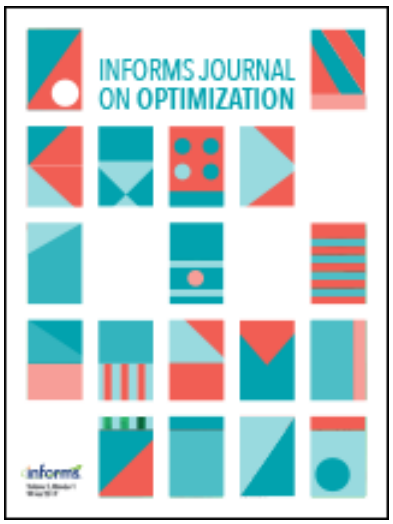


INFORMS Journal on Optimization

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

Identifying Exceptional Responders in Randomized Trials: An Optimization Approach

Dimitris Bertsimas, Nikita Korolko, Alexander M. Weinstein

To cite this article:
Dimitris Bertsimas, Nikita Korolko, Alexander M. Weinstein (2019) Identifying Exceptional Responders in Randomized Trials: An Optimization Approach. INFORMS Journal on Optimization 1(3):187-199. https://doi.org/10.1287/ijoo.2018.0006

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

INFORMS JOURNAL ON OPTIMIZATION
Vol. 1, No. 3, Summer 2019, pp. 187–199
ISSN 2575-1484 (print), ISSN 2575-1492 (online)

Identifying Exceptional Responders in Randomized Trials: An Optimization Approach

Dimitris Bertsimas,[a] Nikita Korolko,[a] Alexander M. Weinstein[a]

[a] Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contact: dbertsim@mit.edu, https://orcid.org/0000-0002-1985-1003 (DB); nikita.korolko@alum.mit.edu, https://orcid.org/0000-0001-8306-2271 (NK); amw22@alum.mit.edu, https://orcid.org/0000-0002-1283-6329 (AMW)

Received: January 4, 2018
Revised: June 16, 2018
Accepted: August 19, 2018
Published Online in Articles in Advance: April 16, 2019

https://doi.org/10.1287/ijoo.2018.0006

Abstract. In randomized clinical trials, there may be a benefit to identifying subgroups of the study population for which a treatment was exceptionally effective or ineffective. We present an efficient mixed-integer optimization formulation that can directly find an interpretable subset with maximum (or minimum) average treatment effect. Using both simulated and real data from randomized trials, we demonstrate the effectiveness and stability of the optimization approach in identifying subsets with exceptional response and verifying their statistical significance.

History: This article has been accepted for the INFORMS Journal on Optimization Special Issue on Machine Learning and Optimization.
Brian Denton served as Editor-in-Chief for this article.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/ijoo.2018.0006.

Keywords: fractional programming • mixed-integer optimization • randomized trials • subgroup identification

1. Introduction

Researchers in the medical and social sciences invest substantial resources to implement randomized controlled trials, the gold standard in statistical analysis of response to a treatment. Whether the response measured is biological, economic, social, or otherwise, the hope of randomized trials is to confirm the effectiveness (or harm) of an experimental intervention. When a trial fails to yield a significant result, the investigator may abandon the study of the intervention entirely, despite the initial investment made. In this paper, we present a method that uses optimization to identify subgroups for which an exceptionally large positive or negative response was found.

Such a method could provide great value in the pharmaceutical industry. For instance, in 2010, the expenditure of global pharmaceutical companies on clinical trials for investigational drugs was $32.5 billion, due to multiple phases of clinical trials necessary for a drug approval process that typically take years to complete (Berndt and Cockburn 2013, U.S. Food and Drug Administration 2015). It is estimated that from 2003 to 2011, 60% of Phase III clinical trials for investigational drug indications led to submission of a new drug application or biologic license application to the U.S. Food and Drug Administration, of which 83% proceeded to approval (Hay et al. 2014). Our proposed method could suggest promising subpopulations in which to conduct (or avoid) further testing of a treatment. In this way, there is the potential to increase the value of research and development dollars and realize large-scale economic benefits throughout the healthcare industry. Investigators could also use our method to revisit long-terminated clinical trials and search for opportunities to revive the testing of failed drugs in promising subgroups.

This potential to enhance the value of clinical trials through subgroup identification may exist even for trials that were initially successful. Subgroup identification could point investigators to the prevalence of adverse events arising from the use of new or existing drugs in subpopulations. For approved drugs whose patents are expiring, our method may increase the financial benefits of companies' initial research and development investments by suggesting opportunities for remarketing the drugs to different segments of the population or for different medicinal purposes.

The classical statistical approach to identifying subgroups with distinct treatment effects involves using a Cox proportional hazard model with treatment–covariate interaction terms (Schemper 1988). Citing the disadvantage of needing to specify relevant interactions, Kehl and Ulm (2006) improve on this model by incorporating a bump-hunting procedure based on Friedman and Fisher (1999), which they also contrast with the greedier approach of regression trees (Breiman et al. 1984). Foster et al. (2011) introduce several variations of a "virtual twins" method, based on counterfactual modeling and the application of machine learning approaches, including logistic regression, random forest, and classification and regression trees (CARTs) (Breiman et al. 1984,

Breiman 2001). This virtual twins method is shown to be effective in simulation studies, with positive predictive values (PPVs) ranging from 45% to 60%. Others (Su et al. 2009, Hardin et al. 2013) have presented CART-inspired recursive partitioning approaches to solve the subgroup identification or partitioning problem. Su et al. (2009) show their method is effective in identifying subgroups with significant positive or negative treatment effects from observational data. The Hardin et al. (2013) approach has been used in practice to identify subgroups of patients with type 2 diabetes mellitus for which short-acting insulin may provide a benefit. These bump-hunting and tree-based approaches can identify satisfyingly interpretable subgroups, but they entail greedily solving a sequence of optimization problems.

We present a mixed-integer optimization (MIO) approach to identify a subgroup for which the average treatment effect (ATE) was exceptionally strong or exceptionally weak and that can be defined by a small prespecified number of covariates. When a randomized clinical trial is unable to reject the null hypothesis of no effect, our method may identify a subset for which the treatment was effective, or provide evidence that there is no such subgroup. Even when the randomized clinical trial is conclusive in confirming the effectiveness or harm of a treatment, our method can identify subgroups for which the treatment was particularly effective or particularly ineffective. With the power of MIO, we can find optimal interpretable solutions directly by solving a single global formulation instead of using a greedy, recursive approach.

Whereas we analyze data from randomized trials, MIO has also been used in the related setting of observational studies, where treatment allocation is not determined by the study administrator, but rather by (potentially nonrandom) conditions present in observed data (Zubizarreta 2012, Sauppe and Jacobson 2017). In these studies, MIO is used to match subjects in a treatment group to subjects in a control group to make causal inferences. The goal of the matching is to minimize the distance between empirical distributions of subjects in the two groups, where this distance may be expressed in terms of univariate moments (such as means, variances, and skewness), multivariate moments (such as correlations between covariates), or differences in quantiles and statistics (such as the Kolmogorov–Smirnov statistic). Another approach for observational studies is the balance optimization subset selection (BOSS) model (Nikolaev et al. 2013). In this setting, the matching optimization problem of treatment–control pairs is replaced by a search for a subset of subjects in a control group that features a desired level of balance on the covariate distributions for the treatment and control groups. The matching and BOSS model approaches can be merged by virtue of MIO (Sauppe et al. 2014).

Our study is focused on the randomized trial setting. In Section 2, we formally describe the problem of finding an interpretable subset with optimal treatment response and introduce an explicit optimization formulation with a fractional objective function. We show that the fractional problem can be transformed into a tractable and efficient MIO formulation with $\mathcal{O}(n^2)$ continuous variables and $\mathcal{O}(n)$ binary variables, where n is the number of trial subjects. We describe an algorithm that modifies the MIO formulation for the setting when one wants to find multiple subsets with exceptional treatment response sequentially, such that there is limited intersection between the subsets. We also introduce a simple tree-based heuristic that finds near-optimal solutions instantaneously and can be used in practice to provide feasible warm-start solutions for the MIO.

In Section 3, we present simulation experiments in which the MIO approach is used to identify optimal interpretable subgroups. We evaluate the effectiveness of the algorithm in terms of the rates of identifying true positive and false positive subsets. In Section 4, we apply the MIO approach to data sets from three real randomized controlled trials. In two examples, the method identifies a subset with a statistically significant exceptional response. In the third example, the method finds no subset that has a statistically significant response. Finally, in Section 5, we share some concluding remarks.

2. Identifying Interpretable Optimal Subgroups

In our problem setting, we are given randomized controlled trial data $\mathcal{D} := \{X, \mathbf{T}, \mathbf{v}\}$ for n trial subjects, with covariate matrix $X \in \mathbb{R}^{n \times S}$, binary treatment vector $\mathbf{T} \in \{0,1\}^n$, and treatment response vector $\mathbf{v} \in \mathbb{R}^n$. There are S covariates, or variables of interest, for which we have a complete vector of observations $\mathbf{x}_i$ for each subject $i = 1, \ldots, n$, as represented in the rows of matrix X. Each subject $i = 1, \ldots, n$ has received a treatment assignment to one of two treatment conditions, treatment ($T_i = 1$) or control ($T_i = 0$), and we have observed the subject's response v_i to her respective treatment.

Our objective is to identify a box $\mathcal{B} \subseteq \mathbb{R}^S$ that maximizes, or minimizes, the ATE among subjects whose covariate vectors are contained in the box. We define a box as a (potentially unbounded) subset of the covariate space formed by the intersection of a finite number of half-spaces, whose boundaries are parallel to the axes. The

ATE of the box $\mathcal{B}$ is defined as the average response for the treatment group minus the average response for the control group, when limiting to subjects whose covariate vector $\mathbf{x}_i$ is contained in $\mathcal{B}$:

$$\mathrm{ATE}(\mathcal{B}) := \frac{1}{|\mathcal{T}_1 \cap \mathcal{I}(\mathcal{B})|} \sum_{i \in \mathcal{T}_1 \cap \mathcal{I}(\mathcal{B})} v_i - \frac{1}{|\mathcal{T}_0 \cap \mathcal{I}(\mathcal{B})|} \sum_{i \in \mathcal{T}_0 \cap \mathcal{I}(\mathcal{B})} v_i, \tag{1}$$

where $\mathcal{T}_t := \{i : T_i = t\}$, for $t = 0, 1$, and $\mathcal{I}(\mathcal{B}) := \{i : \mathbf{x}_i \in \mathcal{B}\}$. For the remainder of this paper, we focus on finding $\mathcal{B}$ to maximize the ATE, and we assume that larger values of the response v_i are preferable. However, the formulation generalizes easily to minimizing the ATE.

The requirement that the optimal subset $\mathcal{B}^*$ be a box in the covariate space yields the appealing property of interpretability. Especially in healthcare applications, it is easy for a policy maker to communicate and enforce a treatment policy for a population subset that can be described by a small number of ranges of covariate values. Consider, for instance, a treatment that is exceptionally effective for adults ages 45 to 65, with height greater than six feet and eye color blue. This population subset is described by a fixed interval on age, an interval on height that is open on one end (but implicitly limited by the maximum observed human height), and a closed interval on the space of the binary variable encoding blue eye color. If the decision maker is interested in finding an optimal subset that is not fully connected, our formulation allows for identifying multiple boxes (Section 2.4).

We require the decision maker to specify several parameters that ensure the optimization formulation in Section 2.1 will produce an interpretable box $\mathcal{B}^*$:

First, the decision maker specifies, for each covariate $s = 1, \ldots, S$, a set of K_s hyperplanes orthogonal to that covariate axis. In the optimization formulation, we will select the boundaries for $\mathcal{B}^*$ from this discrete set of hyperplanes. Each hyperplane is defined by a value γ_{sk}, $s = 1, \ldots, S$, $k = 1, \ldots, K_s$, such that

$$\min_i x_{is} - \epsilon = \gamma_{s1} < \gamma_{s2} < \cdots < \gamma_{sK_s} = \max_i x_{is} + \epsilon,$$

where $\epsilon > 0$ is a small perturbation term. In practice, the locations of these hyperplanes could be chosen based on domain knowledge to yield meaningful covariate ranges. For instance, if the covariate is age and the population is human adults, the decision maker could specify hyperplanes at each multiple of 10 years from age 20 to 80. If the covariate is body mass index (BMI), the decision maker may want to specify hyperplanes at widely accepted BMI values that distinguish underweight, normal, overweight, and obese adults. Having a discrete set of choices for these interval boundaries also facilitates modeling the problem using mixed-integer optimization (Section 2.1).

Second, as mentioned above, a population subset is more interpretable if described by a relatively small number of covariate ranges. To model this constraint, we include a parameter $S_0 \leq S$ to be chosen by the decision maker as an upper limit on the number of covariate dimensions along which we can restrict the subset. This means that there can be at most S_0 dimensions for which we choose lower and/or upper bounds more restrictive than the full covariate range $[\gamma_{s1}, \gamma_{sK_s}]$. For instance, if one covariate is age for a human adult population, a decision maker could say that the full human adult population is contained in the interval [0, 125], and therefore could specify $\gamma_{\mathrm{age},1} = 0$ and $\gamma_{\mathrm{age},K_{\mathrm{age}}} = 125$. In this example, we could choose [0, 125] as the age range for the optimal box $\mathcal{B}^*$, and this would not count against the limit S_0 on restricted dimensions. Conversely, selecting intervals such as [0, 20], [100, 125], or [45, 65] would count toward the limit S_0.

Finally, the decision maker should specify upper and lower limits $\underline{N}$ and $\overline{N}$, respectively, on the cardinality of $\mathcal{I}(\mathcal{B}^*)$ such that $\underline{N} \leq |\mathcal{T}_t \cap \mathcal{I}(\mathcal{B}^*)| \leq \overline{N}$, for $t = 0, 1$. These parameters are important because, if the box is too large, then it may not be sufficiently distinct from the full population. If the box is too small, it may define a population subset too small to recommend actionable policies.

We include a section on practical application of the optimal subset identification approach in the online appendix.

2.1. Mixed-Integer Optimization Approach

We formulate the problem of finding the optimal box $\mathcal{B}^*$ as a mixed-integer optimization formulation. The approach is to choose subset boundaries from the discrete set of prespecified hyperplanes to maximize the ATE in the box defined by those boundaries. We include interpretability constraints on both the number of restricted dimensions S_0 and the cardinality of $\mathcal{I}(\mathcal{B}^*)$.

We introduce binary decision variables to model the choice of subset boundaries. Let $\mathbf{L} := \{L_{sk} \,|\, s = 1, \ldots, S; k = 1, \ldots, K_s\}$ be a set of binary decision variables that take value 1 if γ_{sk} is chosen as the lower subset bound for

dimension s and 0 otherwise. Similarly, let $\mathbf{U} := \{U_{sk} \mid s = 1, \ldots, S; k = 1, \ldots, K_s\}$ be a set of binary decision variables that take value 1 if γ_{sk} is chosen as the upper subset bound for dimension s and 0 otherwise.

We introduce auxiliary binary decision variables $z_i := \mathbb{I}\{i \in \mathcal{J}(\mathcal{B})\}$, $i = 1, \ldots, n$, that indicate whether each subject's covariates are contained in the box $\mathcal{B}$ defined by the choices of $\mathbf{L}$ and $\mathbf{U}$. To enforce the limit S_0 on restricted dimensions, we define auxiliary binary indicator variables q_s, which indicate whether covariate dimension s is used to restrict the box $\mathcal{B}$ defined by the choices of $\mathbf{L}$ and $\mathbf{U}$. Both vectors $\mathbf{z}$ and $\mathbf{q}$ are fully determined by the primary decision vectors $\mathbf{L}$ and $\mathbf{U}$, according to constraints (2b)–(2d) and (2g)–(2i), respectively.

We formulate the following fractional MIO to identify the box with the highest ATE:

$$\max_{\mathbf{z},\mathbf{q},\mathbf{L},\mathbf{U}} \quad \frac{\sum_{i\in\mathcal{T}_1} v_i z_i}{\sum_{i\in\mathcal{T}_1} z_i} - \frac{\sum_{i\in\mathcal{T}_0} v_i z_i}{\sum_{i\in\mathcal{T}_0} z_i} \tag{2a}$$

$$\text{s.t.} \quad z_i + \sum_{s=1}^{S}\left[\sum_{k:\gamma_{sk}>x_{is}} L_{sk} + \sum_{k:\gamma_{sk}<x_{is}} U_{sk}\right] \geq 1, \qquad \forall i = 1, \ldots, n, \tag{2b}$$

$$z_i + L_{sk} \leq 1, \qquad \forall s = 1, \ldots, S,\ k = 1, \ldots, K_s, \qquad i : x_{is} < \gamma_{sk}, \tag{2c}$$

$$z_i + U_{sk} \leq 1, \qquad \forall s = 1, \ldots, S,\ k = 1, \ldots, K_s, \qquad i : x_{is} > \gamma_{sk}, \tag{2d}$$

$$\sum_{k=1}^{K_s} L_{sk} = 1, \qquad \forall s = 1, \ldots, S, \tag{2e}$$

$$\sum_{k=1}^{K_s} U_{sk} = 1, \qquad \forall s = 1, \ldots, S, \tag{2f}$$

$$q_s + L_{s1} \geq 1, \qquad \forall s = 1, \ldots, S, \tag{2g}$$

$$q_s + U_{sK_s} \geq 1, \qquad \forall s = 1, \ldots, S, \tag{2h}$$

$$q_s + L_{s1} + U_{sK_s} \leq 2, \qquad \forall s = 1, \ldots, S, \tag{2i}$$

$$\sum_{s=1}^{S} q_s \leq S_0, \tag{2j}$$

$$\underline{N} \leq \sum_{i\in\mathcal{T}_t} z_i \leq \overline{N}, \qquad \forall t = 0, 1, \tag{2k}$$

$$\mathbf{z}, \mathbf{q}, \mathbf{L}, \mathbf{U} \in \{0, 1\}.$$

The objective function (2a) in formulation (2) is equivalent to the ATE of the box uniquely determined by variables $\mathbf{L}$ and $\mathbf{U}$, where the definition of the ATE is given by expression (1). As mentioned earlier, we choose $\mathbf{L}$ and $\mathbf{U}$ to maximize this objective function, but the formulation (2) also permits solution as an ATE minimization problem by simply changing the maximization in the objective function to a minimization.

The constraints in formulation (2) deserve further discussion. First, for any given subject i, if the condition

$$\sum_{s=1}^{S}\left[\sum_{k:\gamma_{sk}>x_{is}} L_{sk} + \sum_{k:\gamma_{sk}<x_{is}} U_{sk}\right] = 0 \tag{3}$$

is met, then z_i must be equal to 1, by constraints (2b). We observe that condition (3) is met if and only if, for *all* covariates $s = 1, \ldots, S$, both of the following statements are true:

1. by our choice of $\mathbf{L}$, we do *not* select any lower bound γ_{sk} for which $\gamma_{sk} > x_{is}$, and
2. by our choice of $\mathbf{U}$, we do *not* select any upper bound γ_{sk} for which $\gamma_{sk} < x_{is}$,

where x_{is} is the value of covariate s for subject i. Taken together, these statements imply that subject i must be in the box defined by the choices of $\mathbf{L}$ and $\mathbf{U}$. Therefore, by construction, the auxiliary variable z_i should be equal to 1 in this case.

Conversely, constraints (2c) and (2d) ensure that z_i must be equal to 0 whenever the choices of $\mathbf{L}$ and $\mathbf{U}$ imply that subject i is outside the box. Specifically, if we select γ_{sk} as a lower bound on dimension s by taking $L_{sk} = 1$, then by constraints (2c) we have $z_i = 0$ for all subjects i for which $x_{is} < \gamma_{sk}$. If we select γ_{sk} as an upper bound by taking $U_{sk} = 1$, then by constraints (2d) we have $z_i = 0$ for all subjects i for which $x_{is} > \gamma_{sk}$.

Constraints (2e) and (2f) indicate that only one lower and upper bound, respectively, can be chosen for each covariate dimension. Constraints (2g)–(2i) encode the desired relationships $q_s = 1 - \mathbb{I}\{L_{s1} = 1 \text{ and } U_{sK_s} = 1\}$, $s = 1, \ldots, S$, so that q_s indicates whether dimension s is used to restrict the box $\mathcal{B}$. This relationship allows us to require that at most S_0 covariate dimensions are used to restrict the subset by adding constraint (2j). Finally, constraints (2k) ensure that the cardinality of $\mathcal{J}(\mathcal{B})$ conforms to the specified limits. Note that these constraints

also ensure that there will be no dimension for which the lower bound chosen exceeds the upper bound, because choosing such bounds would make it impossible to satisfy the lower bound cardinality constraint $\underline{N}$. We discuss how the selection of parameters $\underline{N}$ and $\overline{N}$ affects the quality of the resulting box $\mathcal{B}$ in the practical considerations section of the online appendix.

The problem (2) is a typical fractional 0–1 program, and its structure is similar to those of other problems that involve averages, probabilities, and percentages (Wu 1997). A recent survey by Borrero et al. (2017) extensively describes applications, computational complexity, and exact and approximating solutions methods for fractional 0–1 programming.

2.2. Tractable Transformation of the Fractional Objective

Because formulation (2) has a fractional objective function, (2a), the problem cannot be solved using off-the-shelf commercial solvers. Developing and augmenting ideas from Wu (1997) and Tawarmalani et al. (2002) for linearization of a product of binary variables and change of variables for the denominator, we can transform the objective from fractional to nonfractional by considering the expressions

$$\Theta_t := \frac{1}{\sum_{i\in\mathcal{T}_t} z_i}, \; t = 0, 1. \tag{4}$$

By construction, Θ_t, $t = 0, 1$, are discrete variables that take values in the set $\left\{\frac{1}{\underline{N}}, \frac{1}{\underline{N}+1}, \ldots, \frac{1}{\overline{N}}\right\}$. Therefore, we can represent each discrete variable by the binary expansion

$$\Theta_t = \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j}\theta_j^{(t)}, \; t = 0, 1,$$

where $\theta_j^{(t)}$, $j = \underline{N}, \ldots, \overline{N}$, $t = 0, 1$, are binary variables with $\sum_{j=\underline{N}}^{\overline{N}} \theta_j^{(t)} = 1$, $t = 0, 1$. We can now rewrite the fractional objective function (2a) as a nonfractional, nonlinear expression:

$$\Theta_1 \sum_{i\in\mathcal{T}_1} v_i z_i - \Theta_0 \sum_{i\in\mathcal{T}_0} v_i z_i = \sum_{i\in\mathcal{T}_1} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i z_i \theta_j^{(1)} - \sum_{i\in\mathcal{T}_0} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i z_i \theta_j^{(0)}.$$

To make the objective linear, we introduce additional binary variables ζ_{ij}, $i = 1, \ldots, n$, $j = \underline{N}, \ldots, \overline{N}$. To model the desired relationship $\zeta_{ij} = z_i\theta_j^{(T_i)}$, which is the product of two binary variables, we add three sets of constraints, (5b)–(5d). We now have the linear objective

$$\sum_{i\in\mathcal{T}_1} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij} - \sum_{i\in\mathcal{T}_0} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij}.$$

Finally, to enforce the stated relationship (4), we add the constraints

$$\Theta_t \sum_{i\in\mathcal{T}_t} z_i = \sum_{i\in\mathcal{T}_t} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} \zeta_{ij} = 1, \quad \forall t = 0, 1.$$

Taking together all of these substitutions and constraints, we obtain the following mixed-binary linear optimization formulation, which is equivalent to formulation (2) and can be solved using commercial optimization solvers:

$$\max_{\mathbf{z},\mathbf{q},\mathbf{L},\mathbf{U},\boldsymbol{\zeta},\boldsymbol{\theta}} \quad \sum_{i\in\mathcal{T}_1} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij} - \sum_{i\in\mathcal{T}_0} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij} \tag{5a}$$

$$\text{s.t.} \quad z_i + \sum_{s=1}^{S} \left[\sum_{k:\gamma_{sk}>x_{is}} L_{sk} + \sum_{k:\gamma_{sk}<x_{is}} U_{sk} \right] \geq 1, \qquad \forall i = 1, \ldots, n,$$

$$z_i + L_{sk} \leq 1, \qquad \forall s = 1, \ldots, S, \; k = 1, \ldots, K_s, \qquad i : x_{is} < \gamma_{sk},$$

$$z_i + U_{sk} \leq 1, \qquad \forall s = 1, \ldots, S, \; k = 1, \ldots, K_s, \qquad i : x_{is} > \gamma_{sk},$$

$$\sum_{k=1}^{K_s} L_{sk} = 1, \qquad \forall s = 1, \ldots, S,$$

$$\sum_{k=1}^{K_s} U_{sk} = 1, \qquad \forall s = 1, \ldots, S,$$
$$q_s + L_{s1} \geq 1, \qquad \forall s = 1, \ldots, S,$$
$$q_s + U_{sK_s} \geq 1, \qquad \forall s = 1, \ldots, S,$$
$$q_s + L_{s1} + U_{sK_s} \leq 2, \qquad \forall s = 1, \ldots, S,$$
$$\sum_{s=1}^{S} q_s \leq S_0,$$
$$\underline{N} \leq \sum_{i \in \mathcal{T}_t} z_i \leq \overline{N}, \qquad \forall t = 0, 1,$$
$$\zeta_{ij} \leq \theta_j^{(T_i)}, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{5b}$$
$$\zeta_{ij} \leq z_i, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{5c}$$
$$\zeta_{ij} \geq \theta_j^{(T_i)} + z_i - 1, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{5d}$$
$$\sum_{i \in \mathcal{T}_t} \sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} \zeta_{ij} = 1, \qquad \forall t = 0, 1,$$
$$\sum_{j=\underline{N}}^{\overline{N}} \theta_j^{(t)} = 1, \qquad \forall t = 0, 1,$$
$$0 \leq \zeta_{ij} \leq 1, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N},$$
$$\mathbf{z}, \mathbf{q}, \mathbf{L}, \mathbf{U}, \boldsymbol{\theta} \in \{0, 1\}.$$

Note that $\boldsymbol{\zeta}$ can be included as continuous variables on [0, 1], which improves the computational performance, as the number of binary variables is linear in n. Thus, the formulation includes $\mathcal{O}(n^2)$ continuous variables, with $\mathcal{O}(n)$ binary variables, assuming $\sum_{s=1}^{S} K_s$ is small relative to n.

2.3. Statistical Significance of the Optimal Subset

In all tested instances, we were able to solve formulation (5) to provable optimality or to within a prespecified optimality tolerance through the use of an off-the-shelf commercial optimization solver, such as Gurobi (Gurobi Optimization, Inc. 2016). If a given instance cannot be solved to find a feasible subset that achieves the desired optimality tolerance, the decision maker could lower the resolution of the cut points $\boldsymbol{\gamma}$ and re-solve the formulation until the problem is tractable. Yet, despite the fact we can produce an optimal solution for all problems of practical size, it is unreasonable to expect that every randomized trial has a latent subset in which a significant positive or negative effect is observed.

We propose the use of statistical hypothesis testing to determine whether the average treatment effect within the optimal subgroup is statistically significant. We adopt the null hypothesis that there is no difference in treatment response between the treatment groups.[1] We introduce some robustness to outliers by considering the trimmed mean, that is, the average treatment effect in which the largest 10% and the smallest 10% of response values are discarded from the trial sample.

In testing, we found that this trimmed ATE yielded more stable and trustworthy determinations of positive effect size than the standard ATE. We chose a value of 10% for the trimming parameter to balance representativeness of the data against robustness to outliers. In practice, the decision maker could choose to either increase this percentage to produce more robustness in the significance test or decrease the percentage in cases when extreme values of treatment response are believed to be meaningful. If the optimal subgroup has a statistically significant trimmed ATE, we consider the subset to be viable, and we recommend additional testing of the treatment in subjects who match the subgroup's covariate profile. Otherwise, we discard the optimal solution and determine that there is no need for further study of the treatment in this population. For the computational experiments in Section 3, we use a nonparametric hypothesis testing approach based on the bootstrap (Efron and Tibshirani 1994), with a significance level of $\alpha = 0.01$, chosen to yield a desired balance between true positive rates (TPRs) and false positive rates (FPRs).

In certain instances, one can introduce an additional level of verification by reserving a subset of the data as a test set not to be used when finding the optimal subset. The majority of the data can be used as a training set on which to apply the optimization approach and find an optimal subset. Then, one can identify which

subjects from the reserved test set are in the box defined by the optimal solution, and evaluate the out-of-sample ATE and its significance within the test subset. However, this approach becomes impractical in many settings. Because our algorithm is a subsetting method, the sample size reduction typically associated with training-test split is exacerbated by the fact our exceptional responder identification is further subsetting the data. For instance, if we reserve 20% of the initial data set as a testing data set and the optimal subset contains 25% of the data in the testing set, then our test ATE will be evaluated on only 5% of the data, with perhaps 2.5% in each treatment group. Decision makers should estimate and account for this sample size reduction when deciding whether to make a training-test split.

2.4. Finding Multiple Subsets

Up until now, we solve the problem of finding a single interpretable subset of the data with maximum ATE. Let us assume one wants to find M subsets with the highest ATEs, such that there is limited overlap between the subsets. A greedy approach to this problem is to iteratively solve the MIO formulation from the previous section while adding a constraint on the overlap with previous optimal subsets. For instance, let $Z_1 := \{i \,|\, z_i^* = 1\}$ be the set of data points in the optimal subset found in the first iteration of the optimization. One can then seek a second subset Z_2 with maximal ATE but limited overlap between Z_1 and Z_2, such that

$$\frac{|Z_1 \cap Z_2|}{|Z_1 \cup Z_2|} \leq \rho,$$

for a similarity parameter ρ chosen by the decision maker. The similarity parameter can range from 0 (no overlap) to 1 (full overlap). In the numerical experiments in Section 3.1, we chose $\rho = 0.2$ to produce relatively distinct groups without requiring full mutual exclusion.

To find Z_2 that satisfies this impurity constraint, one can add the following constraint to formulation (5):

$$\sum_{i \in Z_1} z_i \leq \rho \cdot \left(|Z_1| + \sum_{i \notin Z_1} z_i \right), \tag{6}$$

because $|Z_1 \cup Z_2| = |Z_1| + |Z_2 \backslash Z_1|$. Then re-solving the MIO with the added constraint (6) yields the optimal Z_2. More generally, for any $m = 2, \ldots, M$, one can add pairwise impurity constraints:

$$\sum_{i \in Z_\ell} z_i \leq \rho \cdot \left(|Z_\ell| + \sum_{i \notin Z_\ell} z_i \right), \quad \forall \ell = 1, \ldots, m-1.$$

When testing for significance with multiple subsets ($M > 1$), one should be careful to account for multiple comparisons using the Holm–Bonferroni procedure or a similar approach (Holm 1979).

2.5. Recursive Partitioning Heuristic

We present a heuristic inspired by the greedy, recursive partitioning scheme of classification trees (Breiman et al. 1984). The creation of this heuristic is motivated by the desire to quickly obtain near-optimal solutions, either for direct use or as warm-start feasible solutions for formulation (5).

Like a CART, at each branch, the heuristic searches for a one-dimensional covariate split that will greedily maximize an objective function, in this case, the ATE for subjects in the subset. In our heuristic, the split is determined by choosing both a lower and upper bound describing the subset along a single covariate dimension. There are several parameters that govern the choice of split at each branch to ensure heuristic solutions are feasible for formulation (5). First, the number of covariates used to make splits in a given tree must not exceed S_0. Second, for any given tree, we specify a depth parameter d, which serves as an upper bound on the number of branches that can be made in a given tree. Third, at each step of the recursion $c = 1, \ldots, d$, where d is the chosen depth parameter, we gradually decrease lower and upper bounds on the cardinality of the leaf indicating the solution subset. The lower and upper bounds are specified as $\underline{N}^c = \frac{n}{2} \cdot \left(\frac{2\underline{N}}{n}\right)^{\frac{1}{d-c+1}}$ and $\overline{N}^c = \frac{n}{2} \cdot \left(\frac{2\overline{N}}{n}\right)^{\frac{1}{d-c+1}}$, respectively, where n is the full sample size, and $\underline{N}$ and $\overline{N}$ are the subset cardinality bounds; on the last step of the recursion, when $c = d$, we have $\underline{N}^c = \underline{N}$ and $\overline{N}^c = \overline{N}$, so that the heuristic solution is feasible for formulation (5). We define $\delta_H(d)$ to be the ATE in the best subset found after applying the heuristic with depth parameter d.

Rather than generate a single tree yielding a single heuristic solution, we can specify a set of integer depth parameters $\mathcal{D}$ and use the heuristic to find the solution $\delta_H^* := \max_{d \in \mathcal{D}} \delta_H(d)$. For the computational experiments in Section 3, we use $\mathcal{D} = \{1, \ldots, 8\}$. To consider an even larger range of solutions, we can specify S^2 different starting pairs (s_1, s_2), $s_1 = 1, \ldots, S$, $s_2 = 1, \ldots, S$, such that the first branch of the tree must split on dimension s_1

and the second branch must split on dimension s_2. If $\delta_H(d, s_1, s_2)$ is the objective value of the best tree found with these starting points, then we have $\delta_H(d) := \max_{(s_1,s_2)\in\{1,\ldots,S\}^2} \delta_H(d, s_1, s_2)$.

At the end, we take δ_H^* and the corresponding subset lower and upper bounds as our best heuristic solution. In all tested instances, the heuristic found a near-optimal or optimal solution within seconds.

3. Computational Experiments

We conducted simulation experiments with data generated according to several different models relating the response vector $\mathbf{v}$ to treatment vector $\mathbf{y}$ and covariate matrix $\mathbf{X}$. In the base experiment, we had $n = 100$ subjects with four measured covariates, $x_i^{(1)}, \ldots, x_i^{(4)}$, $i = 1, \ldots, n$, drawn independent and identically distributed (i.i.d.) from a continuous uniform distribution over [0, 1]. Subjects were randomly assigned to one of two treatment conditions, $y_i = 1$, indicating assignment to the treatment group, or $y_i = 0$, indicating assignment to the control group, with an equal number assigned to each group. For each experiment, we assume a linear data model

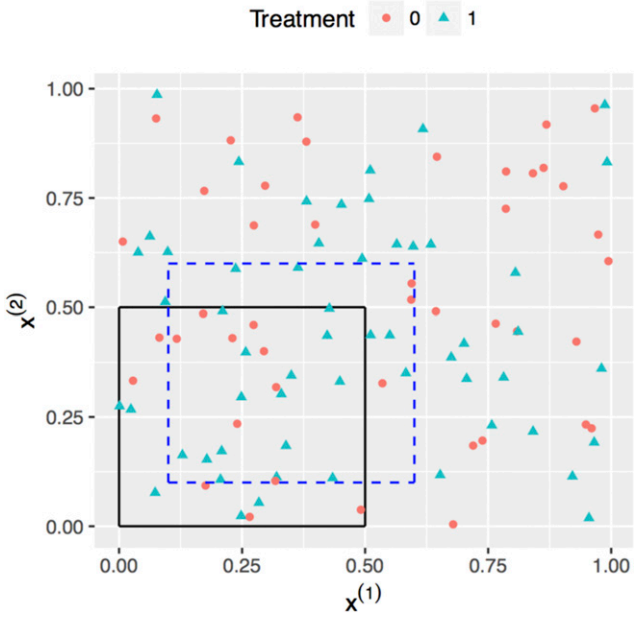
$$v_i = 2 + \delta_0 \cdot y_i \cdot \mathbb{I}\{x_i^{(1)} \le 0.5\} \cdot \mathbb{I}\{x_i^{(2)} \le 0.5\} + \varepsilon_i, \tag{7}$$

where δ_0 is the ground truth treatment effect and ε_i is a noise term. In the base case, we assume $\delta_0 = 2$ and ε_i drawn i.i.d. standard normal. To evaluate the false positive rate of our method, we also considered a modification in which $\delta_0 = 0$.

For 250 unique, random samples, we solved formulation (5) with $S_0 = 2$, $\underline{N} = \lfloor 0.1 \cdot n \rfloor$, $\overline{N} = \lceil 0.3 \cdot n \rceil$, and $\boldsymbol{\gamma}$ specified from 0 to 1 by increments of 0.1 for all dimensions. The computations were implemented using Julia programming language (Bezanson et al. 2012, Lubin and Dunning 2015) and the integer optimization solver Gurobi 6.5 (Gurobi Optimization, Inc. 2016). For each sample, we applied a bootstrapped hypothesis test used by Bertsimas et al. (2015) with significance level $\alpha = 0.01$ to determine whether the subgroup had a statistically significant treatment response.

In the modified case with $\delta_0 = 0$, the algorithm erroneously identified a statistically significant subset 12.4% of the time, which we consider to be the false positive rate. In the base case with $\delta_0 = 2$, the method identified a statistically significant subset in 76.8% of simulations. However, because of noise in the response model (7), the subset identified did not always precisely match the known underlying best subset $\mathcal{I}_0^* := \{i \,|\, x_i^{(1)} \le 0.5, x_i^{(2)} \le 0.5\}$ (Figure 1). To evaluate the true positive rate of our method in the base case, we compared all significant found subsets to the known best subset. The confusion matrix showing the number of subjects within each subset averaged across all simulations is shown in Table 1, for subsets that were found to be statistically

Figure 1. An Example of a Data Sample Projected onto the $(x^{(1)}, x^{(2)})$ Space



Notes. Shape indicates treatment group. The black solid line delineates the boundary of known best subset $\mathcal{I}_0^*$. The blue dashed line delineates the boundary of the found optimal subset.

Table 1. Average Percentage of Subjects in the Found Subset vs. Best Known Subset $\mathcal{I}_0^*$ over 250 Simulations in Base Case When the Found Subset Was Significant

	$i \in \mathcal{I}_0^*$ (%)	$i \notin \mathcal{I}_0^*$ (%)
In found subset ($z_i = 1$)	18.5	4.9
Not in found subset ($z_i = 0$)	6.6	70.0

significant. According to Table 1, the average accuracy, or percentage of subjects for which the classification of the found subset matched the best known classification, was 88.5%; the accuracy should be compared with the baseline prevalence of 74.9% of subjects not in $\mathcal{I}_0^*$.

One way to evaluate the effectiveness of our method is to examine subsets for which the ATE was found to be statistically significant *and* at least 50% of subjects in the found subset ($z_i = 1$) were also in the known best subset ($i \in \mathcal{I}_0^*$), that is, the PPV was greater than 50%. We use the term threshold-based TPR to refer to the percentage of found subsets with a significant positive ATE and a PPV greater than 50%. The threshold-based TPR in the base case was 68.0%. Alternatively, if we simply average the PPV among subsets determined to have statistically significant positive ATEs, we derive the PPV-weighted TPR, which was 60.7% in the base case. When using the heuristic alone, without mixed-integer optimization, the threshold-based TPR was 66.4%, and the PPV-weighted TPR was 59.0%.

We conducted sensitivity analyses with respect to sample size n, covariate dimension S, depth restriction S_0, effect size δ_0, and distribution of the noise term ε_i (Table 2). As expected, the TPR (both threshold-based and PPV-weighted) increased with n and δ_0, and decreased with the variance of ε_i. The TPR was relatively unchanged with respect to S and was very low for $S_0 = 1$, but grew slowly for $S_0 \geq 2$. The FPR was stable and low in all tested instances, although it grew with respect to S, likely because of the increase in the combinatorial space of possible subsets.

We also conducted computational timing experiments to test how tractable the optimization is as n grows. In all tested instances, the heuristic found an optimal or near-optimal warm-start solution within seconds. The time to provable optimality is shown in Table 3 for experiments with $S = 10$ and $\sum_{s=1}^{S} K_s = 110$. We speculate that the use of the warm-start heuristic and preprocessing done by Gurobi cause the computational time to increase at a decreasing rate in n. If the computational time becomes a practical concern (e.g., when the size of the instance is very large), then formulation (2) can be augmented with additional valid inequalities introduced by Tawarmalani et al. (2002) for fractional 0–1 programs. These inequalities may help to obtain tighter relaxations by means of a large number of additional continuous variables and, therefore, improve convergence behavior.

Table 2. Sensitivity Analyses of TPR and FPR of Found Subsets with Varying Sample Size n, Effect Size δ_0, and Distributions of Noise ε_i

Parameter	Value	Percentage of significant subsets (%)	TPR based on threshold (%)	TPR weighted by PPV (%)	FPR (%)
n	100[a]	76.8	68.0	60.7	12.4
	150	96.8	90.8	84.9	16.0
	200	98.8	97.2	91.7	14.4
S	2	72.4	70.0	62.4	5.2
	3	79.2	74.8	65.2	11.2
	4[a]	76.8	68.0	60.7	12.4
	5	81.2	71.6	63.6	16.0
	10	86.8	66.0	60.2	25.2
S_0	1	40.0	23.2	20.8	3.6
	2[a]	76.8	68.0	60.7	12.4
	3	84.0	73.2	63.3	17.6
	4	88.0	76.8	65.6	19.2
δ_0	0				12.4
	1	28.8	17.2	15.9	
	2[a]	76.8	68.0	60.7	
	4	98.8	96.0	87.9	
ε_i distribution	$N(0,1)$[a]	76.8	68.0	60.7	12.4
	$U[-\sqrt{3}, \sqrt{3}]$	68.0	58.8	52.1	12.8

[a]The base case parameters used were $n = 100$, $S = 4$, $S_0 = 2$, $\delta_0 = 2$, and $\varepsilon_i \sim N(0,1)$ (i.i.d.).

Table 3. Average Computational Time to Achieve Provable Optimality by Sample Size n

n	Time (s)
100	33
250	1,070
500	7,280
1,000	14,535

3.1. Multiple Subsets

We conducted an additional experiment in which we generated data according to the response model:

$$v_i = 2 + 6 \cdot y_i \cdot \mathbb{I}\{x_i^{(1)} \le 0.5\} \cdot \mathbb{I}\{x_i^{(2)} \le 0.5\} + 3 \cdot y_i \cdot \mathbb{I}\{x_i^{(1)} > 0.5\} \cdot \mathbb{I}\{x_i^{(2)} > 0.5\} + \varepsilon_i,$$

with ε_i distributed standard normal. We allowed the algorithm to find two subsets ($M = 2$) and measured the significance of each found subset using the Holm procedure with significance level $\alpha = 0.01$. The results of this experiment are shown in Table 4. The method virtually always found the exact subset with $\delta_0 = 6$. With respect to the second subset with $\delta_0 = 3$, the method reliably detected the subset in more than 50% of instances.

4. Real Case Studies

In this section, we present two examples in which we apply the optimization approach to real data sets. In Examples 1 and 2, the method identifies a subset with a statistically significant positive average treatment effect despite an overall nonsignificant negative treatment effect in the study sample; the results are slightly less conclusive in Example 2 than in Example 1. In Example 3, the method does not identify a subgroup with a statistically significant positive average treatment effect. We include both examples to demonstrate our method's ability to discover new insights in clinical trial data, while maintaining appropriate discriminatory power when there is no signal to be found. Note that we cannot evaluate a true positive or false positive rate in these experiments because the underlying presence or absence of a subset of exceptional responders is unknown.

By way of comparison, for each example, we also present the subsets identified by two other heuristic methods. One comparison method is a recursive partitioning method derived from Su et al. (2009); the second comparison method is a bump-hunting method derived from Kehl and Ulm (2006). A detailed description of each of these heuristic algorithms is presented in the online appendix. In all of our computational experiments, both of these methods returned solutions very fast; computational time was under one minute in all instances. In all examples, our method performed as well or better than the best heuristic method in terms of the ATE of the best box found. Even when a heuristic method may produce a solution as good as that of our method (as in Example 1 from Section 4.1), the MIO approach has two advantages. First, our method guarantees optimality: by using the MIO approach, we ensure that we indeed find the best possible box (if it exists). Second, we provide a method to extend the MIO approach to more than one box with a high ATE.

4.1. Example 1: Randomized Placebo-Controlled Trial of Diethylstilbestrol for Late-Stage Prostate Cancer

Diethylstilbestrol is a form of synthetic estrogen that has been used to treat late-stage prostate cancer. Byar and Green (1979) discuss data from a randomized trial testing the effect on survival of diethylstilbestrol at three dosage levels (0.2 mg, 1.0 mg, or 5.0 mg) versus placebo in 502 patients with stage 3 or 4 prostate cancer.[2] For each patient, the researchers recorded the months of follow-up and the patient's mortality status, along with covariates including age, weight, medical history, cancer status, and common laboratory measurements.

Taking the group that received 5.0 mg of estrogen and the placebo group, we analyzed 252 subjects using our optimal subset approach with 12 covariates, $S_0 = 3$, $\underline{N} = 25$, and $\overline{N} = 76$. In the study, the 125 subjects who received the 5.0 mg dose of estrogen had an average survival time of 35.0 months from the time of enrollment until death or end of study follow-up, whereas the 127 subjects in the placebo group had average survival of 35.3 months. The average survival in the treatment group was 0.3 months shorter in the treatment group than

Table 4. TPR with Respect to Finding Each of Two Known Underlying Subsets

Subset	δ_0	Percentage of significant subsets (%)	TPR based on threshold (%)	TPR weighted by PPV (%)
$\mathcal{I}_1^* := \{i \mid x_i^{(1)} \le 0.5,\ x_i^{(2)} \le 0.5\}$	6	100.0	97.2	89.1
$\mathcal{I}_2^* := \{i \mid x_i^{(1)} > 0.5,\ x_i^{(2)} > 0.5\}$	3	57.2	54.0	50.7

in the placebo group, but the effect was nonsignificant using the bootstrap hypothesis testing approach discussed in Section 3 with significance level $\alpha = 0.01$. This hypothesis testing approach and significance level are used to test for effect significance throughout the current section. Applying our approach, we found an optimal box containing 59 subjects (30 in the estrogen group, 29 in the placebo group). Within this subset, average survival was 42.5 months in the treatment group versus 24.3 months in the placebo group, an average treatment effect of 18.2 additional months of survival. The effect was statistically significant with $p = 0.001$. The subset was defined by patients who have stage 4 cancer, no history of cardiovascular disease, and diastolic blood pressure of 70 mmHg or above at time of measurement. The results suggest that further testing of the 5.0 mg diethylstilbestrol treatment is warranted in subjects meeting these specific criteria.

We analyzed the same data set using two heuristic methods described in the online appendix, with the same parameter settings we used for our method. Method 1 (Su et al. 2009) found a box with a much lower ATE than that found using our method (9 additional months of survival, compared with 18.2 added months of survival with our method). Method 2 (Kehl and Ulm 2006) found the exact same box as our method. Both solutions were statistically significant. In this example, our method performed equally well as the best heuristic method.

4.2. Example 2: Randomized Controlled Trial of Periodontal Therapy to Prevent Preterm Birth

In the obstetrics and periodontal therapy study, researchers studied whether mechanical periodontal therapy administered to pregnant women between 13 and 17 weeks of gestation affected birth outcomes, including gestational age and the presence of congenital anomalies (Hodges and Michalowicz 2013).[3] The 413 subjects in the treatment group underwent scaling and root planing before 21 weeks, underwent monthly tooth polishing, and received oral hygiene instruction. Of those, 410 subjects underwent scaling and root planing after delivery. The study found that although periodontal treatment improved periodontal outcomes and did not cause harm, the treatment did not significantly alter birth outcomes.

We applied our optimal subset approach to a randomly selected training set of 500 subjects with 11 covariates, including age, body mass index, presence of hypertension and diabetes, use of alcohol and tobacco, number of previous pregnancies, and measures of periodontal disease. The parameters used were $S_0 = 3$, $\underline{N} = 50$, and $\overline{N} = 150$. We considered four different birth outcomes (early birth before 37 weeks, presence of congenital anomalies, and Apgar scores at one and five minutes) and found a significant subset only for the presence of congenital anomalies. Our randomly selected training set included 247 subjects who received the treatment, among whom there were nine congenital anomalies (3.6%), and 253 subjects from the control group, among whom there were six congenital anomalies (2.4%); the average treatment effect was +1.2%, a net increase in anomalies for the treatment group, which was a statistically significant difference at $p < 0.001$. Our method identified a subset including subjects of age 22 to 30 with BMIs of 26 or higher and fewer than 20 teeth qualifying as having periodontal disease. The subset included 58 subjects who received the treatment, among whom there was one anomaly (1.7%) and 51 subjects from the control group, among whom there were five anomalies (9.8%). In this subset, the prevalence of congenital anomalies was lower in the treatment group by 8.1%, an average treatment effect that was statistically significant at $p < 0.001$.

We also evaluated outcomes among the remaining 323 subjects not included in the training set. This validation set included 166 subjects from the treatment group with four observed anomalies (2.4%), and 157 subjects from the control group with one observed anomaly (0.6%). The overall average treatment effect in the validation set was +1.7%, with a higher prevalence of anomalies in the treatment group. Within the validation set, there were 83 subjects who met the criteria to be included in the box identified by our method. In the validation set, the box included 43 treatment group subjects with zero observed anomalies and 40 control group subjects with zero observed anomalies. Thus, out of sample, the average treatment effect in the box was zero, which was a statistically significant difference ($p < 0.001$) compared with the 1.7% higher treatment-group prevalence observed for the overall validation set.

Although the reduction in the prevalence of anomalies for training-set subjects in the box does not fully hold in the validation set, we conclude there is moderate potential that mechanical periodontal therapy can reduce congenital abnormalities among women aged 22 to 30 with BMIs of 26 or higher and fewer than 20 teeth with periodontal disease.

When using other heuristic methods with the same parameter settings, the solutions found were worse than those found using our method. Method 1 (Su et al. 2009) found a box with a much worse ATE than that found using our method, as the prevalence of congenital anomalies actually increased by 1.7% in this method's solution; the poor performance is likely explained by the fact that Method 1 was developed for a slightly different purpose than our method, as well as parameter settings that may not have been optimal for this method. Method 2 (Kehl and Ulm 2006) found a box with a worse ATE than that found by our method (a 7.7% decrease

in congenital anomalies versus an 8.1% reduction for our method); unlike our method's solution, the box found by Method 2 did not have a statistically significant ATE. In this example, our method outperformed the best heuristic method.

4.3. Example 3: Randomized Placebo-Controlled Trial of D-Penicillamine for Primary Biliary Cirrhosis of the Liver

Primary biliary cirrhosis (PBC) of the liver is a rare chronic disease that leads to death. Between 1974 and 1984, the Mayo Clinic conducted a placebo-controlled trial of the drug D-penicillamine in 312 PBC patients, as described in Fleming and Harrington (1991).[4] For each patient, the researchers recorded the number of days between study registration and the earlier of death, liver transplantation, or the end of the study in July 1986. There were 16 covariates measured at the time of registration, including age, sex, disease stage, presence of associated conditions, and various hematological laboratory measurements, such as serum bilirubin, serum cholesterol, albumin, and platelet counts. Analysis of the study found that there was no significant difference in survival time between the treatment and placebo groups. Among 154 subjects in the placebo group, the average survival time from study enrollment was 1,996.9 days. Among 158 subjects in the treatment group, the average survival was 18.7 days longer, at 2,015.6 days. This result was not significant under the null hypothesis using the bootstrap hypothesis test with significance level $\alpha = 0.01$.

We applied our optimal subset approach to determine whether there was an interpretable subset of the population for which the drug may have had a significant effect on survival. Because of some missing data, we used 14 of the 16 covariates. We sought an interpretable subset with $S_0 = 3$, $\underline{N} = 20$, and $\overline{N} = 60$. The optimal interpretable subset included 65 subjects who met all three of the following conditions: had not exhibited spider angiomas, had serum bilirubin between 0.75 mg/dl and 1.5 mg/dl, and had a prothrombin time of no more than 11.1 seconds. In the optimal box, the average survival time among 33 subjects in the treatment group was 2,910.6 days, which was 854.7 days longer than the average survival of 2,055.9 days among 32 subjects in the placebo group. We considered the null hypothesis that the treatment effect in the subset did not differ from that in the overall sample, which we observed to be 18.8 days of added survival. The p-value was 0.03, which was not significant at $\alpha = 0.01$. Therefore, we determined that the subset may have been a false positive and did not warrant further investigation as a possible group of exceptional responders.

We also analyzed this data set using the two other heuristic methods. Both methods, Methods 1 and 2, found boxes with worse ATEs than that found with our method (404 added days of survival and 724 added days of survival, respectively, compared with 855 added days of survival for our method). None of the solutions were statistically significant, including the one produced by our method.

5. Discussion

In this paper, we show that the problem of identifying one or more interpretable subsets of a trial population with best (or worst) average treatment response can be modeled and efficiently solved using mixed-integer linear optimization. We use variable substitution and binary expansion to transform the fractional objective function into a linear function that is tractable in practical instances. We present an approach for determining whether the found subset is statistically significant. We also introduce a tree-based heuristic that finds near-optimal solutions quickly. In simulated and real-world scenarios, we demonstrate that the method finds subgroups worthy of further investigation, while minimizing the rate of false positive subsets. Further research is warranted to explore the use of the optimization approach on nonrandomized data from observational studies, or in trials where more than two treatment conditions are administered.

Endnotes

[1] In settings where the overall study sample had a statistically significant nonzero treatment response, one may want to adopt a different null hypothesis that the treatment effect in the subgroup does not differ from the overall treatment effect in the study population.

[2] The Byar and Green (1979) data are available at http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

[3] The Hodges and Michalowicz (2013) data are available at http://conservancy.umn.edu/handle/11299/160551.

[4] The Fleming and Harrington (1991) data are available online at https://www4.stat.ncsu.edu/~boos/var.select/pbc.html.

References

Berndt ER, Cockburn IM (2013) Price indexes for clinical trial research: a feasibility study. NBER Working Paper No. 18918, National Bureau of Economic Research, Cambridge, MA.

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.

Bezanson J, Karpinski S, Shah VB, Edelman A (2012) Julia: A fast dynamic language for technical computing. Working paper, Massachusetts Institute of Technology, Cambridge.
Borrero JS, Gillen C, Prokopyev OA (2017) Fractional 0–1 programming: Applications and algorithms. *J. Global Optim.* 69(1):255–282.
Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.
Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees* (CRC Press, Boca Raton, FL).
Byar D, Green S (1979) The choice of treatment for cancer patients based on covariate information. *Bull. Cancer* 67(4):477–490.
Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).
Fleming TR, Harrington DP (1991) *Counting Processes and Survival Analysis*, vol. 169 (John Wiley & Sons, Hoboken, NJ).
Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Statist. Medicine* 30(24):2867–2880.
Friedman JH, Fisher NI (1999) Bump hunting in high-dimensional data. *Statist. Comput.* 9(2):123–143.
Gurobi Optimization, Inc. (2016) *Gurobi Optimizer Reference Manual* (Gurobi Optimization, Inc., Beaverton, OR).
Hardin DS, Rohwer RD, Curtis BH, Zagar A, Chen L, Boye KS, Jiang HH, Lipkovich IA (2013) Understanding heterogeneity in response to antidiabetes treatment: A post hoc analysis using sides, a subgroup identification algorithm. *J. Diabetes Sci. Tech.* 7(2):420–430.
Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. *Nature Biotech.* 32(1):40–51.
Hodges JS, Michalowicz BS (2013) Public-use data from the obstetrics and periodontal therapy (opt) study, a randomized trial of periodontal therapy to prevent pre-term birth. Working paper, University of Minnesota, Minneapolis.
Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian J. Statist.* 6(2):65–70.
Kehl V, Ulm K (2006) Responder identification in clinical trials with censored data. *Comput. Statist. Data Anal.* 50(5):1338–1355.
Lubin M, Dunning I (2015) Computing in operations research using Julia. *INFORMS J. Comput.* 27(2):238–248.
Nikolaev AG, Jacobson SH, Cho WKT, Sauppe JJ, Sewell EC (2013) Balance optimization subset selection (BOSS): An alternative approach for causal inference with observational data. *Oper. Res.* 61(2):398–412.
Sauppe JJ, Jacobson SH (2017) The role of covariate balance in observational studies. *Naval Res. Logist.* 64(4):323–344.
Sauppe JJ, Jacobson SH, Sewell EC (2014) Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS J. Comput.* 26(3):547–566.
Schemper M (1988) Non-parametric analysis of treatment–covariate interaction in the presence of censoring. *Statist. Medicine* 7(12):1257–1266.
Su X, Tsai C-L, Wang H, Nickerson DM, Li B (2009) Subgroup analysis via recursive partitioning. *J. Machine Learn. Res.* 10(February):141–158.
Tawarmalani M, Ahmed S, Sahinidis NV (2002) Global optimization of 0-1 hyperbolic programs. *J. Global Optim.* 24(4):385–416.
U.S. Food and Drug Administration (2015) The drug development process—Step 3: Clinical research. Accessed January 21, 2019, http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm.
Wu T-H (1997) A note on a global approach for general 0–1 fractional programming. *Eur. J. Oper. Res.* 101(1):220–223.
Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* 107(500):1360–1371.