure, the x-axis is the value of the original objective and the y-axis is the value of the surrogate objective. The ideal case to make a good approximation from the surrogate model to the original problem is that the dots are roughly aligned along the diagonals. The results in Figures 5 and 6 show that the surrogate model is a good approximation to the

Figure 1. (Color online) Objective Values of the Original Problem with Different n and p Values



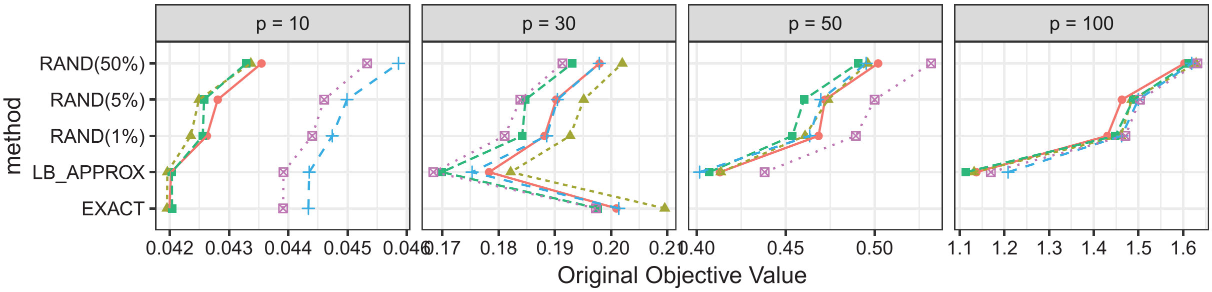
Note. Each color represents the results from one realization of H.

Figure 2. (Color online) Objective Values of the Surrogate Formulation with Different n and p Values



Note. Each color represents the results from one realization of H.

Figure 3. (Color online) Objective Values of the Original Problem with $n = 300$



Note. Each color represents the results from one realization of H

original problem when p is relatively small compared with n, whereas the approximation is not accurate if p is relatively large compared with n.

In summary, the comparison via surrogate objective shows that the optimal design given by the proposed solution approaches greatly improves random designs, and the comparison via the original objective indicates that the approximation of the original objective is accurate, especially for the cases with a larger ratio between n and p.

4.2. Value of Treatment-Covariate Interaction

For precision medicine, it is crucial to investigate the performance of the proposed approaches at the individual level. Notice that our optimization problem is formulated to optimize the worst-case scenario among all covariates. Therefore, it does not necessarily guarantee that the optimal designs can achieve better accuracy for every single individual. Throughout this subsection, we only consider the optimal design solutions to the model with treatment-covariate interactions that are obtained by the LB_APPROX approach, as we have seen in Section 4.1 that the LB_APPROX approach yields superior performance. To investigate the performance at the individual level, we compute the variance of $\mathbf{z}^\top\hat{\boldsymbol{\beta}}(\mathbf{x}, H)$ associated with the resulting optimal design and compare it with the mean variance associated with random designs for a randomly generated set of patient information. Given a patient information vector $\mathbf{z}_0$, the expected variance of the estimated interaction effect from random designs is

$$\mathrm{E}_{\mathbf{x}}\left[\mathbf{z}_0^\top \Sigma_\beta(\mathbf{x}, H)\,\mathbf{z}_0\right] = \mathbf{z}_0^\top\left[\mathrm{E}_{\mathbf{x}}\Sigma_\beta(\mathbf{x}, H)\right]\mathbf{z}_0, \tag{16}$$

where the expectation is taken with respect to the random design $\mathbf{x}$ and can be approximated empirically, for example, via a Monte Carlo sample. Given an optimal design, $\mathbf{x}^*$, for example, computed by approach EXACT or LB_APPROX, the variance of the interaction effect is

$$\mathbf{z}_0^\top \Sigma_\beta(\mathbf{x}^*, H)\mathbf{z}_0. \tag{17}$$

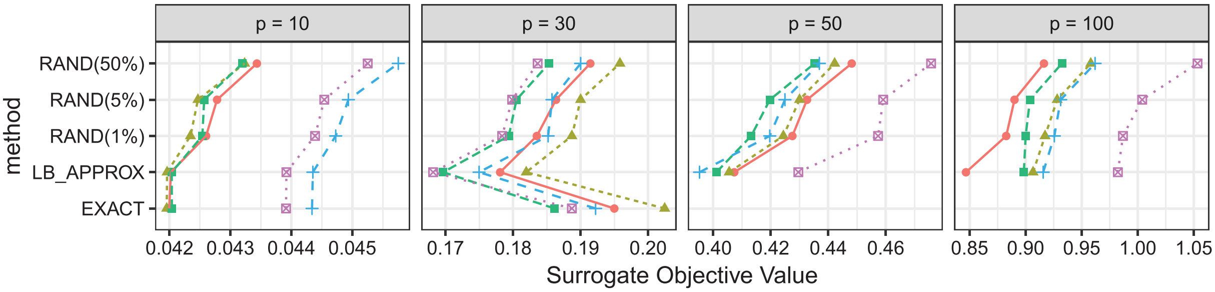
For a given vector $\mathbf{z}_0$ of patient information, the percentage of variance reduction yielded by the optimal design $\mathbf{x}^*$ compared with random designs can be expressed by

$$100 \times \frac{\mathbf{z}_0^\top\left[\mathrm{E}_{\mathbf{x}}\Sigma_\beta(\mathbf{x}, H)\right]\mathbf{z}_0 - \mathbf{z}_0^\top \Sigma_\beta(\mathbf{x}^*, H)\mathbf{z}_0}{\mathbf{z}_0^\top\left[\mathrm{E}_{\mathbf{x}}\Sigma_\beta(\mathbf{x}, H)\right]\mathbf{z}_0}\%. \tag{18}$$

We randomly generate 1,000 random designs to empirically estimate the above variance reduction measure for 1,000 randomly generated patient information vector $\mathbf{z}_0$ s. Empirically, this variance reduction indicates the extent at which the accuracy of the estimated interaction effect in (1) is improved by the proposed approach compared with a random allocation, which in turn shows the value of the proposed approach in accurately selecting personalized treatment.

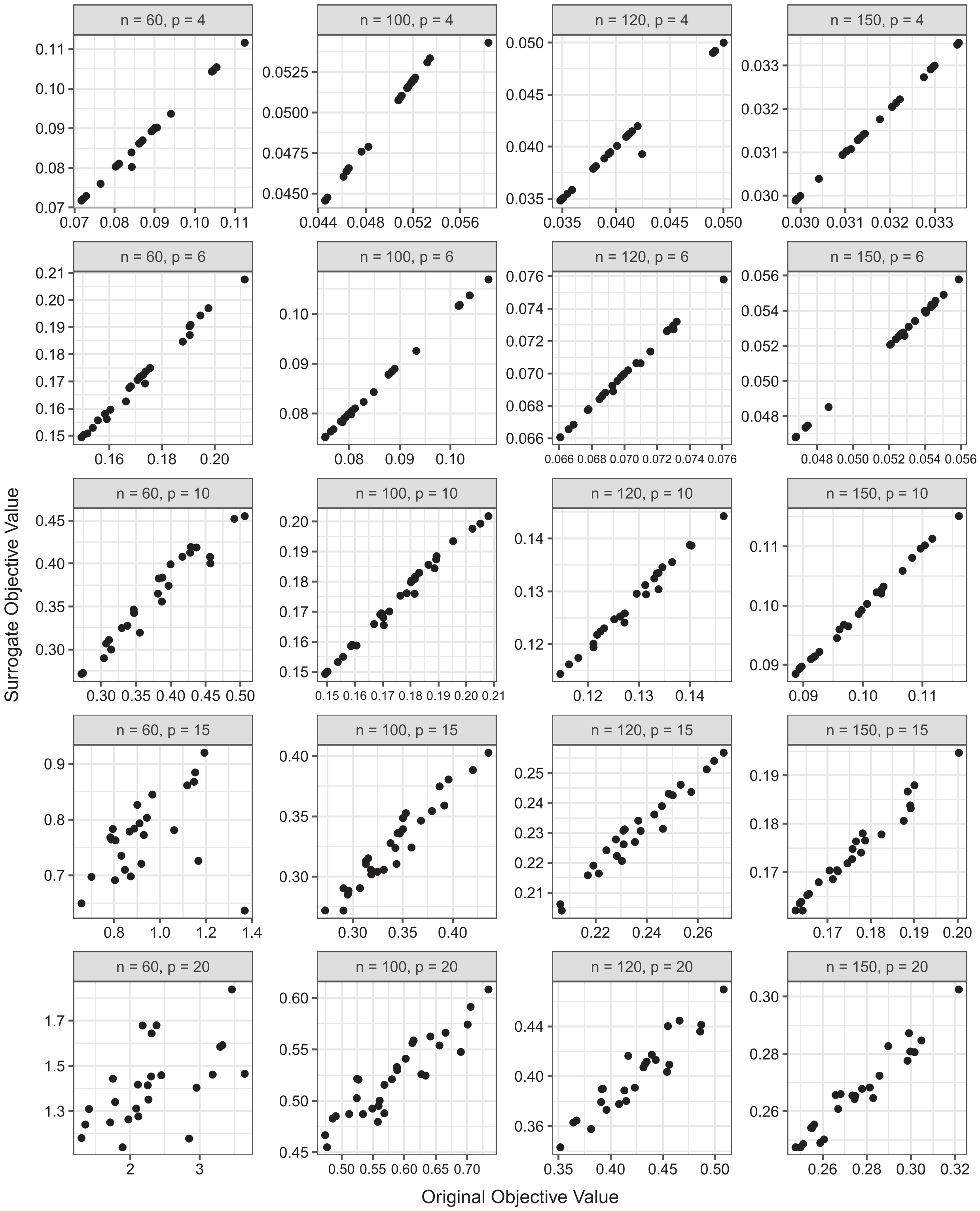
We now evaluate the variance reduction measure in (18) of optimal designs with respect to the mean variance from random designs in Figure 7. According to

Figure 4. (Color online) Objective Values of the Surrogate Model with $n = 300$



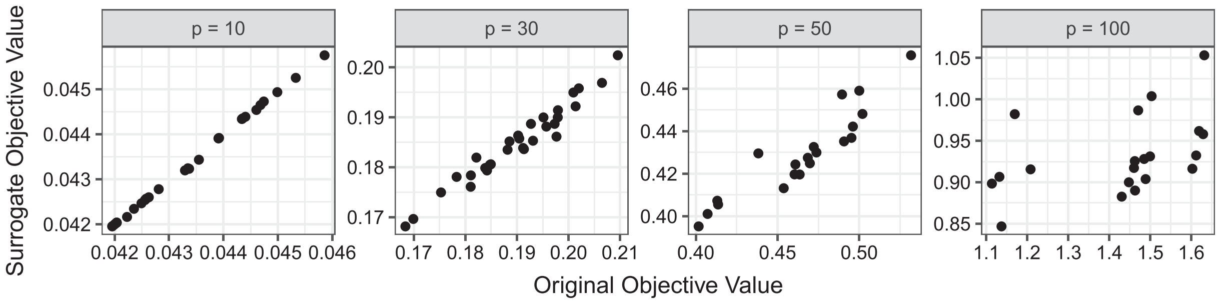
Note. Each line represents the results from one realization of H.

Figure 5. Original Objective Value Verse Surrogate Objective Value for the Instances in Figure 1



the results in Figures 1–4, we see that the performances with different realizations are consistent. To make the comparison concise, we use the same realization for each p and n in Figure 7 for all approaches. By comparing (16) and (17) over 1,000 randomly generated patient instances $\mathbf{z}_0$, almost 100% patients achieve smaller variance of the estimated interaction effect by using the optimal design from LB_APPROX. As shown in the bottom panel of Figure 7, the percentage of variance reduction ranges from 5%–50% for

Figure 6. Original Objective Value Verse Surrogate Objective Value for the Instances in Figure 3



different individuals. The variance reduction results from ADDITIVE are shown in the top panel of Figure 7; for most cases, the median variance reduction is around zero, which indicates that optimal design from the ADDITIVE approach is not necessarily leading to variance reduction compared with the random design. Over different cases, 50%–90% of patient covariates achieve a smaller variance of the estimated interaction effect by using our proposed optimal design compared with random design, whereas the ADDITIVE approach usually leads to a variance reduction for less than 50% of the patients. This comparison demonstrates the value of adding the interaction between treatment and covariates in optimal designs for precision medicine clinical trials.

We next compare different approaches by the probability of correct selection for each individual. Given the covariates information $\mathbf{z}_0$, $\mathbf{z}_0^\top\hat{\boldsymbol\beta}(\mathbf{x},H)$ follows a normal distribution with mean $\mathbf{z}_0^\top\boldsymbol\beta$ and variance $\mathbf{z}_0^\top\Sigma_\beta(\mathbf{x}^*,H)\mathbf{z}_0$. Then the probability of correct selection can be expressed by

$$P(\hat{x}(\mathbf{z}_0)=x(\mathbf{z}_0)) = \begin{cases} P\left(\mathbf{z}_0^\top\hat{\boldsymbol\beta}\,(\mathbf{x},H)\ge 0\right) & \text{if} \quad x(\mathbf{z}_0)=1\\ P\left(\mathbf{z}_0^\top\hat{\boldsymbol\beta}\,(\mathbf{x},H)\le 0\right) & \text{if} \quad x(\mathbf{z}_0)=-1\end{cases}$$

$$= \Phi\left(\frac{|\mathbf{z}_0^\top\boldsymbol\beta/\sigma|}{\sqrt{\mathbf{z}_0^\top\Sigma_\beta(\mathbf{x},H)\mathbf{z}_0/\sigma}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. We can see that the probability of correct selection monotonically decreases as the variance (i.e., our original objective) increases. Also, it monotonically increases as the signal-to-noise ratio $|\mathbf{z}_0^\top\boldsymbol\beta/\sigma|$ increases.

In Figure 8, we compare different design approaches under different levels of the signal-to-noise ratio. For each level, we generate 1,000 patient covariates as described earlier and group these patients according to their quantiles of variances. To demonstrate how the results change with the signal-to-noise ratio, we generate a constant vector $\boldsymbol\beta/\sigma$ and scale the values of its entry by 0.01, 0.05, and 0.1 to produce different levels of signal-to-noise ratio. We see from Figure 8 that as either the signal-to-noise ratio or the sample sizes n increases, the probability of correct selection increases for all approaches. The advantage of the proposed optimal design approach (LB_APPROX) compared with the ADDITIVE is more significant when the signal-to-noise ratio is small and/or the sample size is small. These results reflect the value of modeling the interaction terms

Figure 7. Percentage of Variance Reduction of the Design Attained from Different Optimal Design Approaches with Respect to the Mean of the Random Designs

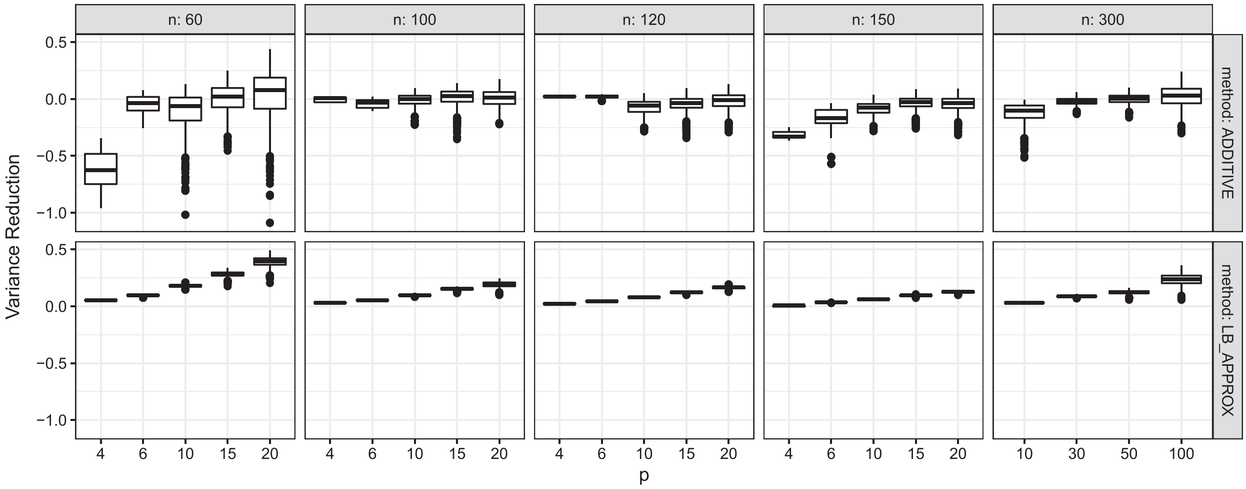
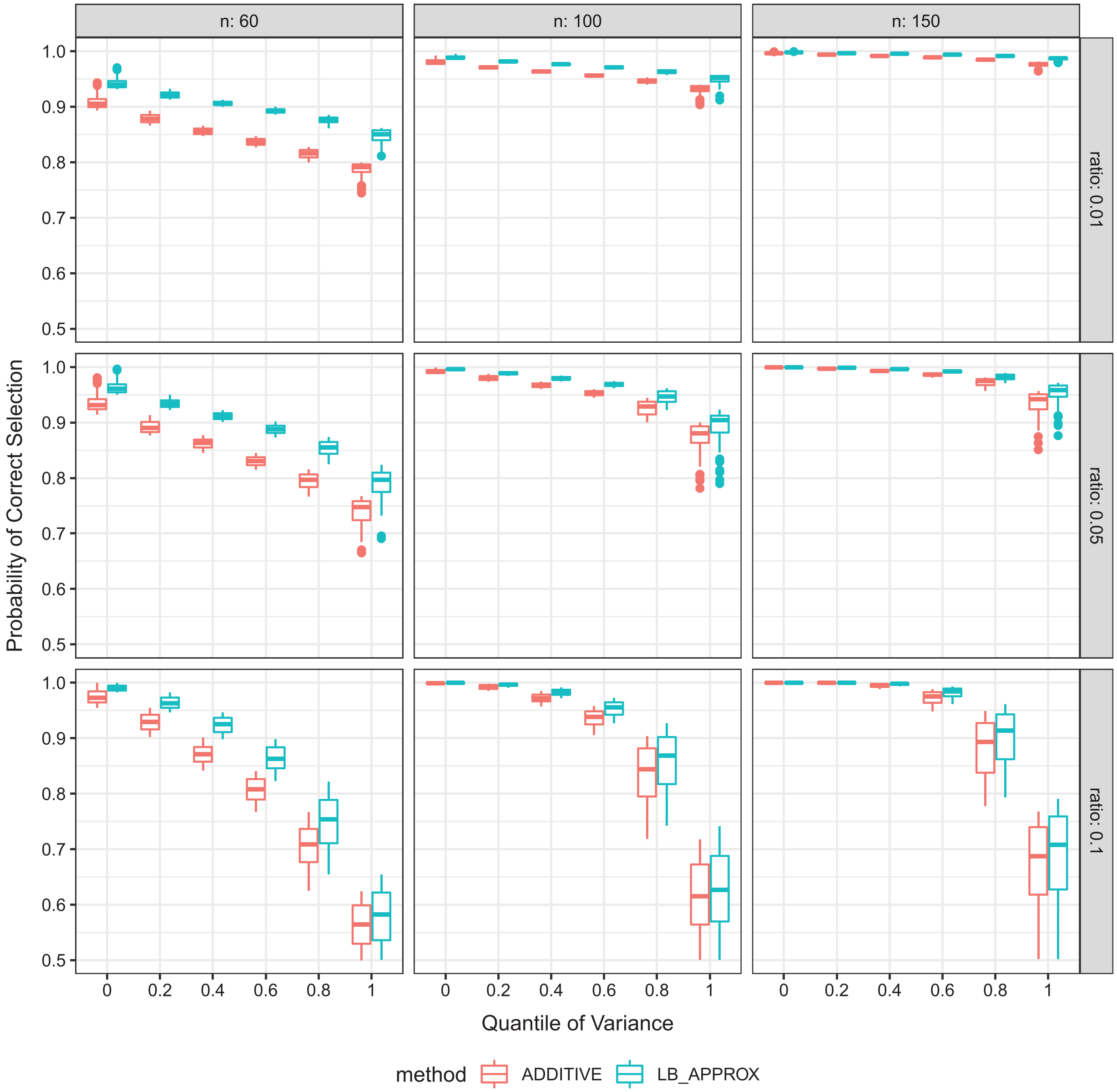


Figure 8. (Color online) Probability of Correct Selection of Different Design Approaches for the Cases with $p = 20$



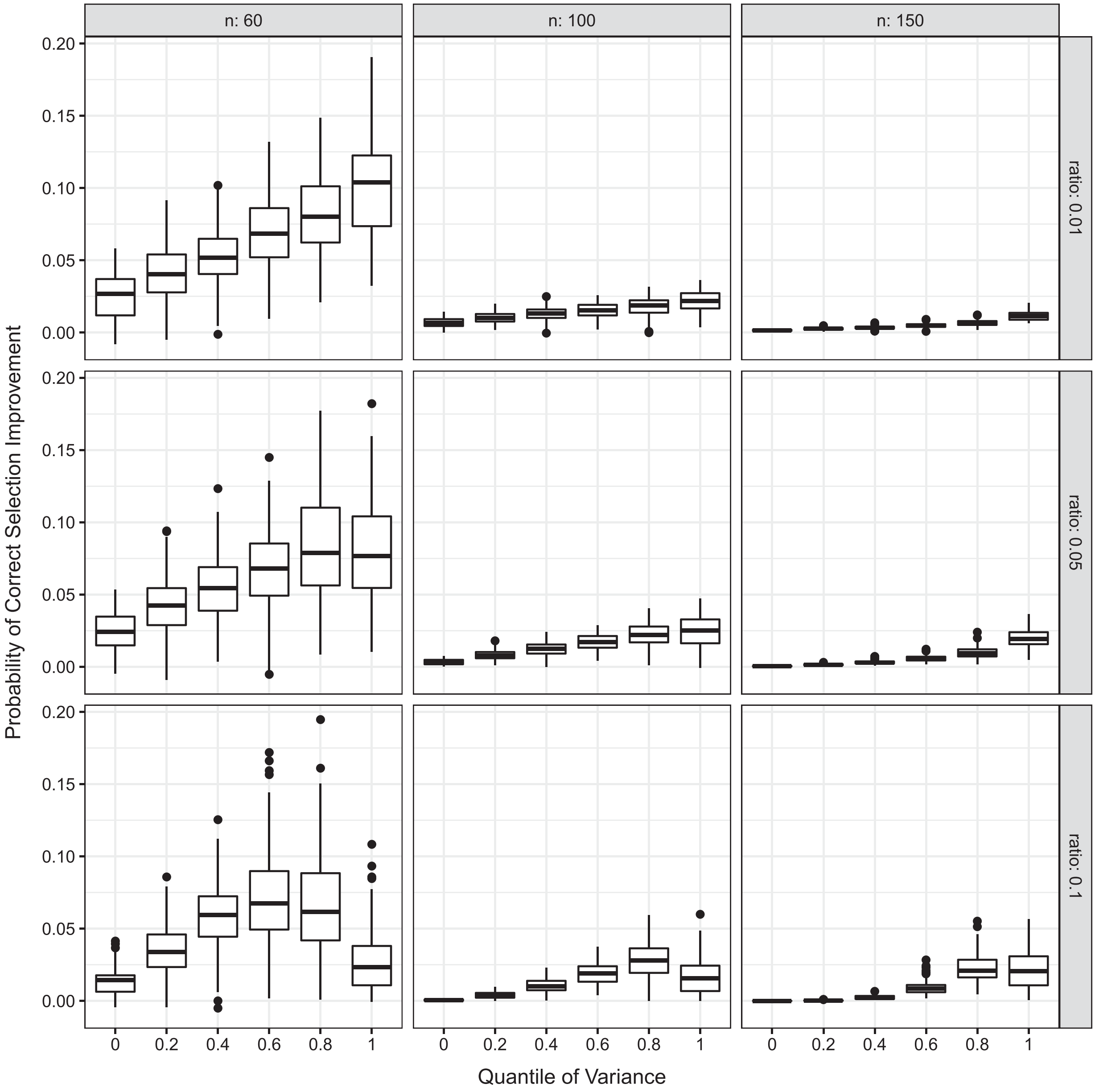
between treatment effect and patient covariates in precision medicine clinical trial designs, especially in cases when the variance tends to be large because of a small signal-to-noise ratio and/or small sample size.

To make this comparison more explicit, in Figure 9, we depict the improvement in probability of correct selection obtained by LB_APPROX compared with ADDITIVE:

$$\frac{\Phi\left(\frac{|\mathbf{z}_0^\top\boldsymbol{\beta}/\sigma|}{\sqrt{\mathbf{z}_0^\top\Sigma_\beta(\mathbf{x}^I,H)\mathbf{z}_0/\sigma}}\right)-\Phi\left(\frac{|\mathbf{z}_0^\top\boldsymbol{\beta}/\sigma|}{\sqrt{\mathbf{z}_0^\top\Sigma_\beta(\mathbf{x}^A,H)\mathbf{z}_0/\sigma}}\right)}{\Phi\left(\frac{|\mathbf{z}_0^\top\boldsymbol{\beta}/\sigma|}{\sqrt{\mathbf{z}_0^\top\Sigma_\beta(\mathbf{x}^A,H)\mathbf{z}_0/\sigma}}\right)},$$

where $\mathbf{x}^I$ represents the optimal design resulting from our proposed model that includes the interaction term between treatment and patient covariates (i.e., LB_APPROX) and $\mathbf{x}^A$ represents the optimal design resulting from ADDITIVE. We see from Figure 9 that LB_APPROX achieves improvement in the percentage of correct selection with respect to ADDITIVE for over 98% and 80% of patient covariates, respectively. We also see that the improvement is more significant for covariates corresponding to higher variances than the ones corresponding to lower variances. This phenomenon is attributed to the min-max optimal design formulation, which targets the worst-case covariates in terms of their corresponding variances. The superiority of the performance of LB_APPROX throughout this section highlights the value of incorporating the treatment-covariate interaction information into the model.

Figure 9. Improvement in the Probability of Correct Selection of the Optimal Design (LB_APPROX) Over the Model Without Interaction (i.e, ADDITIVE) for $p = 20$



5. Case Study

Warfarin is an anticoagulant medication, which is used to treat blood clots. In the United States, more than 30 million patients were prescribed warfarin in 2010 (ClinCalc 2016). However, taking an incorrect dose of warfarin can cause significant adverse effects (Wysowski et al. 2007). Therefore, there has been significant interest from the medical community to improve dose prescription strategies by taking patients' covariates into account. In particular, the International Warfarin Pharmacogenetics Consortium collected clinical and genetic data from 5,700 patients who were treated with warfarin (Pharmacogenetics Consortium 2009). This data set was used to design a personalized dosing algorithm and it is publicly available. Their analysis shows that the following covariates are significant: age, height, weight, race, use of enzyme inducers, use of amiodarone, VKORC1, and CYP2C9. Specifically, the VKORC1 gene provides the instructions to produce an enzyme that activates clotting proteins, and the CYP2C9 gene provides the instructions to produce an enzyme that helps protein processing. This result confirms that genetic factors can play a notable role in optimal warfarin dosage (White 2010).

In our case study, we consider the optimal design of the aforementioned trial retrospectively. That is, if the decision makers were to design the trial with the covariates that they observed, what would have been the optimal way? Recall that our goal in the clinical trial optimal design is to gain the maximum improvement on the statistical power of the best personalized treatment identification for a large variety of patients with heterogeneous covariate information, which can potentially be achieved by reducing the variance of the estimates of individualized treatment effects. We use this data set to test different policies for optimal design purposes, but it does not necessarily mean that the trial was designed based on optimal design principles. Note that in the case study, there were three dosages: low (≤21 mg per week), medium (>21 and <49 mg per week), and high (≥49 mg per week). Because we only consider two levels, we extract the data for patients that had a low and high prescription. In addition, the end point is the maximal response, which makes our framework applicable to this setting. We assume that the covariates are those that are considered significant in the literature and mentioned above. Following the results given by Pharmacogenetics Consortium (2009), age is categorized to nine groups ([10,20), [20,30), … , [90,-)), height to three groups ([0, 160), [160, 180), [180, -)), weight to three groups ([0,60), [60, 90), [90, -)), race to four groups (White, Asian, Black, and others), use of enzyme inducer is binary (Yes, No), use of amiodarone is binary (Yes, No), VKORC1 to three groups (A/A, A/G, G/G), and CYP2C9 to six groups (*1/*1, *1/*2, *1/*3, *2/*2, *2/*3, *3/*3). By excluding the patients with missing/censored data, we have the data for 1,476 patients and 21 covariates.

For this large data set, the EXACT algorithm is computationally intractable; therefore, we only compare the performances of LB_APPROX with RAND. The results are provided in Table 1. We observe that LB_APPROX outperforms RAND in terms of both the original and the surrogate model objective value. This observation is consistent with the results shown in Section 4. With over 1,000 randomly generated patient information $\mathbf{z}_0$, LB_APPROX achieves variance reduction for all 1,000 patients; the percentage of variance reduction of LB_APPROX ranges from 1%–8%. This modest variance reduction is somewhat expected, because when n is much larger than p, as is the case here, random design can already perform well in fitting the linear Model (1), which does not leave much room for further improvement.

Table 1. Objective Values of the Real Data Set with 1,476 Patients

Method	Original objective value	Surrogate objective value
LB_APPROX	0.8265 ([b]1%)	0.8265 ([b]1%)
RAND (1%[a])	0.8316	0.8314
RAND (5%)	0.8340	0.8337
RAND (50%)	0.8522	0.8502

[a] The percentage in the parenthesis refers to the percentile among all RAND solutions.

[b] The 1% indicates that the objective value is smaller than the 1% percentile among the 100 RAND solutions, that is, it gives the smallest objective.

Table 2. Objective Values of the Real Data Set with 100 Patients

Method	Original objective value	Surrogate objective value
EXACT	6.9999 ([b]1%)	6.8642 ([b]1%)
LB_APPROX	7.1715 ([b]1%)	7.1396 (1%)
RAND (1%[a])	7.5634	7.2928
RAND (5%)	7.8390	7.4735
RAND (50%)	9.2347	8.2819

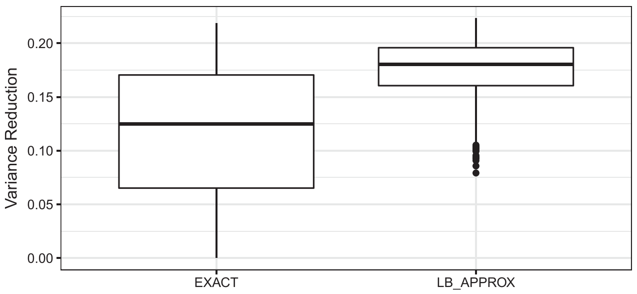
[a] The percentage in the parenthesis refers to the percentile among all RAND solutions.

[b] The 1% indicates that the objective value is smaller than the 1% percentile among the 100 RAND solutions, that is, it gives the smallest objective.

To evaluate the performance of EXACT and compare it with alternative approaches on this real data set, we truncate the problem size by randomly selecting 100 patients to conduct a relatively small-scale experiment. To ensure that there is no numerical issue in evaluating the true objective, we only include 17 columns of the covariates matrix out of 21 in this experiment. The results are given in Table 2. The results show that the EXACT algorithm outperforms other algorithms in terms of both the original model and the surrogate model objective value. The LB_APPROX algorithm is located at or under the 1% quantile among the objective values generated from 100 random designs.

With over 1,000 randomly generated patient information $\mathbf{z}_0$, LB_APPROX reduces the variance of all

Figure 10. Percentage of Variance Reduction of the Design Attained from LB_APPROX and EXACT with Respect to the Mean of the Random Designs for the Real Example with $n = 100$



1,000 patients with respect to the mean variance, whereas EXACT reduces the variance of 896 patients with respect to the mean variance. We show the boxplot of the percentage of variance reduction for both LB_APPROX and EXACT in Figure 10.

6. Conclusion

This study introduced a novel model to incorporate patient covariates into treatment effect as significant evidence is established in precision medicine literature that patients may respond differently to a treatment. We studied the optimal design of two-armed clinical trials using the introduced model, which helps practitioners design clinical trials that more accurately estimate treatment effects. Our extended model posed significant challenges in the optimization problems that emanated from optimal design of such experiments, which has optimization over design and patient covariates simultaneously. In particular, we minimized (over design) the maximum (over patient covariates) variance of the estimated individualized treatment effect, which is a min-max bilevel mixed-integer nonlinear program. We proposed a solution methodology by replacing the variance of the estimated individualized treatment effect with its natural approximation, motivated by asymptotically balanced trials. We proposed an exact algorithm to solve the surrogate optimization problem via reformulation and decomposition techniques. In addition, we created a lower bound for the inner optimization problem and solved the outer optimization over the lower bound. We tested our algorithms on hypothetical and real-world data sets. Our numerical analysis concluded the following insights: (1) The quality of approximation for the surrogate objective function is high if the number of covariates is small compared with the total number of patients in the trial, that is, low-dimensional settings. (2) Our proposed algorithms outperformed the standard (re)randomization techniques used in the optimal design literature. Our result echoes that of Bertsimas et al. (2015), which showed the power of optimization over randomization in a different optimal design setting. Through the comparison with optimal design from the model in Bhat et al. (2020), we also demonstrated the value of including interaction between treatment and patient covariates. (3) Our lower bounding algorithm produced high-quality solutions with respect to the surrogate and original objective function across all settings. This observation suggests that the lower bounding algorithm, which is easy to implement via off-the-shelf solvers, can be used instead of the proposed exact algorithm in practice.

Furthermore, our modeling approach to incorporate patient covariates into treatment effect can be generalized to other settings. In particular, our framework generalizes the A-B testing framework in Bhat et al. (2020) in incorporating side information into treatment effect. Therefore, it can be used for other applications, such as e-commerce, on-line advertising, and assortment, where one seeks to investigate covariate-dependent treatment effect. Furthermore, our framework is flexible in incorporating a variety of operational constraints. Specifically, one can add constraints on design variables **x** into the outer optimization problem and all of our methodology still holds. This is an appealing feature because practitioners may face some limitations in the design phase upfront, and our methodology can provide robust solutions in those settings. We note, however, that these constraints should not invalidate the balanced structure of the design as the approximation technique that we employ may no longer hold without this structure.

Finally, we remark on some future research directions. First, it is important to consider how to extend the optimal design of two-armed trials to multiarm trials. By incorporating more than two treatments in experimental design, the resulting optimal design objective and decisions will be different from the ones that we present in this paper. It is desirable to investigate new optimization approaches to tackle this challenge. Second, it is of interest to theoretically investigate the range of n and p such that our proposed approximation is near optimal. This can be achieved by, for example, creating a theoretical lower bound on the objective function, which we leave for future study. Third, it is interesting to extend the framework for clinical trials where the objective of the decision maker is to identify the minimum effective dose, maximum tolerable dose, or 95% effective dose, which is considered in Phase II clinical trials.

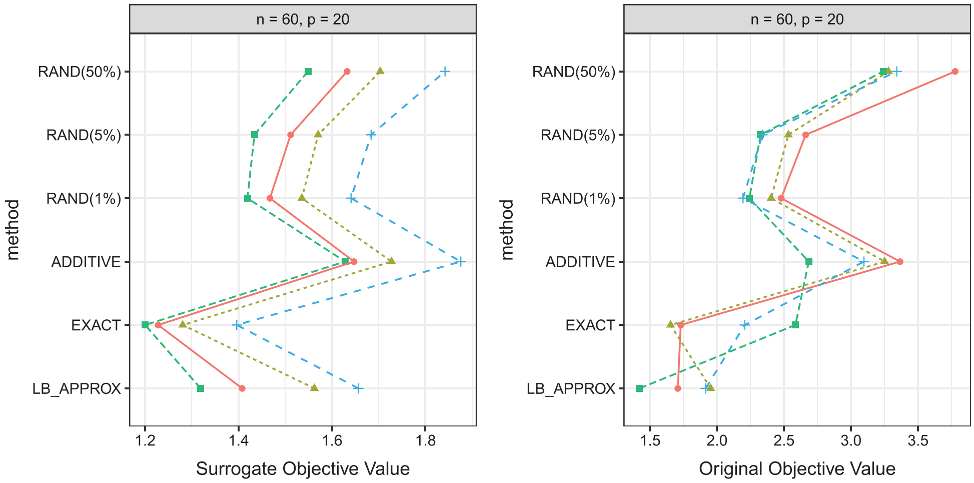
Acknowledgments

The authors acknowledge the valuable comments from the editors and three anonymous reviewers.

Appendix. Additional Numerical Comparison

We first provide the numerical results for the case with covariates information generated as continuous variables. Specifically, the first column of H is loaded by ones, and we randomly generate the entries of the remaining $p-1$ columns as independent standard normal random variables. The remaining experimental settings are the same as in Section 4.1. The results are provided in Figure A.1, which demonstrate that although the lower bound of LB_APPROX is developed based on the assumption that the covariates are discrete, the performance of this approach on continuous covariates is still superior compared with ADDITIVE and RAND designs.

Figure A.1. (Color online) Objective Values of Different Methods with Continuous Covariates Matrix



References

Atkinson A (2015) Optimum designs for two treatments with unequal variances in the presence of covariates. *Biometrika* 102(2):494–499.

Berry DA (2006) Bayesian clinical trials. *Nature Rev. Drug Discovery* 5(1):27–36.

Berry DA, Mueller P, Grieve AP, Smith M, Parke T, Blazek R, Mitchard N, Krams M (2002) Adaptive Bayesian designs for dose-ranging drug trials. Gatsonis C, Kass RE, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M, eds. *Case Studies in Bayesian Statistics*, Lecture Notes in Statistics, vol. 162 (Springer, New York), 99–181.

Bertsimas D, Johnson M, Kallus N (2015) The power of optimization over randomization in designing experiments involving small samples. *Oper. Res.* 63(4):868–876.

Bhat N, Farias VF, Moallemi CC, Sinha D (2020) Near optimal A-B testing. *Management Sci.* 66(10):4477–4495.

ClinCalc (2016) Number of prescriptions over time. Accessed March 1, 2021, https://clincalc.com/DrugStats/Drugs/Warfarin.

DeNegre S (2011) Interdiction and discrete bilevel linear programming. Unpublished doctoral thesis, Lehigh University, Bethlehem, PA.

Fisher RA (1936) Design of experiments. *BMJ* 1(3923):554–554.

Giffin RB, Lebovitz Y, English RA, et al (2010) *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary* (National Academies Press, Washington, DC).

Hayden E (2015) California unveils "precision-medicine" project. *Nature News*. https://www.nature.com/news/california-unveils-precision-medicine-project-1.17324.

Kallus N (2018) Optimal a priori balance in the design of controlled experiments. *J. Roy. Statist. Soc. Statist. Methodology Ser. B* 80(1):85–112.

Kosorok MR, Laber EB (2019) Precision medicine. *Annual Rev. Statist. Appl.* 6:263–286.

Kotas J, Ghate A (2018) Bayesian learning of dose–response parameters from a cohort under response-guided dosing. *Eur. J. Oper. Res.* 265(1):328–343.

Laber EB, Zhao YQ, Regh T, Davidian M, Tsiatis A, Stanford JB, Zeng D, Song R, Kosorok MR (2016) Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Statist. Medicine* 35(8):1245–1256.

Le Tourneau C, Lee JJ, Siu LL (2009) Dose escalation methods in phase I cancer clinical trials. *J. National Cancer Inst.* 101(10): 708–720.

Majewski IJ, Bernards R (2011) Taming the dragon: Genomic biomarkers to individualize the treatment of cancer. *Nature Medicine* 17(3):304–312.

Morris M, Dean A, Stufken J, Bingham D (2015) History and overview of design and analysis of experiments. *Handbook of Design and Analysis of Experiments* (Chapman and Hall/CRC Press, Boca Raton, FL), 21–80.

Pharmacogenetics Consortium (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England J. Medicine* 360(8):753–764.

Press WH (2009) Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proc. Natl. Acad. Sci. USA* 106(52):22387–22392.

Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. *Ann. Statist.* 39(2):1180.

Rencher AC, Schaalje GB (2008*) Linear Models in Statistics* (John Wiley & Sons, New York).

Schork NJ (2015) Personalized medicine: Time for one-person trials. *Nature* 520(7549):609–611.

Sertkaya A, Wong HH, Jessup A, Beleche T (2016) Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials* 13(2):117–126.

Shi C, Song R, Lu W, Fu B (2018) Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *J. Roy. Statist. Soc. Statist. Methodology Ser. B* 80(4):681–702.

Singh M, Xie W (2020) Approximation algorithms for D-optimal design. *Math. Oper. Res.* 45(4):1512–1534.

Tang Y, Richard JPP, Smith JC (2016) A class of algorithms for mixed-integer bilevel min–max optimization. *J. Global Optim.* 66(2):225–262.

Tufts (2014) Cost to develop and win marketing approval for a new drug is $2.6 billion. Accessed March 1, 2021, http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study.

White PJ (2010) Patient factors that influence warfarin dose response. *J. Pharmacy Practice* 23(3):194–204.

Wu CJ, Hamada MS (2011) *Experiments: Planning, Analysis, and Optimization*, vol. 552 (John Wiley & Sons, New York).

Wysowski DK, Nourjah P, Swartz L (2007) Bleeding complications with warfarin use: A prevalent adverse effect resulting in regulatory action. *Arch. Internal Medicine* 167(13):1414–1419.