From Data to Decisions in Healthcare: An Optimization Perspective

by

Alexander Michael Weinstein

B.A., Yale University (2009)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author .....
Sloan School of Management
May 15, 2017

Certified by .....
Dimitris Bertsimas
Boeing Leaders for Global Operations Professor
Co-Director, Operations Research Center
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

From Data to Decisions in Healthcare:

An Optimization Perspective

by

Alexander Michael Weinstein

Submitted to the Sloan School of Management
on May 15, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

The past few decades have seen many methodological and technological advances in optimization, statistics, and machine learning. What is still not well understood is how to combine these tools to take data as inputs and give decisions as outputs.

The healthcare arena offers fertile ground for improvement in data-driven decision-making. Every day, medical researchers develop and test novel therapies via randomized clinical trials, which, when designed efficiently, can provide evidence for efficacy and harm. Over the last two decades, electronic medical record systems have become increasingly prevalent in hospitals and other care settings. The growth of these and other data sources, combined with the aforementioned advances in the field of operations research, enable new modes of study and analysis in healthcare. In this thesis, we take a data-driven approach to decision-making in healthcare through the lenses of optimization, statistics, and machine learning.

In Parts I and II of the thesis, we apply mixed-integer optimization to enhance the design and analysis of clinical trials, a critical step in the approval process for innovative medical treatments. In Part I, we present a robust mixed-integer optimization algorithm for allocating subjects to treatment groups in sequential experiments. By improving covariate balance across treatment groups, the proposed method yields statistical power at least as high as, and sometimes significantly higher than, state-of-the-art covariate- adaptive randomization approaches. In Part II, we present a mixed-integer optimization approach for identifying exceptional responders in randomized trial data. In practice, this approach can be used to extract added value from costly clinical trials that may have failed to identify a positive treatment effect for the general study population, but could be beneficial to a subgroup of the population.

In Part III, we present a personalized approach to diabetes management using electronic medical records. The approach is based on a k-nearest neighbors algorithm. By harnessing the power of optimization and machine learning, we can improve patient outcomes and move from the one-size-fits-all approach that dominates the medical landscape today, to a personalized, patient-centered approach.

Thesis Supervisor: Dimitris Bertsimas
Title: Boeing Leaders for Global Operations Professor
Co-Director, Operations Research Center

Acknowledgments

First and foremost, I would like to thank Professor Dimitris Bertsimas, my thesis advisor, mentor, and friend. I have been inspired by his vision, his positivity, and the strength of his convictions, especially regarding the types of problems the field of operations research should seek to address. Dimitris has gone out of his way to nurture my confidence in my own abilities and motivate me to achieve more than I could have imagined during my PhD.

Many thanks are due to my co-authors and collaborators from the ORC student body. I spent many hours working side by side with Nikita Korolko; as a team, we were greater than the sum of our parts and helped motivate each other to produce our best work through our friendship and intellectual curiosity. Our fruitful collaboration resulted in two papers: 1) "Covariate-Adaptive Optimization in Online Clinical Trials," submitted to *Biometrics*; and, 2) "Identifying Exceptional Responders in Randomized Trials: An Optimization Approach." Thanks also to Daisy Zhuo and Nathan Kallus, my co-authors on "Personalized Diabetes Management Using Electronic Medical Records," published in *Diabetes Care*. Nathan's wisdom and thoughtfulness set a great example and strengthened the paper. Daisy's support was invaluable in checking my work, sharing alternative perspectives, and always contributing a sunny, positive attitude; it was fun presenting our work together at the American Diabetes Association conference in New Orleans. Thanks to Lauren Berk, Julia Yan, and Nataly Youssef, my collaborators on the Sapient project; our camaraderie made the project a much more enjoyable experience.

I would like to thank all the students from the ORC, who were an incredible source of support and friendship throughout my five years. In particular, I am grateful to the students with whom I collaborated on research, coursework, preparation for qualifying exams, and TAs: Jerry Kung (many times over), He Wang, Michael Beeler, Christina Epstein, Berk Ustun, Yaron Shaposnik, John Silberholz, and others; I truly would not have made it through the program without all of you. Many senior students were terrific role models and sources of wisdom and friendship for me, among them Kris

Johnson Ferreira, Allison O'Hair, Angie King, David Fagnan, Nathan, Vishal Gupta, Yehua Wei, Iain Dunning, Andre Calmon, Adam Elmachtoub, Joel Tay, Swati Gupta, Chaitanya Bandi, Nataly, John, and others. Nishanth Mundru was a great friend throughout; our regular lunches were always a welcome escape from it all. Thanks also to my fellow REFS, Kris, Joel, Jerry, Sebastien Martin, Arthur Delarue, and Julia, and to my fellow INFORMS officers, Velibor Misic, Anna Papush, and Alex Sahyoun.

Thank you to the ORC faculty, among whom were great teachers, role models, and resources. Special thanks to my thesis committee, Dimitris, Colin Fogarty, and Vivek Farias, and to Asu Ozdaglar who served on my general examination committee. Colin was particularly helpful in sharing his technical expertise for "Covariate-Adaptive Optimization in Online Clinical Trials." Thanks also to David Simchi-Levi who was a terrific supervisor during the first two years of my PhD while I worked on dynamic pricing problems.

In support of "Personalized Diabetes Management Using Electronic Medical Records," thanks are due to Dr. Michael Kane of the Massachusetts Institute of Technology for sharing clinical expertise in the progression and treatment of diabetes, and to Dr. William Adams of Boston University Clinical and Translational Science Institute for sharing clinical expertise and assisting with the interpretation of EMR. I also acknowledge the use of the Boston Medical Center i2b2 database in the diabetes work.

Finally, I cannot sufficiently express my gratitude to my family and friends from outside of MIT who supported me throughout five years full of twists and turns. Thanks to my parents, Amy and David, who have done so much to support and encourage my educational pursuits and to shape the person I have become. Thank you to my brothers, Natey and Garrett, my Grandma Aileen, and my Grandpa Seymour. Bobie would have loved to see me graduate from MIT and she is dearly missed. Thanks also to my new parents-in-law, Steve and Susan, and my brother-in-law, Will. And my last and greatest thanks are due to my wife, Jesse, who has been a constant partner, supporter, motivator, and friend throughout.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent decades have brought many methodological and technological advances in optimization, statistics, and machine learning. What is still not well understood is how to combine these tools to take data as inputs and give decisions as outputs.

The healthcare arena offers fertile ground for improvement in data-driven decision-making. Every day, medical researchers develop and test novel therapies via randomized clinical trials, which, when designed efficiently, can provide evidence for efficacy and harm. Over the last two decades, electronic medical record systems have become increasingly prevalent in hospitals and other healthcare settings. The growth of these and other data sources, combined with the aforementioned advances in the field of operations research enable new modes of study and analysis in healthcare.

In this thesis, we take a data-driven approach to decision-making in healthcare through the lenses of optimization, statistics, and machine learning. In Section 1.1, we introduce two practical applications in which we use analytics to make an impact on decision-making in healthcare. In Section 1.2, we briefly discuss some of the methodological approaches that underlie our work. Finally, in Section 1.3, we summarize the main contributions of the studies presented in Chapters 2, 3, and 4 of the thesis. Chapter 5 contains some concluding remarks.

1.1 Practical Challenges in Healthcare

The thesis considers two broad areas of decision-making in healthcare. In Chapters 2 and 3 of the thesis, we apply mixed-integer optimization to enhance the *design and analysis of clinical trials*, a critical step in the approval process for innovative medical treatments. In Chapter 4, we approach the problem of *personalized medicine* through the lens of machine learning and optimization.

1.1.1 Design and analysis of clinical trials

In Chapter 2 of the thesis, we consider a problem in the design of clinical trials, namely, how to allocate subjects sequentially to treatment groups. The classical approach of randomized assignment can yield an accidental bias identified by Efron [1971], in which there is an imbalance in the distributions of known or hidden covariates across treatment groups. State-of-the-art methods address this accidental bias via covariate-adaptive randomization [Rosenberger and Sverdlov, 2008, Antognini and Zagoraiou, 2011, Kapelner and Krieger, 2014]. We take a different perspective by applying robust-mixed integer optimization to develop the first covariate-adaptive optimization algorithm. By improving covariate balance across treatment groups, the proposed method yields statistical power at least as high as, and sometimes significantly higher than, state-of-the-art covariate- adaptive randomization approaches.

In Chapter 3, we shift our attention to a problem in the analysis of clinical trial data. Because of the high cost of clinical trials, there may be value in identifying subgroups of the study population for which an exceptionally large positive or negative response was observed. We present the first mixed-integer optimization approach for identifying exceptional responders in randomized trial data. In practice, this approach could extract value from costly clinical trials that may have failed to identify a positive treatment effect for the general study population, but could be beneficial to a subgroup of the population.

1.1.2 Personalized medicine

In Chapter 4, we present a personalized approach to diabetes management using electronic medical records (EMR). Existing clinical guidelines for the treatment of type 2 diabetes are largely one-size-fits-all [Rodbard et al., 2009]. We propose a novel approach that harnesses the power of machine learning to improve patient outcomes and shift the landscape of diabetes care to a personalized, patient-centered focus. Our approach is based on a k-nearest neighbors algorithm that uses historical data from existing EMR systems to make personalized treatment recommendations based on patient-specific demographic, medical, and treatment history factors. The algorithm's recommendations, and the evidence supporting these recommendations, can be summarized for providers in an intuitive, interactive software dashboard. Using data from Boston Medical Center, we demonstrate the algorithm is effective in improving post-treatment levels of glycated hemoglobin, a measure of blood glucose level. Our personalized approach to diabetes management can serve as a model for how to use machine learning and increasingly available data from historical records to improve care across the disease spectrum.

1.2 Analytical Background

Our work draws on several modes of analysis that have been enabled by recent developments in mathematical research and computational technology. First, we use mixed-integer optimization (MIO) to model problems in the design and analysis of clinical trials in Chapters 2 and 3. Methodological and software development in the implementation of MIO solvers, such as Gurobi [Gurobi Optimization, Inc., 2016], combined with massive speed-ups in computational processing power, have made it possible to find optimal or near-optimal solutions to MIO formulations in timeframes that are practical for real-world settings, such as ours. Second, to handle uncertainty in the data inputs when designing clinical trials, we incorporate techniques from the burgeoning field of robust optimization in Chapter 2. Third, the ability to draw statistical inferences from a randomized trial or observational EMR data runs

throughout this work. Specifically, in Chapters 2, 3, and 4, we draw on the potential outcomes framework of Rosenbaum and Rubin [1983] as a way to understand individuals' responses to treatment. Finally, from the machine learning literature, we draw on algorithms for supervised learning, including k-nearest neighbors, regularized linear regression, classification and regression trees (CART), and random forests. These methods are used in Chapter 4 to generate personalized treatment recommendations. Recursive partitioning schemes, such as CART, are also relevant to the discussion of how to identify exceptional responders in Chapter 3.

1.2.1 Mixed-integer optimization

In Chapters 2 and 3, we use mixed-integer optimization, specifically mixed-binary linear optimization, to model problems related to the design and analysis of clinical trials. In mixed-binary linear optimization, the aim is to find an optimal vector solution $(\mathbf{x}, \mathbf{y})$ to problems of the following form:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \mathbf{c}'\mathbf{x} + \mathbf{d}'\mathbf{y} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b} \\ & \mathbf{x}, \mathbf{y} \geq 0 \\ & \mathbf{x} \in \{0,1\}^p, \end{aligned}$$

where $\mathbf{A}$ and $\mathbf{B}$ are data matrices, $\mathbf{b}, \mathbf{c}$, and $\mathbf{d}$ are data vectors, $\mathbf{x}$ is a decision vector of dimension p that takes binary values, and $\mathbf{y}$ is a decision vector that can take continuous values.

The use of binary decision variables $\mathbf{x}$ makes it possible to indicate a choice between two alternatives, such as whether or not to assign an individual to a given treatment group. There are many modeling techniques in integer optimization that enable one to define complex relationships between variables [Bertsimas and Tsitsiklis, 1997]. For instance, in Chapter 3, we use mixed-binary optimization to indicate whether a subject is located in the best subset of exceptional responders as determined

by the choice of lower and upper bounds for the subset in each covariate dimension.

Cutting planes and branch-and-bound algorithms can be used to solve complex mixed-integer linear optimization problems efficiently using software, such as Gurobi [Gurobi Optimization, Inc., 2016].

1.2.2 Robust optimization

In Chapter 2, we tackle an online optimization problem in which we aim to assign clinical trial subjects to treatment groups in order to balance covariates across groups without knowing *a priori* the covariate data of future subjects. We show that this optimization under uncertainty can be modeled using MIO combined with robust optimization. In robust optimization, one has a vector of uncertain variables $\mathbf{w}$, which are known with high probability to take values in the uncertainty set $\mathcal{U}$. The goal of robust optimization is to solve the problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{w}\in\mathcal{U}} f(\mathbf{x}, \mathbf{w}),$$

where $\mathcal{X}$ is the feasible set for the outer minimization problem.

Depending on the specific properties of the formulation, various techniques exist to solve this problem tractably [Ben-Tal et al., 2002, Bertsimas et al., 2011, Bandi and Bertsimas, 2012]. In Chapter 2, we use a choice of uncertainty set that allows us to solve the sub-problem, $\max_{\mathbf{w}\in\mathcal{U}} f(\mathbf{x}, \mathbf{w})$, explicitly for any fixed $\mathbf{x}$. Therefore, we can model the sequential assignment problem as a deterministic optimization problem that accounts for future uncertainty in the data.

1.2.3 Causal inference

Throughout this thesis, we refer to the potential outcomes framework of Rosenbaum and Rubin [1983]. Using this framework of causal inference, each subject, indexed by $i = 1, \ldots, n$, receives a treatment assignment $T_i \in \mathcal{T}$, where $\mathcal{T}$ is a set of treatment alternatives. For instance, in a placebo-controlled randomized clinical trial, we may have $\mathcal{T} = \{0, 1\}$, where $T_i = 1$ would indicate subject i was assigned to the treat-

ment group, and $T_i = 0$ would indicate subject i was assigned to the control group. According to the potential outcomes framework, we observe each subject's response v_i to treatment T_i, but we cannot observe the counterfactual response(s), which the subject would have experienced if she had been assigned to a treatment other than T_i. In the example where $\mathcal{T} = \{0, 1\}$, each subject i has a pair of potential outcomes $(v_i^{(1)}, v_i^{(0)})$, where the superscript indicates treatment (1) or control (0). We have the following relationship between the observed response and the potential outcomes: $v_i = v_i^{(1)} T_i + v_t^{(0)}(1 - T_i)$.

We adopt the potential outcomes framework in two different contexts. In Chapters 2 and 3, we apply a hypothesis testing approach with Fisher's sharp null hypothesis [Fisher, 1935], which states that each individual subject would have had the same response to treatment regardless of which treatment she was assigned. In our example, under the sharp null hypothesis, we have $v_i = v_i^{(1)} = v_i^{(0)}$, $i = 1, \ldots, n$. Thus, under the sharp null hypothesis, we not only observe the true response to treatment, but also, by construction, we learn each subject's complete set of potential outcomes.

The second context in which we apply potential outcomes is in Chapter 4 when making personalized treatment recommendations. In this setting, we cannot observe a subject's counterfactual outcomes under alternative treatment options. To impute these unobserved potential outcomes, we test and evaluate various algorithms for supervised machine learning.

1.2.4 Supervised machine learning

Supervised machine learning algorithms are used to estimate unknown response values based on a separate labeled set of training data. Training data consist of pairs $(\mathbf{w}_i, v)$, where $\mathbf{w}_i$ is a vector of side information, referred to as covariates, and v_i is a scalar response value, referred to as a label. By using the labeled training data to learn relationships between the covariates and response, one can then apply these learned functions to make a prediction, $\hat{v}_j$, for the unknown response value given an unlabeled covariate vector $\mathbf{w}_j$. When the supervised learning problem involves predicting a continuous response value, we refer to the problem as regression; when the response

comes from a discrete set of values or classes, such as a binary variable, we refer to the problem as classification.

For instance, in a classical supervised learning approach called multiple linear regression, one uses labeled data to learn the *coefficient vector* $\boldsymbol{\beta}$ that minimizes the error terms ϵ_i among the training data in the following linear function:

$$v_i = \boldsymbol{\beta}' \mathbf{w}_i + \epsilon_i.$$

In Chapter 4, we test several common algorithmic approaches in supervised learning to impute unknown potential outcomes in response to various pharmacological regimens for treatment of diabetes. First, we test a simple algorithm called k-nearest neighbors, which estimates the response, v_j, by taking the average response among a set of k labeled subjects who are closest to the unlabeled subject j, as measured by some pre-defined distance metric on the covariate space [Cover and Hart, 1967]. Second, we consider a regularized version of multiple linear regression called LASSO regression [Tibshirani, 1996]. Finally, we consider random forests, which is an ensemble method that generates multiple regression trees [Breiman et al., 1984] and then takes the mean prediction across trees [Breiman, 2001]. The recursive partitioning scheme that underlies classification and regression trees [Breiman et al., 1984] also inspires our development of a fast heuristic to identify exceptional responders in Chapter 3.

1.3 Main Contributions

Our contributions in this thesis can be summarized as follows, listed by chapter.

Chapter 2. Covariate-Adaptive Optimization in Online Clinical Trials

- We present a robust mixed-integer optimization algorithm that achieves strong covariate balance in the sequential allocation of subjects to treatment groups in randomized clinical trials.

- Our method yields statistical power at least as high as, and sometimes significantly higher than, state-of-the-art covariate-adaptive randomization approaches. In one example with a nonlinear covariate-response relationship, our method achieved a desired level of statistical power at a sample size 25-50% smaller than other state-of-the art methods.

- The optimization approach compares favorably with state-of-the-art methods in regard to common experimental biases, including selection bias and accidental bias with respect to observed and hidden covariates.

- The CA-RO algorithm is computationally tractable for instances of practical size, despite taking the form of a nonlinear mixed-integer optimization problem that cannot be solved using off-the-shelf commercial solvers. Our choice of uncertainty set allows us to extract a closed-form solution for the robust constraints. Hence, we are able to solve the optimization by enumeration with computational complexity that is independent of the sample size.

Chapter 3. Identifying Exceptional Responders in Randomized Trials via Mixed-Integer Optimization

- We present an optimization approach for identifying the subset in randomized trial data with the largest or smallest average treatment effect.

- Despite a fractional objective function, we show that the problem can be transformed into a tractable mixed-integer linear optimization problem.

- The approach has many practical applications in the analysis of randomized trials. Investigators could use our method to revisit long-terminated clinical trials and search for opportunities to revive the testing of failed drugs in promising subgroups. 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.

- In simulation experiments where a subset with large positive treatment response was known to exist, our method correctly identified the subset with high accuracy and high positive predictive value despite additive noise. When no such subset existed, our method falsely identified a significant subset in only a small percentage of cases.

- Using data from a randomized clinical trial for the use of estrogen in treating late-stage prostate cancer, we find a subgroup in which the treatment effect in terms of survival days is substantially larger than for the general study population.

Chapter 4. Personalized Diabetes Management Using Electronic Medical Records

- Our study is the first of its kind to use machine learning to develop an algorithm for personalized treatment recommendations using electronic medical records.

- Based on simulations with data from Boston Medical Center, we estimate that the use of our algorithm could improve outcomes for patients with type 2 diabetes by reducing post-treatment glycated hemoglobin levels relative to current practice.

- The algorithm can be integrated into existing EMR systems to dynamically suggest personalized treatment paths for each patient based on historical records.

- We prototype an intuitive, interactive dashboard that summarizes the evidence for each recommendation, including the expected distribution of outcomes under alternative treatments.

- We believe this integrated, interactive approach has the potential to reshape medical practice across the disease spectrum.

Chapter 2

Covariate-Adaptive Optimization in Online Clinical Trials

This work has been submitted for review at Biometrics, with co-authors Dimitris Bertsimas and Nikita Korolko.

Pharmaceutical companies spend tens of billions of dollars each year to operate multi-year clinical trials needed for the approval of new drugs. In this chapter, we present a novel covariate-adaptive optimization algorithm for online allocation in clinical trials that leverages robust mixed-integer optimization. In all tested scenarios, the proposed method yields statistical power at least as high as, and sometimes significantly higher than, state-of-the-art covariate-adaptive randomization approaches. We present a setting in which our algorithm achieves a desired level of power at a sample size 25-50% smaller than that required with randomization-based approaches. Correspondingly, we expect that covariate-adaptive optimization could substantially reduce both the duration and operating costs of clinical trials in many commonly observed settings, while maintaining computational efficiency and protection against experimental bias.

2.1 Introduction

The annual expenditure of global pharmaceutical companies on research and development was $46.4 billion in 2010; of this, $32.5 billion was due to the high expense of clinical trials [Berndt and Cockburn, 2013]. Drug approval requires multiple phases of clinical trials that typically take years to complete [U.S. Food and Drug Administration, 2015]. We present an algorithm that can decrease the sample size needed to conduct a clinical trial by as much as 25-50% in certain settings, and thus substantially reduce both the cost of conducting clinical trials and the time it takes for novel effective therapies to reach patients.

The study of treatment allocation for controlled experiments dates back to Fisher [1935]. Randomization has been favored historically as a way to control for selection bias. However, randomization can yield another accidental bias identified by Efron [1971], in which there is an imbalance in the distributions of known or hidden covariates across randomly assigned treatment groups. There have been many attempts in the literature to address this accidental bias in both the offline and online allocation settings. For the offline problem, some prominent mechanisms are pairwise matching [Rosenbaum and Rubin, 1985, Greevy et al., 2004], rerandomization [Morgan and Rubin, 2012], and the finite selection model [Morris, 1979]. Bertsimas et al. [2015] used an alternative offline optimization-based approach.

For the online sequential allocation problem, Rosenberger and Sverdlov [2008] provide an excellent review of the available heuristics for covariate-adaptive randomization, including prestratification and biased coin designs. Many of the existing heuristics stem from variations of the biased coin design first introduced by Efron [1971], including nonrandomized minimization [Taves, 1974], randomized minimization [Pocock and Simon, 1975], and designs that attempt to minimize the variance of the treatment effect [Atkinson, 1982] or minimize loss of information [Antognini and Zagoraiou, 2011]. These biased coin designs outperform pure randomization and represent the current state of the art for online allocation. More recently, Kapelner and Krieger [2014] introduced a pooled sequential matching algorithm, which discards

covariate data as soon as each subject is matched. Bhat, Farias, and Moallemi [2015] propose a dynamic programming algorithm for sequential allocation that comes with computational challenges for which they provide an approximation algorithm.

In this paper, we develop a novel covariate-adaptive optimization mechanism for online allocation, which outperforms state-of-the-art covariate-adaptive randomization methods. We extend the offline mixed-integer optimization (MIO) approach presented in Bertsimas et al. [2015] to the online setting in which patients arrive sequentially and each patient's covariate data cannot be observed until the time of her arrival. The new algorithm takes the form of a sequence of mixed-integer nonlinear optimization problems. The uncertainty about future subjects is modeled by robust optimization techniques with a quadratic uncertainty set [Ben-Tal, Nemirovski, and Roos, 2002, Bertsimas, Brown, and Caramanis, 2011].

The new method, henceforth referred to as the covariate-adaptive robust optimization (CA-RO) algorithm, delivers the following benefits:

1. In all tested scenarios, the CA-RO method achieved statistical power at least as high as, and sometimes significantly higher than, covariate-adaptive randomization (CA-RAND) approaches. We present an example of a nonlinear covariate-response setting for which the CA-RO method achieved a desired level of statistical power at a sample size 25-50% smaller than that required with the best CA-RAND approach.

2. We present theoretical and empirical evidence that the optimization approach compares favorably with CA-RAND methods with respect to three advantages of complete randomization described by Efron [1971]: freedom from selection bias; freedom from accidental bias with respect to observed and hidden covariates; and, a reasoned basis for inference.

3. The algorithm is sufficiently general to produce assignments among multiple groups $p = 1, \ldots, m$ with multiple observed covariates per subject. The CA-RO algorithm can also be extended to the setting where it is possible to aggregate subjects into small clusters of size r prior to making group assignments.

4. The CA-RO algorithm is computationally tractable for instances of practical size, despite taking the form of a nonlinear mixed-integer optimization problem that cannot be solved using off-the-shelf commercial solvers. Our choice of uncertainty set allows us to extract a closed-form solution for the robust constraints. Hence, we are able to solve the optimization by enumeration with complexity $\mathcal{O}(m^r)$, which does not depend on the sample size N. In all observed instances, CA-RO provides the decision-maker with a high-quality assignment recommendation instantaneously via enumeration.

The rest of the paper is organized as follows. In Section 2.2, we briefly revisit the optimization-based allocation algorithm for the offline setting from Bertsimas, Johnson, and Kallus [2015] This offline algorithm will form the basis for the online CA-RO approach we develop in Section 2.3. At the end of Section 2.3, we present computational results from experiments demonstrating the effectiveness of CA-RO in reducing between-group covariate imbalance. In Section 2.4, we provide empirical evidence that the CA-RO algorithm achieves a high level of statistical power with much smaller sample size as compared to CA-RAND methods when the covariate-response relationship is nonlinear. In Section 2.5, we discuss the experimental bias and inference properties of the CA-RO approach and demonstrate that it compares favorably with CA-RAND methods. Section 2.6 contains concluding remarks.

2.2 Offline Optimization Approach

In this section, we describe a MIO approach to assign groups for the setting when pre-treatment covariate values of all subjects are known ahead of time (Bertsimas et al., 2015). The decision-maker knows *a priori* the total number of subjects N in the experiment and the respective covariates $\mathbf{w} = (w_1, \ldots, w_N)$ of all subjects. Thus, she can make treatment allocations using this full information. This may be the case, for example, in laboratory cancer drug testing on mice.

The decision-maker will assign $k := N/m$ subjects to each of $m \geq 2$ treatment groups. The objective of the assignment is to minimize the maximum discrepancy

between any two groups in the weighted sum of the first and second moments of the covariates. Without loss of generality, we assume that the vector of covariates $\mathbf{w}$ is normalized and has zero sample mean and unit sample variance. The parameter ρ regulates the relative weight of the first and the second moments. The binary decision variables are $\mathbf{x} = \{x_{ip} \,|\, i = 1, \ldots, N,\, p = 1, \ldots, m\}$, where $x_{ip} = 1$ if subject i is assigned to group p, and $x_{ip} = 0$, otherwise. We can express the mean and second moment of each of the groups $p \in \{1, \ldots, m\}$ as follows:

$$\mu_p(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{N} w_i \, x_{ip} \quad \text{and} \quad \sigma_p^2(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{N} w_i^2 \, x_{ip}.$$

Hence, the optimal offline assignment can be found using the following MIO problem, which we henceforth refer to as the OPT algorithm:

$$\begin{aligned}
\min_{\mathbf{x}} \max_{p<q} \quad & |\mu_p(\mathbf{x}) - \mu_q(\mathbf{x})| + \rho |\sigma_p^2(\mathbf{x}) - \sigma_q^2(\mathbf{x})| \quad = \\
\min_{\mathbf{x},d} \quad & d \\
\text{s.t.} \quad & \forall p < q = 1, \ldots, m : \\
& \qquad d \geq \mu_p(\mathbf{x}) - \mu_q(\mathbf{x}) + \rho \sigma_p^2(\mathbf{x}) - \rho \sigma_q^2(\mathbf{x}) \\
& \qquad d \geq \mu_p(\mathbf{x}) - \mu_q(\mathbf{x}) + \rho \sigma_q^2(\mathbf{x}) - \rho \sigma_p^2(\mathbf{x}) \\
& \qquad d \geq \mu_q \mathbf{x}) - \mu_p(\mathbf{x}) + \rho \sigma_p^2(\mathbf{x}) - \rho \sigma_q^2(\mathbf{x}) \\
& \qquad d \geq \mu_q(\mathbf{x}) - \mu_p(\mathbf{x}) + \rho \sigma_q^2(\mathbf{x}) - \rho \sigma_p^2(\mathbf{x}) \\
& x_{ip} \in \{0, 1\} \\
& \sum_{i=1}^{N} x_{ip} = k, \quad \forall p = 1, \ldots, m \\
& \sum_{p=1}^{m} x_{ip} = 1, \quad \forall i = 1, \ldots, N \\
& x_{ip} = 0 \quad \forall i < p.
\end{aligned}$$

The final constraint reduces the redundancy due to permutation symmetry in group numbering.

In all tested scenarios from Bertsimas et al. [2015], the OPT method generates groups with covariate discrepancy that is exponentially lower in the group size k than those created by randomization. The expected average covariate discrepancy decreases from $O(k^{-1/2})$ for randomization to $O(2^{-k})$ for the OPT algorithm. Furthermore, the OPT algorithm demonstrates exceptional precision in estimating small treatment effects and superior statistical power given a fixed treatment effect.

For the remainder of this paper, the OPT algorithm will serve as a prescient benchmark for the performance of methods in the setting of sequential online allocation.

2.3 Covariate-Adaptive Optimization Algorithms

In this section, we introduce the proposed CA-RO algorithm, show that the the algorithm is tractable, develop an extension in which aggregation of decisions is allowed, and describe the results of empirical experiments comparing the covariate balance of CA-RO versus CA-RAND methods.

2.3.1 CA-RO algorithm

To extend the model from Section 2.2 to the online setting, we consider the problem of N subjects arriving sequentially. The decision-maker knows *a priori* the number of subjects k that will be assigned to each of $m \geq 2$ treatment groups, such that $N = km$.

At each time-step $t = 1, \ldots, N$, where t indexes both the period and the subject, the decision-maker observes the covariate vector $\mathbf{w}_t \in \mathbb{R}^S$, where S is the number of covariates observed for each subject. We assume that this sequence of random covariate vectors is exchangeable, such that any ordering of the subjects' arrival is equally likely. The decision-maker then sets a decision $\{x_{tp}\}_{p=1}^{m} \in \{0,1\}^m$, where $x_{tp} = 1$, if the decision-maker assigns subject t to group $p \in \{1, \ldots, m\}$, and $x_{tp} = 0$, otherwise. In the CA-RO algorithm, the choice of $\{x_{tp}\}_{p=1}^{m}$ is made by solving one instance of robust MIO formulation (3.1) at each time-step. The data for the opti-

mization at time-step t include the covariate observations $\{\mathbf{w}_i\}_{i=1}^{t}$ and assignments $\hat{\mathbf{x}} := \{\hat{x}_{ip} \,\big|\, i = 1, \ldots, t-1,\, p = 1, \ldots, m\} \in \{0,1\}^{(t-1)\times m}$ made at all previous time-steps. We define expressions for the sample mean $\bar{\mathbf{w}}_t$ and the empirical covariance matrix Σ_t at time-step t as follows:

$$\bar{\mathbf{w}}_t := \frac{1}{t}\sum_{i=1}^{t} \mathbf{w}_i \quad \text{and} \quad \Sigma_t := \frac{1}{t}\sum_{i=1}^{t} (\mathbf{w}_i - \bar{\mathbf{w}}_t)(\mathbf{w}_i - \bar{\mathbf{w}}_t)^\top.$$

We also define uncertain parameters $\tilde{\mathbf{w}} := \{\tilde{\mathbf{w}}_i \in \mathbb{R}^S\}_{i=t+1}^{N}$, which represent the unknown covariates for future subjects.

The objective of the CA-RO algorithm is to produce m groups whose covariate distributions are as similar as possible. We measure the proximity between two groups p and q in terms of the mean μ_p^s and *approximated* variance σ_p^s of group $p = 1, \ldots, m$ with respect to covariate $s = 1, \ldots, S$. At time-step $1 \le t \le N$, these sample statistics are defined as follows:

$$\mu_p^s := \frac{1}{k}\Big\{ \sum_{i=1}^{t-1} w_i^s \hat{x}_{ip} + w_t^s x_{tp} + \sum_{i=t+1}^{N} \tilde{w}_i^s x_{ip} \Big\},$$

$$\sigma_p^s := \frac{1}{k}\Big\{ \sum_{i=1}^{t-1} (w_i^s - \bar{w}_t^s)^2 \hat{x}_{ip} + (w_t^s - \bar{w}_t^s)^2 x_{tp} + \sum_{i=t+1}^{N} (\tilde{w}_i^s - \bar{w}_t^s)^2 x_{ip} \Big\},$$

where $\mathbf{x} := \{x_{ip} \in \{0,1\} \,\big|\, i = t, \ldots, N,\, p = 1, \ldots, m\}$ are the binary assignment decision variables. We model the decision at each time-step $t = 1, \ldots, N$ by the following optimization problem:

$$\min_{\mathbf{x}} \max_{p<q} \sum_{s=1}^{S} |\mu_p^s - \mu_q^s| + \rho|\sigma_p^s - \sigma_q^s|. \tag{2.1}$$

Given that the values of future covariates $\tilde{\mathbf{w}}$ are unknown, we employ robust optimization (Ben-Tal et al., 2002) to model formulation (2.1) under uncertainty:

$$\min_{\mathbf{x},\mathbf{M},\mathbf{V},z} \quad z$$

$$
\begin{aligned}
\text{s.t.} \quad & z \geq \sum_{s=1}^{S} M_{pq}^{s} + \rho V_{pq}^{s}, \qquad \forall p < q \\
& M_{pq}^{s} \geq \mu_{p}^{s} - \mu_{q}^{s}, \qquad \forall p < q,\ s = 1, \ldots, S, \quad \forall \tilde{\mathbf{w}} \in U_w \\
& M_{pq}^{s} \geq \mu_{q}^{s} - \mu_{p}^{s}, \qquad \forall p < q,\ s = 1, \ldots, S, \quad \forall \tilde{\mathbf{w}} \in U_w \\
& V_{pq}^{s} \geq \sigma_{p}^{s} - \sigma_{q}^{s}, \qquad \forall p < q,\ s = 1, \ldots, S, \quad \forall \tilde{\mathbf{w}} \in U_w \\
& V_{pq}^{s} \geq \sigma_{q}^{s} - \sigma_{p}^{s}, \qquad \forall p < q,\ s = 1, \ldots, S, \quad \forall \tilde{\mathbf{w}} \in U_w \\
& \sum_{i=1}^{t-1} \hat{x}_{ip} + x_{tp} + \sum_{i=t+1}^{N} x_{ip} = k, \quad \forall p = 1, \ldots, m \\
& \sum_{p=1}^{m} x_{ip} = 1, \quad \forall i = t, \ldots, N
\end{aligned}
\tag{2.2}
$$

In this formulation, we use the uncertainty set U_w defined as follows:

$$
U_w = \Big\{ \tilde{\mathbf{w}} \in \mathbb{R}^{(N-t)\times S} \quad \Big| \quad \tilde{\mathbf{w}}_i = \bar{\mathbf{w}}_t + (\Sigma_t)^{\frac{1}{2}}\, \boldsymbol{\varepsilon}_i,\ i = t+1, \ldots, N, \boldsymbol{\varepsilon} \in U_{\varepsilon} \Big\},
$$

where perturbation vector $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_{t+1}, \ldots, \boldsymbol{\varepsilon}_N)$ belongs to the ellipsoidal uncertainty set U_{ε}:

$$
U_{\varepsilon} = \Big\{ \boldsymbol{\varepsilon} \in \mathbb{R}^{(N-t)\times S} \quad \Big| \quad ||\boldsymbol{\varepsilon}||_2 = \sqrt{\sum_{i=t+1}^{N} \sum_{s=1}^{S} (\varepsilon_i^s)^2} \leq \Gamma \sqrt{(N-t)S} \Big\}. \tag{2.3}
$$

The robustness parameter Γ controls the size of the ellipsoid and represents the level of conservatism of the uncertainty set. In order to protect against experimental biases, we suggest that Γ should be chosen independently at random for each time-step (see Section 2.5).

Formulation (3.1) takes the form of a mixed-binary quadratic robust optimization problem with conic uncertainty set, which cannot be solved using off-the-shelf commercial solvers. We overcome this computational challenge by finding an efficient way to solve the following auxiliary optimization problems with respect to the uncertain variables $\tilde{\mathbf{w}}$:

$$
\max_{\tilde{\mathbf{w}} \in U_w} \left(\mu_p^s - \mu_q^s \right) \quad \text{and} \quad \max_{\tilde{\mathbf{w}} \in U_w} \left(\sigma_p^s - \sigma_q^s \right). \tag{2.4}
$$

The objectives of optimization problems (2.4) are to maximize a linear or quadratic function, respectively, over an ellipsoid. The unique structure of these problems allows us to derive closed-form solutions by applying Karush-Kuhn-Tucker conditions along with eigenvalue optimization. Therefore, formulation (3.1) is equivalent to a mixed-binary optimization problem, described in Section 2.3.2, that can be solved via simple enumeration of m scenarios.

2.3.2 Tractability of the CA-RO algorithm

Lemma 1. Consider robust optimization problem (3.1) with ellipsoidal uncertainty set U_ε as defined in (2.3). To find the optimal objective value of this discrete optimization problem and the optimal current assignment at time t, it is sufficient to inspect the following easily specified set $\mathcal{X}$ consisting of not more than m points:

$$\begin{aligned}\mathcal{X} := \bigcup_{p=1}^{m} \{\, & x_{tp} = 1; \\ & x_{tq} = 0, \quad \forall q = 1, \ldots, m,\ q \neq p; \\ & x_{iu} = 0, \quad \forall i = t+1, \ldots, N,\ u = 1, \ldots, m; \\ & \sum_{i=1}^{t-1} \hat{x}_{ip} + x_{tp} \leq k \,\}.\end{aligned}$$

Proof. In order to model the constraints for each p, q, s from optimization problem (3.1) that should hold for all possible realizations of uncertain vector $\tilde{\mathbf{w}} \in U_w$, we will find a closed-form solution to the following auxiliary optimization problems, repeated from (2.4):

$$\max_{\tilde{\mathbf{w}} \in U_w} (\mu_p^s - \mu_q^s) \quad \text{and} \quad \max_{\tilde{\mathbf{w}} \in U_w} (\sigma_p^s - \sigma_q^s).$$

Step 1. Optimization of the linear term.

Let us define a parameter $\tilde{\Gamma} := \Gamma^2 (N - t) S$, where Γ is the robustness parameter from (2.3). Then, for any fixed values of p, q, s, and $\tilde{\Gamma}$, we consider the optimization

problem

$$\max_{\tilde{\mathbf{w}} \in U_w} (\mu_p^s - \mu_q^s). \tag{2.5}$$

We have

$$k(\mu_p^s - \mu_q^s) = \sum_{i=1}^{t-1} w_i^s (\hat{x}_{ip} - \hat{x}_{iq}) + w_t^s (x_{tp} - x_{tq}) + \sum_{i=t+1}^{N} \tilde{w}_i^s (x_{ip} - x_{iq}),$$

where only the last term of the right-hand side depends on uncertain $\tilde{\mathbf{w}}$. Therefore, we need to solve the following optimization problem for fixed values of components of $\mathbf{x}$:

$$\begin{aligned}\max_{\tilde{\mathbf{w}} \in U_w} \sum_{i=t+1}^{N} \tilde{w}_i^s (x_{ip} - x_{iq}) &= \max_{\boldsymbol{\varepsilon} \in U_\varepsilon} \sum_{i=t+1}^{N} (\bar{w}_t^s + \mathbf{v}_{(s)}^\top \boldsymbol{\varepsilon}_i)(x_{ip} - x_{iq}) \\ &= \bar{w}_t^s \sum_{i=t+1}^{N} (x_{ip} - x_{iq}) + \max_{\boldsymbol{\varepsilon} \in U_\varepsilon} \sum_{i=t+1}^{N} (\mathbf{v}_{(s)}^\top \boldsymbol{\varepsilon}_i)(x_{ip} - x_{iq}).\end{aligned}$$

where $\mathbf{v}_{(s)}$ denotes the s-th row of the matrix $(\Sigma_t)^{\frac{1}{2}}$. The optimization problem

$$\max_{\boldsymbol{\varepsilon} \in U_\varepsilon} \sum_{i=t+1}^{N} (\mathbf{v}_{(s)}^\top \boldsymbol{\varepsilon}_i)(x_{ip} - x_{iq})$$

can be rewritten in the following form:

$$\begin{aligned} \max_{\boldsymbol{\varepsilon}} \quad & (\mathbf{a}^{pqs})^\top \boldsymbol{\varepsilon} \\ \text{s.t.} \quad & \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \leq \tilde{\Gamma}, \end{aligned} \tag{2.6}$$

where vector $\mathbf{a}^{pqs}$ of dimension $(N-t) \times S$ is defined by $(\mathbf{a}^{pqs})_{is'} = (x_{ip} - x_{iq})(\Sigma_t^{\frac{1}{2}})_{ss'}$ for $i = t+1, \ldots, N$, $s' = 1, \ldots, S$.

Application of the Karush-Kuhn-Tucker conditions yields that the optimal value

of the optimization problem (2.6) is equal to

$$\sqrt{\tilde{\Gamma}} \cdot \|\mathbf{a}^{pqs}\|_2 = \sqrt{\tilde{\Gamma}} \sqrt{\sum_{s'=1}^{S} \sum_{i=t+1}^{N} (x_{ip} - x_{iq})^2 ((\Sigma_t^{\frac{1}{2}})_{ss'})^2} = \sqrt{\tilde{\Gamma}} \|\mathbf{v}_{(s)}\|_2 \sqrt{\sum_{i=t+1}^{N} (x_{ip} - x_{iq})^2}.$$

The last factor can be simplified and expressed in terms of the current time-step decision variables as follows:

$$\sum_{i=t+1}^{N} (x_{ip} - x_{iq})^2 = \sum_{i=t+1}^{N} (x_{ip}^2 - 2x_{ip}x_{iq} + x_{iq}^2) = \sum_{i=t+1}^{N} (x_{ip} + x_{iq}) = 2k - \sum_{i=1}^{t-1} (\hat{x}_{ip} + \hat{x}_{iq}) - (x_{tp} + x_{tq}),$$

where the second equality is due to the fact that x_{ip} and x_{iq} are binary variables with $x_{ip}x_{iq} = 0$. Thus, the analysis of optimization problem (2.5) allows us to write a closed-form counterpart of the linear terms in (3.1) that depends only on the current time-step decision variables x_{tp} for $p = 1, \ldots, m$, such that:

$$\begin{aligned} M_{pq}^s \geq \mu_p^s - \mu_q^s, \quad \forall \tilde{\mathbf{w}} \in U_w \quad \Longleftrightarrow \\ k\, M_{pq}^s \geq \sum_{i=1}^{t-1} (w_i^s - \bar{w}_t^s)(\hat{x}_{ip} - \hat{x}_{iq}) + (w_t^s - \bar{w}_t^s)(x_{tp} - x_{tq}) + \\ + \sqrt{\tilde{\Gamma}} \|\mathbf{v}_{(s)}\|_2 \sqrt{2k - \sum_{i=1}^{t-1} (\hat{x}_{ip} + \hat{x}_{iq}) - (x_{tp} + x_{tq})}. \end{aligned}$$

Step 2. Optimization of the variance term.

Similarly to Step 1, we fix values of p, q, s and $\tilde{\Gamma}$ and consider the optimization problem

$$\max_{\tilde{\mathbf{w}} \in U_w} (\sigma_p^s - \sigma_q^s). \tag{2.7}$$

As before, only the term representing the future time periods depends on the uncertain parameters $\boldsymbol{\varepsilon}$. Therefore, the primary goal of this step is to find a closed-form solution to the auxiliary optimization problem

$$\max_{\boldsymbol{\varepsilon} \in U_\varepsilon} \sum_{i=t+1}^{N} (\mathbf{v}_{(s)}^\top \boldsymbol{\varepsilon}_i)^2 (x_{ip} - x_{iq}) = \max_{\|\boldsymbol{\varepsilon}\|_2^2 \leq \tilde{\Gamma}} \boldsymbol{\varepsilon}^\top A \boldsymbol{\varepsilon} = \tilde{\Gamma} \cdot \lambda_{\max}(A). \tag{2.8}$$

In (2.8), $\lambda_{\max}(A)$ denotes the maximum eigenvalue of the square block matrix A:

$$A = \begin{bmatrix} (x_{t+1,p} - x_{t+1,q})B & 0 & 0 & \dots & 0 \\ 0 & (x_{t+2,p} - x_{t+2,q})B & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & (x_{Np} - x_{Nq})B \end{bmatrix},$$

where matrix $B = \mathbf{v}_{(s)}(\mathbf{v}_{(s)})^\top$.

The maximum eigenvalue $\lambda_{\max}(A)$ depends not only on the values of $\mathbf{x}$, but also on the dimension S of the covariate space.

- **Case 1.** $S \geq 2$. In this case, the eigenvalues of matrix B are 0 and $\|\mathbf{v}_{(s)}\|_2^2$, and the maximum eigenvalue of matrix A can be determined as a function of $\mathbf{x}$ as follows:

$$\lambda_{\max}(A) = \begin{cases} 0, & \text{if } x_{ip} - x_{iq} \leq 0 \text{ for all } i = t+1, \dots, N, \\ \|\mathbf{v}_{(s)}\|_2^2, & \text{if } x_{ip} - x_{iq} = 1 \text{ for at least one } i = t+1, \dots, N. \end{cases}$$

By construction, the condition $x_{ip} - x_{iq} = 1$ for at least one $i = t+1, \dots, N$ holds if and only if group p is not full after the current time-step assignment, i.e.,

$$k - \sum_{i=1}^{t-1} \hat{x}_{ip} - x_{tp} \geq 1.$$

Thus, optimization problem (2.8) has the following closed-form solution that depends only on the current time-step decision variables:

$$\max_{\boldsymbol{\varepsilon} \in U_\varepsilon} \sum_{i=t+1}^{N} (\mathbf{v}_{(s)}^\top \boldsymbol{\varepsilon}_i)^2 (x_{ip} - x_{iq}) = \tilde{\Gamma} \cdot \|\mathbf{v}_{(s)}\|_2^2 \cdot \mathbb{I}\Big\{k - \sum_{i=1}^{t-1} \hat{x}_{ip} - x_{tp} \geq 1\Big\}.$$

Now we can exploit the closed-form solution for optimization problem (2.7) within (3.1), as follows:

$$V_{pq}^s \geq \sigma_p^s - \sigma_q^s, \quad \forall \tilde{\mathbf{w}} \in U_w \quad \Longleftrightarrow$$

$$k\,V^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i - \bar{w}^s_t)^2(\hat{x}_{ip} - \hat{x}_{iq}) + (w^s_t - \bar{w}^s_t)^2(x_{tp} - x_{tq}) + \\ + \tilde{\Gamma} \cdot \|\mathbf{v}_{(s)}\|_2^2 \cdot \mathbb{I}\Big\{k - \sum_{i=1}^{t-1}\hat{x}_{ip} - x_{tp} \geq 1\Big\}.$$

- **Case 2.** $S = 1$. In this case, matrix B is one-dimensional and its only eigenvalue is $\|\mathbf{v}_{(s)}\|_2^2$. Hence,

$$\lambda_{\max}(A) = \begin{cases} \|\mathbf{v}_{(s)}\|_2^2, & \text{if } x_{ip} = 1 \text{ for at least one } i = t+1, \ldots, N, \\ -\|\mathbf{v}_{(s)}\|_2^2, & \text{if } x_{ip} = 0 \text{ and } x_{iq} = 1 \text{ for all } i = t+1, \ldots, N, \\ 0, & \text{if } x_{ip} = 0 \text{ for all } i = t+1, \ldots, N \text{ and} \\ & \quad x_{iq} = 0 \text{ for at least one } i = t+1, \ldots, N. \end{cases}$$

This is equivalent to the formulation: $\lambda_{\max}(A) = \|\mathbf{v}_{(s)}\|_2^2 \cdot \Theta_{pq}(\hat{\mathbf{x}}, \mathbf{x})$, where

$$\Theta_{pq}(\hat{\mathbf{x}}, \mathbf{x}) = \begin{cases} 1, & \text{if } k - \sum\limits_{i=1}^{t-1}\hat{x}_{ip} - x_{tp} \geq 1, \\ -1, & \text{if } k - \sum\limits_{i=1}^{t-1}\hat{x}_{ip} - x_{tp} = 0 \text{ and } \sum\limits_{i=1}^{t-1}\hat{x}_{iq} + x_{tq} + (N-t) = k, \\ 0, & \text{if } k - \sum\limits_{i=1}^{t-1}\hat{x}_{ip} - x_{tp} = 0 \text{ and } \sum\limits_{i=1}^{t-1}\hat{x}_{iq} + x_{tq} + (N-t) > k. \end{cases} \tag{2.9}$$

Thus, optimization problem (3.1) modeling the CA-RO algorithm with ellipsoidal uncertainty set has the following closed form for $S \geq 2$:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{M},\mathbf{V},z} \quad & z \\ \text{s.t.} \quad & z \geq \sum_{s=1}^{S} M^s_{pq} + \rho V^s_{pq}, \qquad \forall p < q \\ & \forall p < q, \quad s = 1, \ldots, S: \\ & k\,M^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i - \bar{w}^s_t)(\hat{x}_{ip} - \hat{x}_{iq}) + (w^s_t - \bar{w}^s_t)(x_{tp} - x_{tq}) + \end{aligned}$$

$$+\sqrt{\tilde{\Gamma}}\|\mathbf{v}_{(s)}\|_2\sqrt{2k-\sum_{i=1}^{t-1}(\hat{x}_{ip}+\hat{x}_{iq})-(x_{tp}+x_{tq})}$$

$$k\,M^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i-\bar{w}^s_t)(\hat{x}_{iq}-\hat{x}_{ip})+(w^s_t-\bar{w}^s_t)(x_{tq}-x_{tp})+ \tag{2.10}$$

$$+\sqrt{\tilde{\Gamma}}\|\mathbf{v}_{(s)}\|_2\sqrt{2k-\sum_{i=1}^{t-1}(\hat{x}_{ip}+\hat{x}_{iq})-(x_{tp}+x_{tq})}$$

$$k\,V^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i-\bar{w}^s_t)^2(\hat{x}_{ip}-\hat{x}_{iq})+(w^s_t-\bar{w}^s_t)^2(x_{tp}-x_{tq})+$$

$$+\tilde{\Gamma}\cdot\|\mathbf{v}_{(s)}\|_2^2\cdot\mathbb{I}\Big\{k-\sum_{i=1}^{t-1}\hat{x}_{ip}-x_{tp}\geq 1\Big\}$$

$$k\,V^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i-\bar{w}^s_t)^2(\hat{x}_{iq}-\hat{x}_{ip})+(w^s_t-\bar{w}^s_t)^2(x_{tq}-x_{tp})+$$

$$+\tilde{\Gamma}\cdot\|\mathbf{v}_{(s)}\|_2^2\cdot\mathbb{I}\Big\{k-\sum_{i=1}^{t-1}\hat{x}_{iq}-x_{tq}\geq 1\Big\}$$

$$\sum_{i=1}^{t-1}\hat{x}_{ip}+x_{tp}\leq k,\quad \forall p=1,\ldots,m$$

$$\sum_{p=1}^{m}x_{tp}=1$$

$$x_{ip}\in\{0,1\},\quad \forall i=t,\ldots,N,\ p=1,\ldots,m.$$

The second-to-last constraint guarantees that no group will be assigned more than k subjects and is therefore a sufficient replacement for the second-to-last constraint of formulation (3.1).

A similar formulation for the case $S=1$ is given by

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{M},\mathbf{V},z}\quad & z\\
\text{s.t.}\quad & z\geq\sum_{s=1}^{S}M^s_{pq}+\rho V^s_{pq},\qquad \forall p<q\\
& \forall p<q,\quad s=1,\ldots,S:\\
& k\,M^s_{pq}\geq\sum_{i=1}^{t-1}(w^s_i-\bar{w}^s_t)(\hat{x}_{ip}-\hat{x}_{iq})+(w^s_t-\bar{w}^s_t)(x_{tp}-x_{tq})+
\end{aligned}$$

$$+ \sqrt{\tilde{\Gamma}}\|\mathbf{v}_{(s)}\|_2\sqrt{2k - \sum_{i=1}^{t-1}(\hat{x}_{ip} + \hat{x}_{iq}) - (x_{tp} + x_{tq})}$$

$$k\, M^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i - \bar{w}^s_t)(\hat{x}_{iq} - \hat{x}_{ip}) + (w^s_i - \bar{w}^s_t)(x_{iq} - x_{ip}) + \tag{2.11}$$

$$+ \sqrt{\tilde{\Gamma}}\|\mathbf{v}_{(s)}\|_2\sqrt{2k - \sum_{i=1}^{t-1}(\hat{x}_{ip} + \hat{x}_{iq}) - (x_{tp} + x_{tq})}$$

$$k\, V^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i - \bar{w}^s_t)^2(\hat{x}_{ip} - \hat{x}_{iq}) + (w^s_t - \bar{w}^s_t)^2(x_{tp} - x_{tq}) +$$

$$+ \tilde{\Gamma} \cdot \|\mathbf{v}_{(s)}\|_2^2 \cdot \Theta_{pq}(\hat{\mathbf{x}}, \mathbf{x})$$

$$k\, V^s_{pq} \geq \sum_{i=1}^{t-1}(w^s_i - \bar{w}^s_t)^2(\hat{x}_{iq} - \hat{x}_{ip}) + (w^s_t - \bar{w}^s_t)^2(x_{tq} - x_{tp}) +$$

$$+ \tilde{\Gamma} \cdot \|\mathbf{v}_{(s)}\|_2^2 \cdot \Theta_{qp}(\hat{\mathbf{x}}, \mathbf{x})$$

$$\sum_{i=1}^{t-1} \hat{x}_{ip} + x_{tp} \leq k, \quad \forall p = 1, \ldots, m$$

$$\sum_{p=1}^{m} x_{tp} = 1$$

$$x_{ip} \in \{0, 1\}, \quad \forall i = t, \ldots, N,\ p = 1, \ldots, m,$$

where $\Theta_{pq}(\hat{\mathbf{x}}, \mathbf{x})$ and $\Theta_{qp}(\hat{\mathbf{x}}, \mathbf{x})$ are as defined in (2.9).

Formulations (2.10) and (2.11) depend only on current time-step decisions x_{tp}, for $p = 1, \ldots, m$. Given that these variables are binary and the subject with index t must be assigned to exactly one group, it is sufficient to inspect the set $\mathcal{X}$, with cardinality at most m, to solve (3.1) for the optimal current assignment. □

2.3.3 Aggregated CA-RO algorithm

The development of a partially online method is motivated by the opportunity presented when multiple subjects enroll in a clinical trial within a short period of time. Under these circumstances, the decision-maker may be able to make a joint decision regarding the simultaneous assignment of this sub-cohort of subjects to treatment groups.

For this analysis, we will distinguish the notion of time from the arrival of subjects. Time will be indexed by periods $t = 1, \ldots, T$. Subjects will be indexed separately by $i = 1, \ldots, N$ with covariate vectors $\mathbf{w}_i \in \mathbb{R}^S$, where $N \geq T$. Both the number of periods T and the number of subjects N are known *a priori*. We assume that at time t the decision-maker has observed the covariate values for $r_t \geq 1$ unassigned subjects who have arrived during time period t. Let us define $n_t := \sum_{j=1}^{t} r_j$ to represent the number of subjects who have arrived as of time t. We also introduce the vector $\mathbf{r}_t := \{r_j\}_{j=1}^{t}$. We can then define the following expressions to represent the sample mean and approximated variance of group p with respect to covariate s at time t:

$$\mu_p^s(\mathbf{r}_t) = \frac{1}{k}\Big\{ \sum_{i=1}^{n_{t-1}} w_i^s \hat{x}_{ip} + \sum_{i=n_{t-1}+1}^{n_t} w_i^s x_{ip} + \sum_{i=n_t+1}^{N} \tilde{w}_i^s x_{ip} \Big\}, \quad \text{and}$$

$$\sigma_p^s(\mathbf{r}_t) = \frac{1}{k}\Big\{ \sum_{i=1}^{n_{t-1}} (w_i^s - \bar{w}_t^s)^2 \hat{x}_{ip} + \sum_{i=n_{t-1}+1}^{n_t} (w_i^s - \bar{w}_t^s)^2 x_{ip} + \sum_{i=n_t+1}^{N} (\tilde{w}_i^s - \bar{w}_t^s)^2 x_{ip} \Big\}.$$

In the aggregated CA-RO algorithm, we solve formulation (3.1) at each time-step t, but we replace the expressions μ_p^s and σ_p^s with their generalized counterparts $\mu_p^s(\mathbf{r}_t)$ and $\sigma_p^s(\mathbf{r}_t)$, respectively. The optimal solutions $\{x_{ip}^* \in \{0,1\} \,|\, i = n_t - r_t + 1, \ldots, n_t,\ p = 1, \ldots, m\}$ to the corresponding MIO problem are used to make the assignments at period t for r_t subjects. The problem can be solved at time t by enumeration with complexity $\mathcal{O}(m^{r_t})$, and is therefore computationally tractable for instances of practical size.

If the aggregation level is uniform across time such that $r_t = r$ for all $t = 1, \ldots, T-1$ and $r_T = N - (T-1)r$, we define the CA-RO(r) algorithm with aggregation level r. We observe that the CA-RO(1) algorithm is equivalent to the fully online CA-RO algorithm and the CA-RO(N) algorithm is equivalent to the OPT algorithm from Section 2.2.

It is reasonable to assume that larger values of the aggregation parameter r lead to better performance of the partially online algorithm in terms of both covariate balance and statistical power. With a higher level of aggregation, the decision-maker has more information at the time of each decision. In Section 2.3.5, we provide

empirical evidence for this relationship.

2.3.4 Practical considerations

When using the CA-RO algorithm in practice, we suggest a few modifications and parameter selection guidelines.

1. At the beginning of the assignment process, group indices $p = 1, \ldots, m$ can be randomly assigned to each of the treatment conditions. In this way, the CA-RO algorithm is used to identify groups that are well-balanced with respect to observed covariates, but plays no role in determining which group should receive which treatment.

2. In the objective of formulation (3.1), the parameter ρ controls the tradeoff between imbalance in the sample mean and the approximated variance. In practice, to facilitate an intuitive choice of ρ, it is convenient to substitute the objective $\max_{p<q} \sum_{s=1}^{S} \left[M_{pq}^{s} + \rho \sqrt{V_{pq}^{s}} \right]$, which puts the expressions for first and second moments on the same scale. This substitution of a nonlinear objective is tractable because we are able to solve the optimization efficiently by enumeration. In the experiments that follow, we use this nonlinear objective with $\rho = 6$, which we found to yield strong results across many instances that were robust to perturbations of ρ.

3. At the beginning of the time horizon, we ensure that all groups have been randomly assigned at least one subject before we apply the optimization in formulation (3.1).

4. Toward the end of the time horizon, we set the robustness parameter $\Gamma = 0$ so as to make our algorithm more greedy and avoid overly conservative assignment decisions.

In all tested experiments, the CA-RO(r) algorithms for $r \in \{1, 3, 5\}$ produced assignment recommendations instantaneously, which suggests that the method can

be used not just for clinical trials, but also for settings requiring real-time decisions, such as Internet applications.

2.3.5 Empirical performance

In this subsection, we evaluate the empirical performance of the CA-RO algorithm. First, we review four state-of-the-art CA-RAND methods, which serve as benchmarks for the CA-RO algorithm. Second, we compare the performance of the CA-RO algorithm at various aggregation levels with pure randomization and these CA-RAND methods.

When evaluating the performance of CA-RO, we consider pure randomization (RAND) along with the matching on-the-fly algorithm of Kapelner and Krieger [2014] (KK), and three biased coin designs: the minimization method of Pocock and Simon [1975] (PS), the D_A-optimal design of Atkinson [1982] (DA), and the covariate-adaptive biased coin design of Antognini and Zagoraiou [2011] (AZ). The biased coin design methodology with $m = 2$, generically defined as

$$\phi_t = \Pr\Big(x_{t1} = 1 \Big| \, \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{t-1};\, \mathbf{w}_1, \ldots, \mathbf{w}_t\Big) = F\Big(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{t-1};\, \mathbf{w}_1, \ldots, \mathbf{w}_t\Big),$$

forms the basis of the PS, DA and AZ methods, with function $F(\cdot)$ defined separately for each method. For the PS method,

$$\phi_t = \begin{cases} \frac{1}{2}, & \text{if } D(t) = 0, \\ p, & \text{if } D(t) < 0, \\ 1 - p, & \text{if } D(t) > 0, \end{cases}$$

where p is the bias parameter and $D(t)$ represents the covariate imbalance between

the two groups after $t-1$ subjects have been assigned. For the DA method,

$$\phi_t = \frac{(1-\zeta)^2}{(1-\zeta)^2 + (1+\zeta)^2}, \quad \text{where} \quad \zeta = \frac{w_t \sum_{i=1}^{t-1} w_i (x_{i1} - x_{i2})}{\sum_{i=1}^{t-1} w_i^2}$$

for the case of one-dimensional covariates. For the AZ method, we have $\phi_t = G_j(D_t(\omega_j))$, where

$$G_j(\zeta) = \begin{cases} \frac{1}{2}, & 0 \leq \zeta \leq 1, \\ (\zeta^J + 1)^{-1}, & \zeta > 1, \end{cases} \qquad \text{and} \quad G_j(-\zeta) + G_j(\zeta) = 1, \, \forall \zeta \in \mathbb{Z},$$

for discrete levels of the covariate space indexed by $j = 0, \ldots, J$. In this description, $D_t(\omega_j)$ denotes the imbalance between the two groups within the level ω_j. In the KK method, subjects are either randomized to treatment groups or paired via a matching criterion based on the pairwise Mahalanobis distance. In the latter case, the new paired subject is assigned to the treatment opposite its pair in order to balance the groups.

We now discuss the empirical performance of the various algorithms with respect to covariate balance. For $N \in \{20, 60, 100\}$ and $m = 2$, we simulated 3,000 unique sets of covariate values drawn *i.i.d.* from a standard normal distribution. We evaluated nine algorithms - RAND, PS, DA, AZ, KK and CA-RO(r) (with four different values of r) - to measure the average worst pairwise difference in generalized moments across groups (Table 2.1). For this and all subsequent experiments when evaluating the CA-RO algorithm at any level of aggregation, we chose the robustness parameter Γ in uncertainty set (2.3) independently and uniformly at random from the interval [0.5, 4] at each time-step. In terms of the discrepancy in the first moment, CA-RO was always among the best methods. The discrepancy in the second moment, which closely approximates the discrepancy in the variance in this setting, was always lower for CA-RO than for the best CA-RAND method. The discrepancy in higher moments, as well as generalized moments of $\log(|w|)$ and $1/w$, for CA-RO methods was

always comparable with other CA-RAND algorithms. As we expected, the advantage of optimization increases with r, and the offline OPT algorithm has starkly better performance than other approaches.

Table 2.1: Average between-group absolute discrepancy in moments under allocation algorithms with $m = 2$ and $S = 1$.

		Moment						
N	Algorithm	1	2	3	4	5	$\log(\lvert w\rvert)$	$1/w$
20	RAND	0.358	0.498	1.321	2.955	8.108	0.387	13.521
	PS	0.260	0.509	1.120	2.872	7.389	0.524	13.150
	DA	0.167	0.616	0.931	3.172	6.953	0.719	13.492
	AZ	0.286	0.553	1.228	3.070	7.881	0.558	13.143
	KK	0.221	0.416	1.046	2.706	7.186	0.305	13.307
	CA-RO(1)	0.250	0.269	1.179	2.196	7.759	0.335	13.439
	CA-RO(3)	0.251	0.226	1.228	2.033	8.022	0.340	13.639
	CA-RO(5)	0.254	0.186	1.224	1.954	7.969	0.343	13.621
	OPT	0.024	0.010	0.960	1.517	7.348	0.354	13.702
60	RAND	0.205	0.292	0.793	1.869	5.396	0.224	10.020
	PS	0.125	0.250	0.625	1.640	4.746	0.257	9.850
	DA	0.092	0.350	0.560	1.935	4.700	0.408	9.990
	AZ	0.125	0.278	0.663	1.788	4.981	0.257	9.763
	KK	0.172	0.274	0.708	1.788	5.144	0.214	9.853
	CA-RO(1)	0.099	0.139	0.629	1.378	4.992	0.214	9.979
	CA-RO(3)	0.095	0.090	0.645	1.176	5.044	0.210	9.999
	CA-RO(5)	0.096	0.067	0.653	1.128	5.112	0.206	10.028
	OPT	0.001	3.33×10^{-4}	0.531	1.046	4.668	0.272	10.255
100	RAND	0.161	0.225	0.604	1.470	4.257	0.177	10.507
	PS	0.083	0.178	0.438	1.242	3.619	0.182	10.251
	DA	0.072	0.274	0.436	1.538	3.816	0.324	10.080
	AZ	0.088	0.190	0.482	1.320	3.839	0.181	9.933
	KK	0.133	0.218	0.544	1.449	4.107	0.171	10.195
	CA-RO(1)	0.066	0.116	0.479	1.132	4.034	0.168	10.223
	CA-RO(3)	0.063	0.073	0.488	0.984	4.091	0.165	10.368
	CA-RO(5)	0.064	0.051	0.485	0.901	4.041	0.163	10.551
	OPT	0.001	1.14×10^{-4}	0.402	0.806	3.692	0.218	10.240

We found similar results from additional experiments in which the covariates were generated from alternative distributions, including uniform and long-tailed Cauchy distributions.

2.4 Statistical Power of CA-RO Algorithm

A common pre-condition for the approval of any clinical trial is to demonstrate that the trial will have a sample size sufficient to make sound statistical inferences with high probability. These inferences include both statistical power, the ability to detect a positive treatment effect when one exists, and a low type I error rate, the ability to correctly identify an ineffective treatment. In classical statistical models, the power of a randomized controlled trial can be derived from the sample size and significance level, given an estimated treatment effect. Randomized allocation can yield an accidental imbalance in covariates between treatment groups that can impact the ability to make experimental inferences. Traditionally, when estimating treatment effects, practitioners have been satisfied to control for this covariate imbalance *a posteriori* via regression methods (Lin, 2013).

We provide strong empirical evidence that such post hoc adjustments may produce suboptimal effect estimation, particularly when the relationship between covariates and response is nonlinear. By testing a variety of covariate-response models, we show that, at any fixed sample size, the statistical power of a clinical trial is at least as high when covariate-adaptive optimization is used rather than covariate-adaptive randomization. In settings where the covariate-response relationship is nonlinear, we observe that the power under the CA-RO algorithm is significantly higher than for state-of-the-art CA-RAND methods. Therefore, in certain settings, the use of covariate-adaptive optimization could allow decision-makers to achieve desired levels of statistical power with significantly smaller sample size as compared with CA-RAND mechanisms. Given the high cost of enrolling human subjects in clinical trials, the ability to achieve needed statistical power with much smaller sample size can result in significant cost savings for the healthcare industry and society at large.

2.4.1 Test for statistical power

In order to compare statistical power under CA-RAND and CA-RO online allocation procedures, we apply a hypothesis testing framework based on simulation (Bertsimas

et al., 2015).

Let us assume there are $m = 2$ groups: a treatment group, which will be administered a given therapy, and a control group, which will be administered a placebo. There are N subjects in the trial, such that $k = N/2$ subjects will be assigned to each of the groups. At each time-step $t = 1, \ldots, N$, the decision-maker observes the values of a covariate vector $\mathbf{w}_t$ and makes a binary assignment x_t, where $x_t = 1$ indicates the treatment group (1) and $x_t = 0$ indicates the control group (0). Let v_t be the response measured after the assigned treatment was administered for subject t. We adopt the potential outcomes framework of Rosenbaum and Rubin [1983], such that each subject has a pair of potential outcomes $(v_t^{(1)}, v_t^{(0)})$, where the superscript indicates treatment or control and only one of these two outcomes can be observed. Under this framework, we have the following relationship between the observed response and the potential outcomes: $v_t = v_t^{(1)} x_t + v_t^{(0)}(1 - x_t)$.

Given v_t for each subject $t = 1, \ldots, N$, we can estimate the average treatment effect $\hat{\delta}$. We adopt two estimators for $\hat{\delta}$, unadjusted and regression-adjusted, from Lin [2013]:

1. $\hat{\delta}_{\text{unadj}} := \frac{1}{k}\Big[\sum_{t=1}^{N} v_t x_t - \sum_{t=1}^{N} v_t(1 - x_t)\Big]$
2. $\hat{\delta}_{\text{adj}} := \beta_x$, where β_x is the estimated coefficient on x_t in the ordinary least squares regression $v_t = \beta_0 + \beta_x x_t + \boldsymbol{\beta}_w^\top \mathbf{w}_t$.

We can also consider a modification of $\hat{\delta}_{\text{adj}}$ in which the linear regression contains quadratic terms in $\mathbf{w}_t$. We hypothesize that this modified estimator may control for covariates in the first and second moments *a posteriori*.

To test the significance of this observed effect $\hat{\delta}$, we adopt Fisher's sharp null hypothesis [Fisher, 1935], which states that every subject $t = 1, \ldots, N$ would have had the same response to treatment regardless of which treatment was assigned. In other words, under the sharp null hypothesis, we have $v_t = v_t^{(1)} = v_t^{(0)}$. Equipped with a complete set of potential outcomes for each subject, we can estimate the average treatment effect under alternative random allocations of subjects $1, \ldots, N$. If we compute the average treatment effect δ_b as our test statistic for each alternative

allocation $b = 1, \ldots, B$, we can then estimate the p-value for our observed $\hat{\delta}$, using a two-sided test, as

$$p = \frac{1}{1+B}\Big(1 + \sum_{b=1}^{B} \mathbb{I}\Big[\big|\delta_b\big| \geq \big|\hat{\delta}\big|\Big]\Big).$$

We reject the null hypothesis if $p \leq \alpha$ for some pre-specified significance level α; otherwise, we accept the null hypothesis.

In order to estimate the statistical power under a given algorithm $\mathbb{A}$, we generate Q random samples of N subjects with covariates drawn *i.i.d.* from a fixed distribution. We apply the hypothesis test described above for all random samples, and measure the number of samples Q_{reject} for which the null hypothesis is rejected. We evaluate the probability that the null hypothesis will be rejected by computing the ratio $\lambda := Q_{\text{reject}}/Q$. If the true treatment effect δ_0 is nonzero, then λ estimates the power of the experiment; otherwise, λ estimates the type I error rate.

The alternative allocations for the hypothesis test can be generated randomly using Monte Carlo simulation to approximate the distribution of possible allocations under random assignment mechanism $\mathbb{A}$. If Monte Carlo simulation does not yield a sufficiently diverse set of allocations within computational limits, one can generate bootstrapped resamples of covariate vectors $\mathbf{w}_t^b$, $t = 1, \ldots, N$ drawn uniformly at random from the set $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ (Efron and Tibshirani, 1994). Based on the observations from the original experiment and under the null hypothesis, we have complete mappings $v^{(1)}(\cdot) : \mathcal{W} \to \mathbb{R}$ and $v^{(0)}(\cdot) : \mathcal{W} \to \mathbb{R}$, which represent the potential outcomes under treatment and control, respectively, for individuals with covariates in $\mathcal{W}$. Therefore, for each subject in a given bootstrapped sample, we observe the response under her random allocation x_t^b as $v_t^b = v^{(1)}(\mathbf{w}_t^b) \cdot x_t^b + v^{(0)}(\mathbf{w}_t^b) \cdot (1 - x_t^b)$.

2.4.2 Computational results

To evaluate the statistical power of the CA-RO algorithm relative to CA-RAND methods, we simulated clinical trials under three different hidden realities, each characterized by a unique model relating treatment response to subject covariates. We

assumed each subject $t = 1, \ldots, N$ had a covariate vector $\mathbf{w}_t = (w_t^1, w_t^2)$ of dimension $S = 2$, whose components were drawn *i.i.d.* from a standard normal distribution. The covariate-response models were as follows:

- Nonlinear (NL): $v_t = \delta_0 x_t + (w_t^1)^2 - (w_t^2)^2 + \epsilon_t$,
- Linear (LIN): $v_t = \delta_0 x_t + 2(w_t^1) + 2(w_t^2) + \epsilon_t$,
- No relationship (NR): $v_t = \delta_0 x_t + \epsilon_t$,

where δ_0 is the ground-truth additive treatment effect and ϵ_t is a Gaussian noise term with mean 0 and standard deviation 0.75.

For each covariate-response model, we evaluated statistical power λ under the CA-RAND and CA-RO algorithms by applying the hypothesis test described in Section 2.4.1 with both estimators $\hat{\delta}_{\text{unadj}}$ and $\hat{\delta}_{\text{adj}}$ (Figure 2-1). We considered $N \in \{40, 80, 120\}$ with $\delta_0 = 0.5$, $Q = 800$, $B = 500$, and significance level $\alpha = 0.05$. For all scenarios, the power of the experiment increases with N.

- In the NR scenario, post hoc regression adjustment does not improve power for any of the methods. All methods yield similar power since there is no benefit from covariate balance.
- Conversely, in the linear response setting (LIN), regression adjustment increases statistical power substantially for all methods. When using the $\hat{\delta}_{\text{unadj}}$ estimator, CA-RO(1) yields higher power relative to randomization and CA-RAND methods. However, post hoc regression adjustment using ordinary least squares, which exactly replicates the covariate-response model with additive treatment effect, reduces the need for the *a priori* covariate balance provided by CA-RO. Power evaluated using $\hat{\delta}_{\text{adj}}$ is equally high across all methods.
- Finally, in the nonlinear response scenario described above (NL), there is virtually no benefit to using regression adjustment. In this setting, CA-RO(1) yields much higher statistical power than pure randomization and CA-RAND methods. The advantage of CA-RO grows with the sample size N.

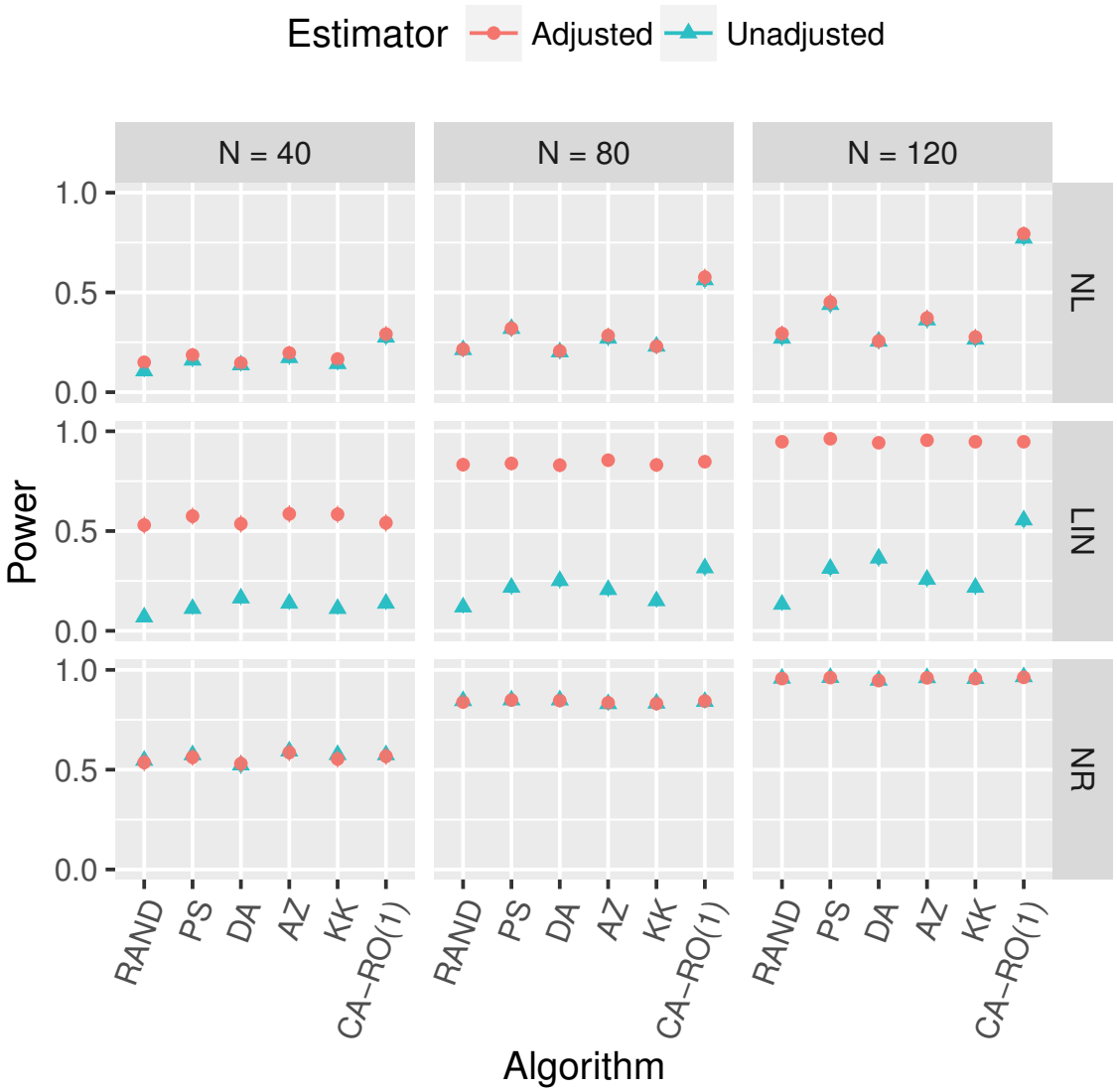


Figure 2-1: Statistical power (with 95% confidence intervals) under CA-RO(1) vs. CA-RAND methods.

Results are shown for $N \in \{40, 80, 120\}$ under various response models (NL, LIN, NR), using both adjusted and unadjusted treatment effect estimators.

We conducted additional experiments under a variety of nonlinear models and found, in all tested scenarios, that CA-RO had power at least as high as (and often higher than) randomization-based methods. The NL scenario is an example in which the benefit of CA-RO was particularly dramatic. When the nonlinear models included terms that were cubic, quadratic, or exponential in the covariates, the CA-RO method had power that exceeded other methods as measured using the $\hat{\delta}_{\text{unadj}}$ estimator; however, the observed power was relatively similar across all methods when including quadratic terms in the modified regression-adjusted $\hat{\delta}_{\text{adj}}$ estimator. It seems that controlling for both linear and quadratic terms via regression can produce some of the benefits of *a priori* covariate balance, but at the cost of using a less transparent effect estimator.

We also ran simulations in which the p-values were estimated using a one-sided test rather than a two-sided test. As one might expect, for fixed δ_0 and distribution of noise ϵ, power was higher for all methods under the one-sided test. However, in all tested scenarios, the CA-RO algorithm maintained a similar advantage relative to other methods.

In Table 2.2, we show the results for the NL setting under the CA-RO(r) assignment mechanism for $N = 40$ with aggregation levels $r \in \{1, 3, 5\}$ along with OPT, which is equivalent to CA-RO(N). As we expect, the power increases with the aggregation level r.

Table 2.2: Statistical power under CA-RO(r) for NL scenario with $N = 40$.

Aggregation level, r	1	3	5	N
Power, λ	29.1%	29.8%	31.9%	36.4%

In Figure 2-1, we show that, under some covariate-response models, CA-RO yields higher power at fixed sample sizes than other methods. This motivates a complementary question: What is the sample size required to achieve a desired level of statistical power? We considered the NL scenario and tested values of δ_0 from 0.75 to 1.75 to estimate $N^*_{\mathbb{A}}(\delta_0)$, the minimum number of subjects per group needed to achieve power of at least 80% when assignment mechanism $\mathbb{A}$ is employed (Figure 2-2a). With a

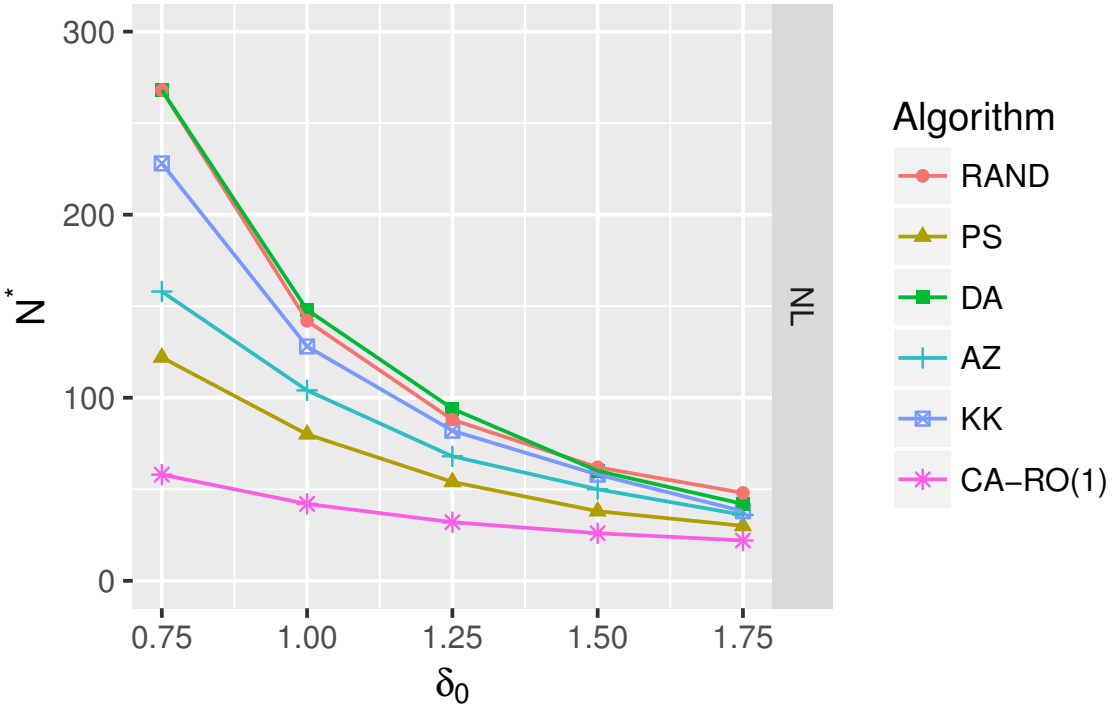
large effect size of $\delta_0 = 1.75$, statistical power of 80% was achieved with a sample size of 22 using the CA-RO(1) algorithm compared with a sample size of 30 using the best CA-RAND method (in this case, PS). With a small effect size of $\delta_0 = 0.75$, the advantage of optimization was even bigger; a sample size of 58 was sufficient to achieve 80% power, compared with a sample size of 122 using the best CA-RAND method (again, PS). For a given treatment effect δ_0, the threshold sample size needed to achieve 80% power under the CA-RO algorithm was reduced by at least 25% relative to the best CA-RAND method (Figure 2-2b). If we consider the NL setting with $\delta_0 = 0.75$ as an example, the CA-RO method may enable the execution and analysis of some clinical trials that would otherwise be infeasible given the prohibitively large sample size required to achieve a sufficient level of statistical power when CA-RAND methods are employed.

Table 2.3 demonstrates that the minimum sample size $N^*_{\mathbb{A}}(\delta_0)$ decreases further as the aggregation level r of CA-RO(r) algorithm grows. Relative to state-of-the-art CA-RAND methods, the CA-RO approach can dramatically reduce the number of subjects enrolled in a trial without sacrificing statistical power.

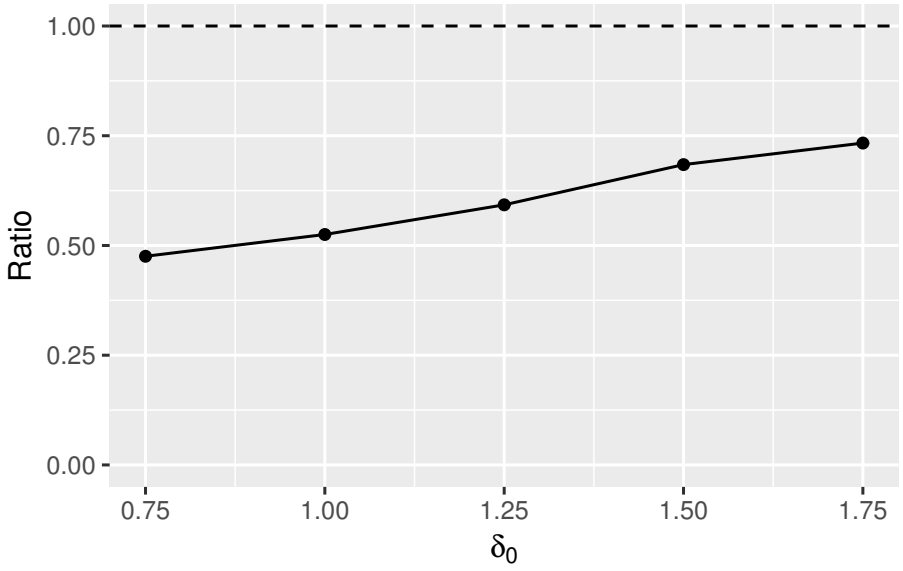
Table 2.3: Minimum number of subjects per group $N^*_{\mathbb{A}}(\delta_0)$ needed for power over 80%.

	Treatment effect, δ_0				
Algorithm, $\mathbb{A}$	0.75	1	1.25	1.5	1.75
RAND	268	142	88	62	48
PS	122	80	54	38	30
DA	268	148	94	60	42
AZ	158	104	68	50	36
KK	228	128	82	58	38
CA-RO(1)	58	42	32	26	22
CA-RO(3)	52	36	28	22	18
CA-RO(5)	48	32	26	22	18
OPT	26	22	18	16	14

We also evaluate the rate of type I errors for CA-RO(1) with $\delta_0 = 0$ with $Q = 800$ and $B = 500$ for $N \in \{40, 80, 120\}$ and for each of the three covariate-response scenarios described above, using the regression-adjusted treatment effect estimator (Table 2.4). We observe that, for each setting, the type I error is a decreasing function



(a) Sample size needed to achieve statistical power of at least 80%.



(b) Ratio of $N^*(\delta_0)$ for CA-RO(1) vs. best CA-RAND algorithm (PS).

Figure 2-2: Sample size needed under CA-RO(1) vs. CA-RAND methods.

of N. Type I error rates for all algorithms tested are shown in Figure 2-3. PS, the CA-RAND method with the best statistical power in this experiment, had a mean type I error rate that was uniformly higher than that produced by CA-RO(1).

Table 2.4: Type I error under CA-RO(1).

	Sample size, N		
Scenario	40	80	120
Nonlinear (NL)	7.1%	4.6%	4.5%
Linear (LIN)	7.0%	6.0%	4.5%
No relationship (NR)	6.5%	5.3%	5.6%

2.5 Unbiasedness of CA-RO Approach

In this section, we provide empirical and theoretical evidence that the CA-RO algorithm introduced in Section 2.3 exhibits the same statistical advantages ascribed to complete randomization by Efron [1971]: freedom from selection bias, freedom from accidental bias with respect to observed and hidden covariates, and a reasoned basis for inference.

2.5.1 Freedom from selection bias

The CA-RO algorithm protects against selection bias, the possibility that an investigator could consciously or unconsciously influence the order of subject enrollment based on deterministic knowledge of the next treatment assignment. Through computer simulation, we demonstrate that, by selecting the robustness parameter Γ independently and uniformly at random with support [0.5, 4] at each time-step, the CA-RO method yields sufficiently random treatment assignments as to protect against this type of selection bias. For N from 30 to 100, we randomly generated 30 unique sequences of covariates $\mathbf{w} \in \mathbb{R}^N$ drawn independently from $\mathcal{N}(0, 1)$. We used the CA-RO(1) algorithm to generate 3,000 random assignments of the N subjects to two groups. We observe that one cannot determine the sequence of future assignments based on knowledge of the algorithm because, on average, the total number of possible alloca-

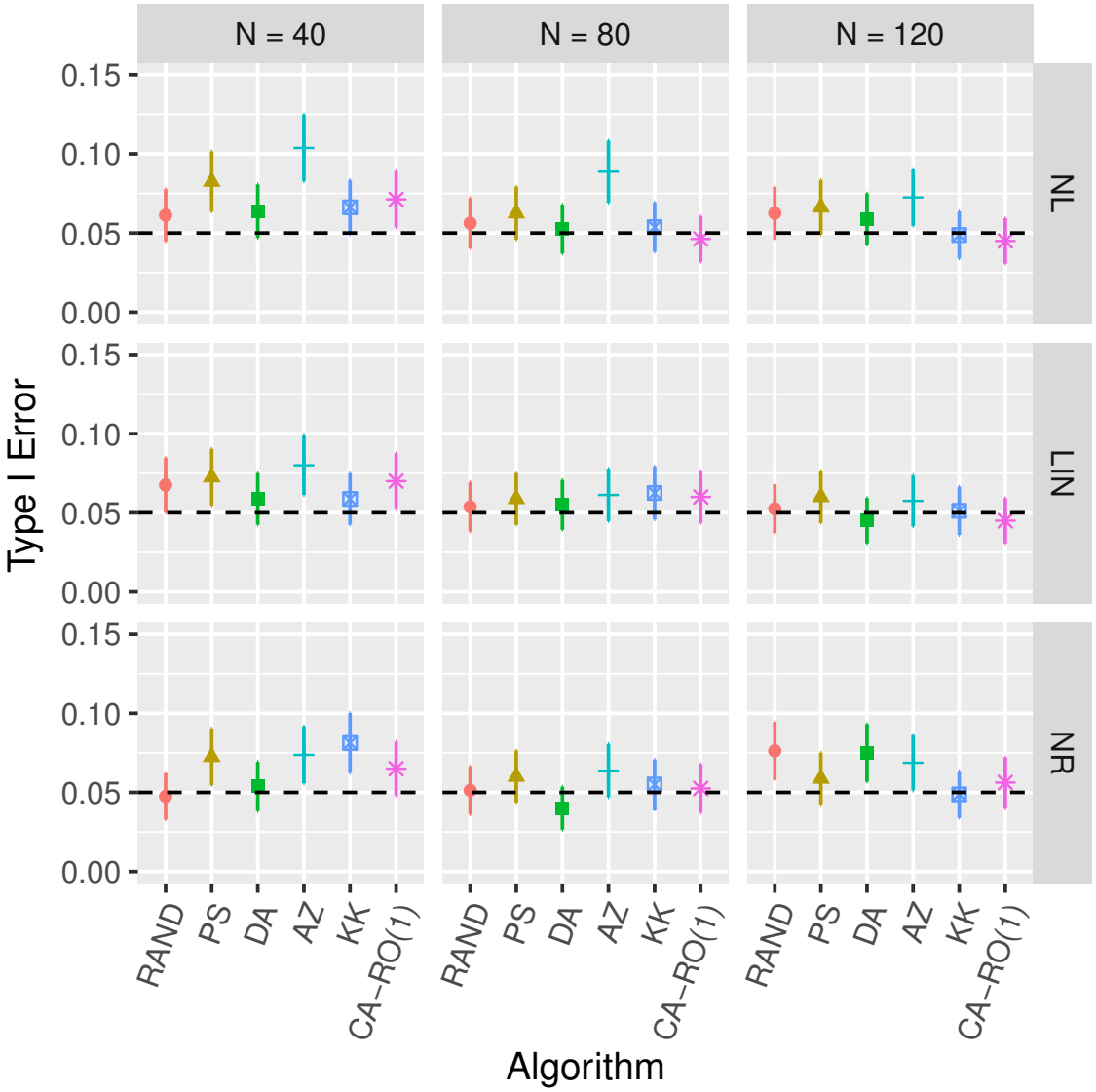
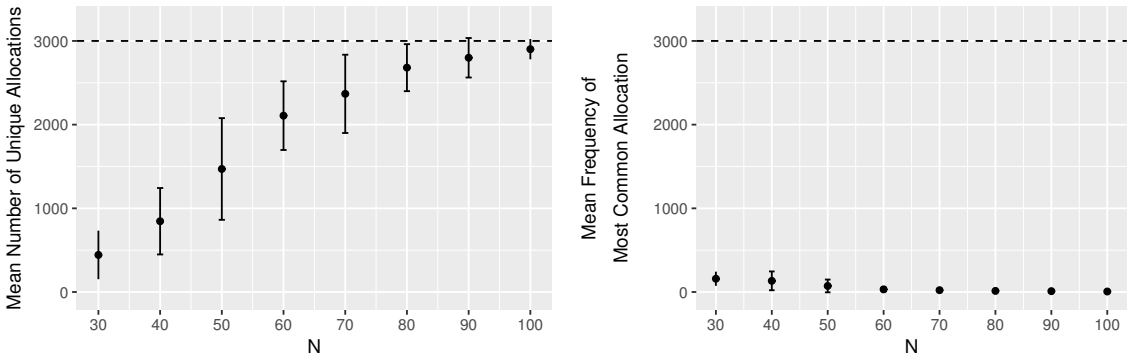


Figure 2-3: Type I error (with 95% confidence intervals) with CA-RO(1) vs. CA-RAND methods.

Results are shown for $N \in \{40, 80, 120\}$ under various response models (NL, LIN, NR), using the adjusted treatment effect estimator. Dashed line indicates 0.05 significance level.

tions is large (Figure 2-4a) and no individual assignment sequence has a likelihood higher than 6% (Figure 2-4b).



(a) Mean (with standard deviation) number of unique allocations among 3,000 simulations.

(b) Mean (with standard deviation) frequency of most common allocation among 3,000 simulations.

Figure 2-4: Analysis of distribution of unique allocations under CA-RO(1).

A concern that directly competes with selection bias is certifiability. When used with a fixed and predefined sequence of robustness parameters Γ_t, $t = 1, \ldots, N$, the CA-RO algorithm is a sequence of deterministic optimization problems, each of which can be reproduced. This reproducibility provides a natural method for certifying *a posteriori* that the algorithm's recommendation was followed, given knowledge of the subjects' covariates and arrival order. If certifiability is deemed to be of greater concern than selection bias in the context of a particular experimental setting, one can apply the CA-RO algorithm with fixed robustness parameters in order to achieve full certifiability of assignments. Conversely, certifiability is not achievable using randomized methods unless the random seed used to initialize the algorithm is provided.

2.5.2 Freedom from accidental covariate imbalance

We have shown in the empirical results from Section 2.3.5 that the CA-RO method produces consistently better balance in the first two moments across groups compared with simple randomization and other existing CA-RAND approaches. In this subsection, we show that, despite only considering the observed covariates $\mathbf{w} \in \mathbb{R}^N$ when making assignment decisions, the CA-RO algorithm provides the same level of

protection as CA-RAND methods against irregular allocation with respect to other, potentially unseen factors.

We consider two natural cases for the dependence of hidden factors on the observed covariates $\mathbf{w}$: no correlation and continuous dependence.

1. If there is a hidden factor that is uncorrelated with observed covariates $\mathbf{w}$, the CA-RO algorithm generates an allocation which is as random with respect to the hidden covariates as that produced by randomized methods.

2. The second case, when the unseen covariate is a continuous function of the observed covariate, warrants further discussion. We see empirically that, when unseen factor f has a polynomial or logarithmic conditional expectation in scalar random variable w, the discrepancy in higher moments and generalized moments $f = \log(|w|)$ and $f = 1/w$ for CA-RO methods is always comparable with (and often lower than) the mismatch produced by CA-RAND algorithms (Table 2.1). In the remainder of this subsection, we present formal theoretical evidence that this empirical relationship extends to the general case of continuous dependence.

To examine this general case, we assume that there are two different assignment algorithms $\mathbb{A}$ and $\mathbb{B}$ (e.g. CA-RO(1) and PS), and an unseen factor f that can be modeled in the form

$$f = g(w) + \epsilon,$$

where $g(\cdot)$ is a Lipschitz function with constant L and ϵ is some noise function. When generating groups of size k by algorithm $\mathbb{A}$, let us denote the maximum discrepancy in means with respect to unseen covariate f by:

$$z_{\mathbb{A}}^{f} := \max_{p<q} \frac{1}{k} \Big| \sum_{i \in \mathrm{I}_p(\mathbb{A})} g(w_i) - \sum_{i \in \mathrm{I}_q(\mathbb{A})} g(w_i) \Big|,$$

where $I_p(\mathbb{A})$, $I_q(\mathbb{A}) \subset \{1, \ldots, N\}$ are disjoint index sets respectively describing groups p and q produced by algorithm $\mathbb{A}$. The maximum discrepancy $z_{\mathbb{B}}^{f}$ between groups generated by algorithm $\mathbb{B}$ is defined analogously.

In Proposition 1, we derive theoretical upper bounds on the the values of $z^f_{\mathbb{A}}$ and $|z^f_{\mathbb{A}} - z^f_{\mathbb{B}}|$, where $\mathbb{A}$ is the CA-RO(1) approach and $\mathbb{B}$ is any CA-RAND method. The first upper bound on $z^f_{\mathbb{A}}$, given by (2.12a), demonstrates that the maximum discrepancy in means with respect to unseen covariate f is controlled by the corresponding discrepancy with respect to the observed covariate w. The second upper bound on $|z^f_{\mathbb{A}} - z^f_{\mathbb{B}}|$, given by (2.12b), indicates that the maximum discrepancy in means with respect to unseen covariate f is as well-controlled under CA-RO(1) as under any other CA-RAND algorithm.

Proposition 1. Let us consider the simplest case where subjects with scalar covariates $w_i, i = 1, \ldots, 2k$ are assigned to $m = 2$ groups. For any assignment algorithms $\mathbb{A}$ and $\mathbb{B}$ that produce groups of equal size k, and for any Lipschitz function $g(\cdot) \in \mathrm{Lip}(L)$, the following inequalities hold:

$$z^f_{\mathbb{A}} \leq L \cdot \theta^*(\mathbb{A}), \tag{2.12a}$$

$$|z^f_{\mathbb{A}} - z^f_{\mathbb{B}}| \leq 2L \cdot \xi^*(\mathbb{A}, \mathbb{B}). \tag{2.12b}$$

In (2.12a), $\theta^*(\mathbb{A})$ is the optimal objective value of the auxiliary pairwise matching problem:

$$\begin{aligned} \theta^*(\mathbb{A}) := \min_{\mathbf{y}} \quad & \tfrac{1}{k} \sum_{i \in \mathrm{I}_1(\mathbb{A})} \sum_{j \in \mathrm{I}_2(\mathbb{A})} |w_i - w_j| \, y_{ij} \\ \text{s.t.} \quad & \sum_{i \in \mathrm{I}_1(\mathbb{A})} y_{ij} = 1, \quad \forall j \in \mathrm{I}_2(\mathbb{A}) \\ & \sum_{j \in \mathrm{I}_2(\mathbb{A})} y_{ij} = 1, \quad \forall i \in \mathrm{I}_1(\mathbb{A}) \\ & y_{ij} \in \{0, 1\}. \end{aligned} \tag{2.13}$$

In (2.12b), we define the value $\xi^*(\mathbb{A}, \mathbb{B}) := \min_{c=1,2} \xi_c(\mathbb{A}, \mathbb{B})$, where $\xi_c(\mathbb{A}, \mathbb{B})$ is the optimal

value of the problem

$$\begin{aligned} \xi_c(\mathbb{A},\mathbb{B}) := \min_{\mathbf{y}} \quad & \sum_{i\in S_1^c}\sum_{j\in S_2^c} |w_i - w_j|\, y_{ij} \\ \text{s.t.} \quad & \sum_{i\in S_1^c} y_{ij} = 1, \quad \forall j \in S_2^c \\ & \sum_{j\in S_2^c} y_{ij} = 1, \quad \forall i \in S_1^c \\ & y_{ij} \in \{0,1\}, \end{aligned} \tag{2.14}$$

and auxiliary sets of indices S_α^β, for $\alpha, \beta = 1, 2$ have the form

$$\begin{aligned} S_1^1 &= \mathrm{I}_1(\mathbb{A}) \cap \mathrm{I}_2(\mathbb{B}) \quad \text{and} \quad S_2^1 = \mathrm{I}_2(\mathbb{A}) \cap \mathrm{I}_1(\mathbb{B}); \\ S_1^2 &= \mathrm{I}_1(\mathbb{A}) \cap \mathrm{I}_1(\mathbb{B}) \quad \text{and} \quad S_2^2 = \mathrm{I}_2(\mathbb{A}) \cap \mathrm{I}_2(\mathbb{B}). \end{aligned}$$

(These sets describe the differences between the groups produced by algorithms $\mathbb{A}$ and $\mathbb{B}$.)

Proof. In order to verify inequality (2.12a) for $m = 2$, we note that optimization problem (2.13) uniquely determines a pairwise matching of sets $\mathrm{I}_1(\mathbb{A})$ and $\mathrm{I}_2(\mathbb{A})$ with minimum average distance between pairs. We denote the resulting pairs as $\{(i_l, j_l) : l = 1, \ldots, k\}$, where k is the number of indices in each set. By the definition of $z_{\mathbb{A}}^f$, we derive

$$z_{\mathbb{A}}^f = \frac{1}{k}\Big| \sum_{i\in \mathrm{I}_1(\mathbb{A})} g(w_i) - \sum_{i\in \mathrm{I}_2(\mathbb{A})} g(w_i)\Big| \le \frac{1}{k}\sum_{l=1}^{k} |g(w_{i_l}) - g(w_{j_l})| \le \frac{L}{k}\sum_{l=1}^{k} |w_{i_l} - w_{j_l}| = L\cdot\theta^*(\mathbb{A}). \tag{2.15}$$

Similar reasoning is applicable for the second inequality (2.12b). First, it is easy to see that the cardinality of both sets S_1^1 and S_2^1 is the same:

$$\gamma := |S_1^1| = |S_2^1|.$$

By symmetry, the cardinalities of the complementary sets are also identical:

$$|S_1^2| = |S_2^2| = k - \gamma.$$

The next step is to express the between-group discrepancies in the means generated by algorithms $\mathbb{A}$ and $\mathbb{B}$ as follows:

$$z^f_{\mathbb{A}} = \frac{1}{k}\Big|\sum_{i\in S^2_1} g(w_i) + \sum_{i\in S^1_1} g(w_i) - \sum_{i\in S^1_2} g(w_i) - \sum_{i\in S^2_2} g(w_i)\Big| = |a+b|.$$

$$z^f_{\mathbb{B}} = \frac{1}{k}\Big|\sum_{i\in S^2_1} g(w_i) - \sum_{i\in S^1_1} g(w_i) + \sum_{i\in S^1_2} g(w_i) - \sum_{i\in S^2_2} g(w_i)\Big| = |a-b|,$$

where

$$a := \frac{1}{k}\Big(\sum_{i\in S^2_1} g(w_i) - \sum_{i\in S^2_2} g(w_i)\Big) \quad \text{and} \quad b := \frac{1}{k}\Big(\sum_{i\in S^1_1} g(w_i) - \sum_{i\in S^1_2} g(w_i)\Big).$$

Hence, analogously to argument (2.15), one may obtain upper bounds:

$$|a| \le L\cdot\xi_2(\mathbb{A},\mathbb{B}) \quad \text{and} \quad |b| \le L\cdot\xi_1(\mathbb{A},\mathbb{B}). \tag{2.16}$$

A simple corollary from the triangle inequality is that, for any a and b,

$$\big||a+b| - |a-b|\big| \le 2\min\big(|a|,|b|\big).$$

This corollary, taken together with (2.16), implies that

$$|z^f_{\mathbb{A}} - z^f_{\mathbb{B}}| \le 2L\cdot\min\big\{\xi_1(\mathbb{A},\mathbb{B}),\, \xi_2(\mathbb{A},\mathbb{B})\big\} \le 2L\cdot\xi^*(\mathbb{A},\mathbb{B}).$$

This proposition has a straighforward extension to the cases of $m > 2$ groups and multidimensional covariates. The proofs have a similar structure to the case considered here, and thus are omitted. □

Having obtained theoretical upper bounds (2.12), we conducted numerical experiments to measure the values of the average pairwise distances $\theta^*(\mathbb{A})$ and $\xi^*(\mathbb{A},\mathbb{B})$ defined in (2.13) and (2.14), respectively. We fixed $\mathbb{A}$ = CA-RO(1) and chose $\mathbb{B}$ from among RAND, PS, DA, AZ and KK, where the randomization methods were modi-

fied to ensure they would produce equal-sized groups at the end of the horizon. We randomly generated populations of size N between 60 and 100, where each subject had a scalar standard normal covariate w_i. After executing both chosen algorithms $\mathbb{A}$ and $\mathbb{B}$, we identified the index sets S_α^β, for $\alpha, \beta = 1, 2$ and solved the auxiliary optimization problems (2.13) and (2.14). After 3,000 simulations, we observed that, in more than 99% of instances, the average discrepancy $\theta^*\big(\text{CA-RO}(1)\big) \leq 0.35$; the corresponding upper bounds for $\xi^*\big(\text{CA-RO}(1), \mathbb{B}\big)$ for various choices of $\mathbb{B}$ are reported in Table 2.5. Given that, by definition, the distances θ^* and ξ^* scale linearly with respect to the covariates w_i, $i = 1, \ldots, N$, one may derive the empirical counterparts of upper bounds (2.12), which hold with high probability:

$$z^f_{\text{CA-RO}(1)} \leq 0.35\, L \cdot \sigma$$
$$\max_{\mathbb{B} \in \{\text{RAND, PS, DA, AZ, KK}\}} |z^f_{\text{CA-RO}(1)} - z^f_{\mathbb{B}}| \leq 0.51\, L \cdot \sigma$$

where σ is the standard deviation of attributes w. The constant 0.51 in the right-hand side of the second bound is derived from the maximum discrepancy among CA-RAND methods in Table 2.5.

Table 2.5: Empirical upper bound on $\xi^*\big(\text{CA-RO}(1), \mathbb{B}\big)$.

Algorithm $\mathbb{B}$	RAND	PS	DA	AZ	KK
$\xi^*\big(\text{CA-RO}(1), \mathbb{B}\big)$	0.255	0.187	0.200	0.185	0.215

The result of Proposition 1 can be extended to the case of general continuous functions $g(\cdot)$ under the assumption that the support $K \subset \mathbb{R}^S$ of covariates $\mathbf{w}$ is a compact set. Indeed, any continuously differentiable function $g(\cdot)$ defined on a compact set K (including any polynomial function) is in a Lipschitz class with $L = \max_{x \in K} |g'(x)|$. Since any continuous function on K can be approximated with arbitrary precision by some polynomial according to the Weierstrass theorem, the upper bounds (2.12) hold for any continuous function $g(\cdot)$ on the set K.

2.5.3 Reasoned basis for inference

The results from Figures 2-4a and 2-4b, which demonstrate the variety of unique allocations that can result under the CA-RO approach, indicate that CA-RO provides a sufficient degree of randomization to be used as a reasoned basis for inference. While the probability distribution of these allocations does not appear to be uniform (see Figure 2-4b), the fact that diverse allocations arise motivates us to conduct randomization-inspired tests for statistical significance such that the power of the CA-RO method can be estimated under various scenarios in Section 2.4.

2.6 Conclusions

In this paper, we introduced a covariate-adaptive optimization algorithm for the problem of online allocation of subjects in randomized controlled trials. Our method leverages robust mixed-integer quadratic optimization to improve upon state-of-the-art covariate-adaptive randomization methods. We demonstrated many desirable properties of the new CA-RO approach, including computational tractability, smaller between-group covariate imbalance as compared with randomization-based methods, and a low potential for common experimental biases. In all tested scenarios, the CA-RO method performed competitively with CA-RAND approaches, and sometimes significantly outperformed these methods, as measured by statistical power. We presented a setting with a nonlinear covariate-response model for which the CA-RO method achieved a desired level of statistical power at a sample size 25-50% smaller than the best CA-RAND method. Thus, the proposed CA-RO algorithm has significant potential to reduce both the cost and duration of clinical trials. The CA-RO algorithm can be used to make assignments to any arbitrary number of treatment groups and for any number of observed covariates. Finally, we constructed an extension of the CA-RO method for the setting in which it is possible to aggregate decision-making. We believe that the proposed CA-RO algorithm is an efficient alternative to covariate-adaptive randomization that can significantly strengthen experimental power in clinical trials and many other disciplines exploiting controlled experiments.

Chapter 3

Identifying Exceptional Responders in Randomized Trials via Mixed-Integer Optimization

This work was completed in collaboration with my co-authors Dimitris Bertsimas and Nikita Korolko.

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. In this chapter, 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.

3.1 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 the 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 benefit of the company's initial research and development investment by suggesting opportunities for re-marketing the drug 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 hazards model with treament-covariate interaction terms [Schemper, 1988]. Citing the disadvantage of needing to specify relevant interactions, Kehl and Ulm [2006] improve upon 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 (CART) [Breiman, 2001, Breiman et al., 1984]. This virtual twins method is shown to be effective in simulation studies, with positive predictive value ranging from 45-60%. Others [Su et al., 2009, Hardin et al., 2013] have presented CART-inspired recursive partitioning approaches to solve the subgroup identification or partioining problem. Su et al. [2009] show their method is effective in identifying subgroups with significant positive or negative treatment effect 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 was exceptionally strong or exceptionally weak and which can be defined by a small pre-specified 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 without the need for recursion or iteration.

In Section 3.2, 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 3.3, 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 3.4, we apply the MIO approach to datasets from two real randomized controlled trials. In one example, the method identifies a subset with a statistically significant exceptional response. In the other example, the method finds no subset that has a statistically significant response. Finally, in Section 3.5, we share some concluding remarks.

3.2 Identifying Interpretable Optimal Subgroups

Let us consider a randomized controlled trial in which subjects, indexed by $i = 1, \ldots, n$, have received treatment assignments T_i to one of two treatment conditions: treatment ($T_i = 1$) or control ($T_i = 0$). We define the sets $\mathcal{T}_t := \{i \,|\, T_i = t\}$, $t = 0, 1$. For each subject, we observe a covariate vector $\mathbf{x}_i \in \mathbb{R}^S$, where S is the number of observed covariates. We also observe treatment responses v_i, $i = 1, \ldots, n$.

Our objective is to find the subset $\mathcal{I}^* \subseteq \{1, \ldots, n\}$ with maximum (or minimum) average treatment value (ATE), where

$$\text{ATE}(\mathcal{I}^*) := \frac{1}{|\mathcal{T}_1 \cap \mathcal{I}^*|} \sum_{i \in \mathcal{T}_1 \cap \mathcal{I}^*} v_i \; - \; \frac{1}{|\mathcal{T}_0 \cap \mathcal{I}^*|} \sum_{i \in \mathcal{T}_0 \cap \mathcal{I}^*} v_i.$$

For the remainder of the paper, we focus on the maximization problem and we assume that larger values of the response v_i are preferable.

To make the definition of subset $\mathcal{I}^*$ interpretable, for each covariate $s = 1, \ldots, S$, we define K_s hyperplanes parallel to the coordinate axes that could be chosen as

boundaries for the box containing the subset. 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. We limit the number of covariate dimensions along which we can restrict the subset to some number $S^0 \leq S$. We also define the quantities $\underline{N}$ and $\overline{N}$, which constrain the cardinality of $\mathcal{I}^*$, such that $\underline{N} \leq |\mathcal{T}_t \cap \mathcal{I}^*| \leq \overline{N}$, $t = 0, 1$.

3.2.1 Mixed-integer optimization approach

In modeling the problem as an optimization problem, the key decisions are to define the boundaries of the subset. 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 bound for dimension s, and 0, otherwise. Similarly, let $\mathbf{U} := \{U_{sk} \,|\, s = 1, \ldots, S, k = 1, \ldots, K_s\}$ be a set of binary decision variables that take value 1, if γ_{sk} is chosen as the upper bound for dimension s, and 0, otherwise. Taken together, these binary decision variables uniquely define a box in the space $\mathbb{R}^S$, which will provide an interpretable subset.

We define auxiliary binary decision variables $z_i := \mathbb{I}\{i \in \mathcal{I}^*\}$, $i = 1, \ldots, n$. To enforce the limit on splitting dimensions, we define auxiliary binary indicator variables q_s, which indicate whether or not covariate dimension s is used to restrict the subset $\mathcal{I}^*$. Both vectors $\mathbf{z}$ and $\mathbf{q}$ are fully determined by the primary decision vectors $\mathbf{L}$ and $\mathbf{U}$, according to constraints eqs. (3.1b) to (3.1d) and eqs. (3.1g) to (3.1i), respectively.

We next describe the following fractional mixed-integer optimization (MIO) to identify the subset with highest ATE:

$$\max_{\mathbf{z},\mathbf{q},\mathbf{L},\mathbf{U}} \quad \frac{\sum_{i\in\mathcal{T}_1} v_i z_i}{\sum_{i\in\mathcal{T}_1} z_i} - \frac{\sum_{i\in\mathcal{T}_0} v_i z_i}{\sum_{i\in\mathcal{T}_0} z_i} \tag{3.1a}$$

$$\text{s.t.} \quad z_i + \sum_{s=1}^{S} \left[\sum_{k\,:\,\gamma_{sk} > x_{is}} L_{sk} + \sum_{k\,:\,\gamma_{sk} < x_{is}} U_{sk} \right] \geq 1, \qquad \forall i = 1, \ldots, n, \tag{3.1b}$$

$$z_i + L_{sk} \leq 1, \qquad \forall s = 1, \ldots, S,\ k = 1, \ldots, K_s, \quad i\ :\ x_{is} < \gamma_{sk}, \tag{3.1c}$$

$$z_i + U_{sk} \leq 1, \qquad \forall s = 1, \ldots, S,\ k = 1, \ldots, K_s, \quad i\ :\ x_{is} > \gamma_{sk}, \tag{3.1d}$$

$$\sum_{k=1}^{K_s} L_{sk} = 1, \qquad \forall s = 1, \ldots, S, \tag{3.1e}$$

$$\sum_{k=1}^{K_s} U_{sk} = 1, \qquad \forall s = 1, \ldots, S, \tag{3.1f}$$

$$q_s + L_{s1} \geq 1, \qquad \forall s = 1, \ldots, S, \tag{3.1g}$$

$$q_s + U_{sK_s} \geq 1, \qquad \forall s = 1, \ldots, S, \tag{3.1h}$$

$$q_s + L_{s1} + U_{sK_s} \leq 2, \qquad \forall s = 1, \ldots, S, \tag{3.1i}$$

$$\sum_{s=1}^{S} q_s \leq S_0, \tag{3.1j}$$

$$\underline{N} \leq \sum_{i \in \mathcal{T}_t} z_i \leq \overline{N}, \qquad \forall t = 0, 1, \tag{3.1k}$$

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

The constraints in formulation (3.1) deserve further discussion. First, for any given subject i, if the following condition is met,

$$\sum_{s=1}^{S} \left[\sum_{k\,:\,\gamma_{sk} > x_{is}} L_{sk} + \sum_{k\,:\,\gamma_{sk} < x_{is}} U_{sk} \right] = 0, \tag{3.2}$$

then z_i must be equal to 1, by constraints (3.1b). We observe that condition (3.2) 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 (3.1c) and (3.1d) 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 (3.1c) 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 (3.1d) we have $z_i = 0$ for all subjects i for which $x_{is} > \gamma_{sk}$.

Constraints (3.1e) and (3.1f) indicate that only one lower and upper bound, respectively, can be chosen for each covariate dimension. Constraints (3.1g), (3.1h), and (3.1i) 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 or not dimension s is used to restrict the space of $\mathcal{I}^*$. This relationship allows us to require that at most S_0 covariate dimensions are used to restrict the subset by adding constraint (3.1j). Finally, constraints (3.1k) ensure that the cardinality of $\mathcal{I}^*$ 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.

3.2.2 Tractable transformation of fractional objective

Because formulation (3.1) has a fractional objective function (3.1a), the problem cannot be solved using off-the-shelf commercial solvers. We can transform the objective from fractional to non-fractional by considering the expressions:

$$\Theta_t := \frac{1}{\sum_{i \in \mathcal{T}_t} z_i}, \; t = 0, 1. \tag{3.3}$$

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 bi-

nary 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 (3.1a) as a non-fractional, 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 eqs. (3.4b) to (3.4d): 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 (3.3), 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 (3.1) and can be solved using commercial optimization solvers:

$$\max_{\mathbf{z},\mathbf{q},\mathbf{L},\mathbf{U},\boldsymbol{\zeta},\boldsymbol{\theta}} \quad \sum_{i\in\mathcal{T}_1}\sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij} - \sum_{i\in\mathcal{T}_0}\sum_{j=\underline{N}}^{\overline{N}} \frac{1}{j} v_i \zeta_{ij} \tag{3.4a}$$

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

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

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

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

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

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

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

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

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

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

$$\zeta_{ij} \leq \theta_j^{(T_i)}, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{3.4b}$$

$$\zeta_{ij} \leq z_i, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{3.4c}$$

$$\zeta_{ij} \geq \theta_j^{(T_i)} + z_i - 1, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N}, \tag{3.4d}$$

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

$$\sum_{j=\underline{N}}^{\overline{N}} \theta_j^{(t)} = 1, \qquad \forall t = 0, 1,$$

$$0 \leq \zeta_{ij} \leq 1, \qquad \forall i = 1, \ldots, n,\ j = \underline{N}, \ldots, \overline{N},$$

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

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.

3.2.3 Statistical significance of optimal subset

Assuming $\boldsymbol{\gamma}$ is defined in a sensible manner, solving formulation (3.4) using an off-the-shelf commercial optimization solver, such as Gurobi [Gurobi Optimization, Inc., 2016], will yield an optimal or near-optimal feasible subset in all practical instances. Yet, 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 Fisher's sharp null hypothesis that there is no difference in treatment response between the treatment groups [Fisher, 1935].[1] We introduce some robustness to outliers by considering the trimmed mean, i.e., 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. 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 3.3, we use a non-parametric 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 and false positive rates.

If the sample size n is sufficiently large, 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

[1] In settings where the overall study sample had a statistically significant non-zero 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.

optimal solution, and evaluate the out-of-sample ATE and its significance within the test subset.

3.2.4 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 small impurity between Z_1 and Z_2, such that:

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

for some pre-defined impurity parameter ρ.

In order to find Z_2 that satisfies this impurity constraint, one can add the following constraint to formulation (3.4):

$$\sum_{i \in Z_1} z_i \leq \rho \cdot \left(\big|Z_1\big| + \sum_{i \notin Z_1} z_i \right), \tag{3.5}$$

since $|Z_1 \cup Z_2| = |Z_1| + |Z_2 \backslash Z_1|$. Then re-solving the MIO with the added constraint (3.5) yields the optimal Z_2. More generally, for any $m = 2, \ldots, M$, one can add pairwise impurity constraints:

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

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].

3.2.5 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 (3.4).

Like 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 in order to ensure heuristic solutions are feasible for formulation (3.4). 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 (3.4). 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 3.3, 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.

3.3 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 *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

$$v_i = 2 + \delta_0 \cdot y_i \cdot \mathbb{I}\{x_i^{(1)} \leq 0.5\} \cdot \mathbb{I}\{x_i^{(2)} \leq 0.5\} + \varepsilon_i, \tag{3.6}$$

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 (3.4) 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, due to noise in the response model (3.6), the subset identified did not always precisely match the known underlying best subset $\mathcal{I}_0^* := \{i \,|\, x_i^{(1)} \leq 0.5,\, x_i^{(2)} \leq 0.5\}$ (Figure 3-1). In order 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 3.1, for subsets that were found to be statistically significant. According to Table 3.1, the average accuracy, or percent 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 $\mathcal{I}_0^*$.

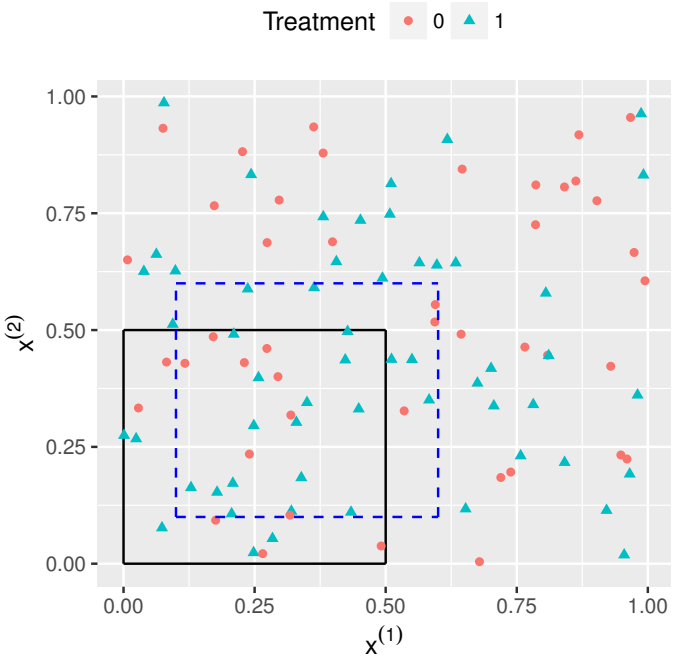


Figure 3-1: An example of a data sample projected onto the $(x^{(1)}, x^{(2)})$ space.

Shape indicates treatment group. Black solid line delineates boundary of known best subset $\mathcal{I}_0^*$. Blue dashed line delineates boundary of found optimal subset.

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

Table 3.1: Average percent of subjects in found subset versus best known subset $\mathcal{I}_0^*$ over 250 simulations in base case, when 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%

found subset ($z_i = 1$) were also in the known best subset ($i \in \mathcal{I}_0^*$), i.e., the positive predictive value (PPV) was greater than 50%. We use the term threshold-based true positive rate (TPR) to refer to the percent 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 ATE, 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 the distribution of the noise term ε_i (Table 3.2). 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 false positive rate (FPR) was stable and low in all tested instances, although it grew with respect to S likely due to 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 3.3 for experiments with $S = 10$ and $\sum_{s=1}^{S} K_s = 110$.

Table 3.2: Sensitivity analyses of true positive rate (TPR) and false positive rate (FPR) of found subsets with varying sample size n, effect size δ_0, and distributions of noise ε_i.

Parameter	**Value**	**Percent subsets significant**	**TPR based on threshold**	**TPR weighted by PPV**	**FPR**
n	100*	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*	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*	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*	76.8%	68.0%	60.7%	–
	4	98.8%	96.0%	87.9%	–
ε_i	$N(0,1)$*	76.8%	68.0%	60.7%	12.4%
distribution	$U[-\sqrt{3},\sqrt{3}]$	68.0%	58.8%	52.1%	12.8%

* Base case parameters used: $n = 100$, $S = 4$, $S_0 = 2$, $\delta_0 = 2$, $\varepsilon_i \sim N(0,1)$ *i.i.d.*

Table 3.3: Average computational time to achieve provable optimality by sample size n.

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

3.3.1 Multiple subsets

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

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

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 3.4. 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.

Table 3.4: True positive rate (TPR) with respect to finding each of two known underlying subsets.

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

3.4 Real Case Studies

In this section, we present two examples in which we apply the optimization approach to real datasets. In Example 1, the method identifies a subset with statistically significant positive average treatment effect despite an overall non-significant negative treatment effect in the study sample. In Example 2, 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.

3.4.1 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 which 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, while 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 in the placebo group, but the effect was non-significant using the bootstrap hypothesis testing approach discussed in Section 3.3 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.

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

3.4.2 Example 2: 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 of 312 PBC patients, as described in Fleming and Harrington [1991].[3] 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 1996.9 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 sharp null hypothesis using the bootstrap hypothesis test with significance level $\alpha = 0.01$.

We applied our optimal subset approach to determine if 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 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 $\alpha = 0.01$.

[3]The Fleming and Harrington [1991] data are available online at https://www.umass.edu/statdata/statdata/data/.

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 conducted additional testing to examine whether we should expect to find a significant subset in other samples from the same population. For 20 different random seeds, we randomly split the 312 subjects into a training set of 200 subjects and a testing set of 112. For each random splitting, we applied our method to find the best subset on the training set, and then evaluated the ATE for that same subset within the testing set. There was only 1 of 20 random splittings for which the out-of-sample ATE in the testing set was positive and statistically significant. Because there was no stable positive result across random splittings, we determined that there is no subset of the study population for which D-penicillamine was effective in improving survival time.

3.5 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 non-randomized data from observational studies, or in trials where more than two treatment conditions are administered.

Chapter 4

Personalized Diabetes Management Using Electronic Medical Records

This work appeared in Diabetes Care, with co-authors Dimitris Bertsimas, Nathan Kallus, and Ying Daisy Zhuo [Bertsimas et al., 2017].

Current clinical guidelines for managing type 2 diabetes do not differentiate based on patient-specific factors. In this chapter, we present a data-driven algorithm for personalized diabetes management that improves health outcomes relative to the standard of care.

We modeled outcomes under 13 pharmacological therapies based on electronic medical records from 1999 to 2014 for 10,806 patients with type 2 diabetes from Boston Medical Center. For each patient visit, we analyzed the range of outcomes under alternative care using a k-nearest neighbor approach. The neighbors were chosen to maximize similarity on individual patient characteristics and medical history that were most predictive of health outcomes. The recommendation algorithm prescribes the regimen with best predicted outcome if the expected improvement from switching regimens exceeds a threshold. We evaluated the effect of recommendations on matched patient outcomes from unseen data.

Among the 48,140 patient visits in the test set, the algorithm's recommendation mirrored the observed standard of care in 68.2% of visits. For patient visits in which

the algorithmic recommendation differed from the standard of care, the mean post-treatment glycated hemoglobin A1c (HbA1c) under the algorithm was lower than standard of care by $0.44 \pm 0.03\%$ ($p \ll 0.001$), from 8.37% under the standard of care to 7.93% under our algorithm.

A personalized approach to diabetes management yielded substantial improvements in HbA1c outcomes relative to the standard of care. Our prototyped dashboard visualizing the recommendation algorithm can be used by providers to inform diabetes care and improve outcomes.

4.1 Introduction

Diabetes is a chronic condition affecting almost 10% of the US population [National Center for Chronic Disease Prevention and Health Promotion, 2014]. Individuals with diabetes experience abnormally high blood glucose levels, which can lead to severe complications such as heart disease, stroke, and kidney failure. The most common form of diabetes is type 2 diabetes, which constitutes 90-95% of all diabetes cases in the US [Centers for Disease Control and Prevention, 2015]. The disease is typically managed through healthy eating, physical activity, oral medication, and/or insulin injections. While there are evidence-based clinical guidelines for glycemic control [Rodbard et al., 2009], how to choose among pharmacological therapies to maximize effectiveness for a given patient is not well understood. There has been growing interest in using clinical evidence to understand the effects of treatments in different type 2 diabetes populations. In a joint statement from 2012, the American Diabetes Association and the European Association for the Study of Diabetes highlighted the need for a patient-centered approach to diabetes management [Inzucchi et al., 2012]. The need for an individualized approach is especially pressing given the variety of disease symptoms, comorbid conditions, pharmacological treatments, individual treatment histories, and other individual characteristics that may inform treatment [Subramanian and Hirsch, 2014].

Evidence suggests that the response to blood glucose regulation agents can differ

among population subgroups. A post-hoc secondary analysis found that African-American pre-diabetic adults responded better to metformin than Caucasian pre-diabetic adults [Zhang and Zhang, 2015]. Another study recommended less aggressive treatments for older patients, as they were more likely to experience severe consequences from hypoglycemia [Ismail-Beigi et al., 2011]. These studies each provide valuable insights with respect to a single subgroup or treatment, but do not offer a decision rule for the general population that providers can easily apply in practice.

Tailoring glycemic management for specific subpopulations can be critical. Among patients with chronic kidney disease, contraindication to metformin needs to be taken into consideration when prescribing medication [Lipska et al., 2011]. Separate glycated hemoglobin (HbA1c) goals may be needed for subgroups or individuals differentiated by age, comorbidities, and other clinical characteristics [Subramanian and Hirsch, 2014]. A personalized treatment recommendation using a quantitative approach could readily incorporate different glycemic targets and contraindications, and thus allow for more systematic management of subgroups.

We provide an algorithm that generates a personalized type 2 diabetes treatment recommendation for any given patient based on evidence from historical outcomes of similar patients drawn from an electronic medical records (EMR) database. EMR analysis allows for pinpoint comparisons of effectiveness because of the abundance of clinical evidence from multiple treatment options administered to a diverse population over long-term patient clinical histories. EMR data combines the large sample sizes found in some insurance claims databases with the depth of longitudinal clinical evidence typically found in clinical trials. One caveat is that EMR data are not controlled via randomization.

Our methodological approach applies machine learning techniques and causal inference to make personalized recommendations based on comparative effectiveness among subpopulations in the EMR database. Machine learning techniques have been increasingly adopted in health care, along with many other fields [Jordan and Mitchell, 2015, Bertsimas and Kallus, 2014, Bertsimas et al., 2016]. Our novel approach leverages the power of analytics and abundant data in the EMR system to improve quality

of care.

The recommendations are personalized by patient characteristics, including age, sex, race, BMI, treatment history, and diabetes progression. We evaluate the effectiveness of the personalized treatment recommendations against the current standard of care by estimating patients' counterfactual outcomes from historical outcomes of similar patients in the EMR database. We develop a prototype clinical support dashboard that provides evidence for the algorithm's recommendations and could guide providers in caring for type 2 diabetes patients in a personalized manner.

4.2 Research Design and Methods

4.2.1 Analytic overview

We modeled outcomes for patients with type 2 diabetes based on EMR data. We divided each patient's medical history into distinct lines of therapy, each characterized by a particular drug monotherapy or combination therapy. Within each line of therapy, we considered patient visits occurring every 100 days. At each visit, the provider decides whether to proceed with the patient's current line of therapy or to recommend an alternative regimen. We developed a non-parametric prescriptive algorithm that provides personalized treatment recommendations. For each patient visit, we used k-nearest neighbor (kNN) regression [Cover and Hart, 1967] to predict the potential HbA1c outcome under each treatment alternative. The nearest neighbors were chosen to control for confounding that may be present in non-randomized data [Rosenbaum and Rubin, 1983] and to maximize similarity on the patient characteristics that were most predictive of outcomes. The algorithm then prescribed the regimen with best predicted outcome, provided the predicted improvement relative to the patient's current regimen exceeded a confidence threshold. The outcome metric was the average HbA1c measurement 75 to 200 days after the visit date. The effect of the prescriptive algorithm was evaluated by comparing the expected HbA1c outcome under our recommended therapy to the observed outcome under the standard of care

(ground-truth) therapy, according to a commonly used matching approach [Imbens and Rubin, 2015]. We conducted additional simulations to ensure that the results were robust to training models on different datasets and using alternative predictive modeling techniques.

4.2.2 Data

Through a partnership with Boston Medical Center (BMC), an academic medical center in Boston, Massachusetts, we obtained EMR for over 1.1 million patients from 1999 to 2014. In this dataset, 10,806 patients met all of the following inclusion criteria:

- Were present in the system for an observation period of at least 1 year;
- Received a prescription for at least one blood glucose regulation agent, including insulin, metformin, sulfonylureas, or one of the other blood glucose regulation agents listed below, and had at least one medical record 100 days prior to the date of this prescription;
- Had at least three recorded laboratory measurements of HbA1c; and,
- Did not have a recorded diagnosis of type 1 diabetes, as defined by the presence of International Classification of Diseases (ICD-9) diagnosis code 250.x1 or 250.x3 combined with the absence of any subsequent prescriptions for oral blood glucose regulation agents. (If the patient received oral blood glucose regulation agents subsequent to one of these diagnosis codes, we assumed the diagnosis record was an error.)

For each patient, we had access to demographic data, including date of birth, sex, and race/ethnicity, and to all BMC EMR data, including a history of drug prescriptions and measurements of height, weight, BMI, and HbA1c, as well as creatinine levels (Table 4.1). Neither the size of the population nor the proportion with good glycemic control changed substantially over the course of the study.

Table 4.1: Demographics, medical history, and treatment history of patients ($N = 10,806$).

Feature	**Mean (SD)**
*Age (years)**	59.7 (13.6)
% Male	42.4%
% Black	58.5%
% Hispanic	15.1%
% White	16.6%
*BMI (kg/m^2)**	33.1 (8.1)
*HbA1c (%)**	7.9 (1.8)
*% with good glycemic control, i.e. HbA1c $\leq$ 7.0%**	37.7%
Years since first treatment in EMR	3.52 (3.66)
Current prescription for metformin†	45.6%
Current prescription for insulin†	30.2%
Contraindicated to metformin‡	17.4%
Number of patients with first visit prior to 2007 (%)	6,175 (57.1%)

* Sample statistics are calculated across all patient visits. Individual patients with longer medical histories may be over-represented in the sample.
† Individuals may have a current prescription for both metformin and insulin.
‡ A patient was considered to be contraindicated to metformin when current serum level of creatinine was greater than 1.5 mg/dL.

4.2.3 Interpreting individual medical histories

We divided each patient's medical history into distinct lines of therapy, each characterized by a particular drug regimen (Figure 4-1). Within each line of therapy, we considered patient visits occurring every 100 days, corresponding to the life cycle of a red blood cell [Franco, 2012]. These patient visits provided the basis for our definition of patient outcomes.

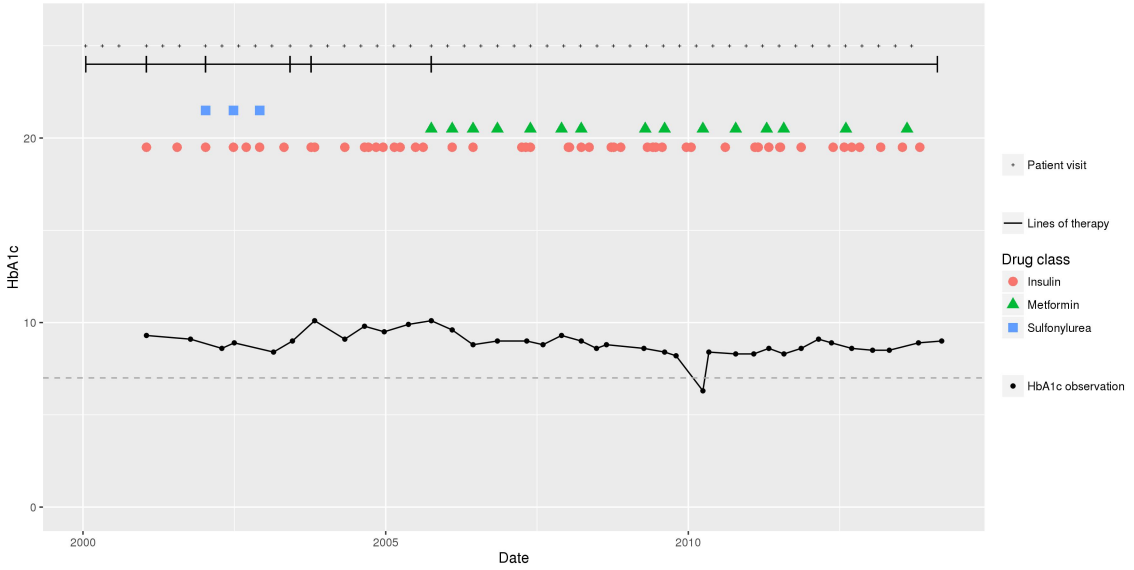


Figure 4-1: Treatment history for a sample patient.

The black points connected by black lines in the lower portion of the figure depict the HbA1c measurements over time for an example patient. The dotted horizontal black line shows a desired HbA1c threshold of 7%. The history of prescriptions this patient has received for each of three drugs, insulin, metformin, and sulfonylureas, is shown in red, green, and blue, respectively. The start and end of each line of treatment is indicated by a vertical tick mark. Finally, above the phase descriptions are small black dots indicating the date of each patient visit.

Lines of Therapy. We developed an algorithm to define precisely when each line of therapy ends and the next line begins according to when the combination of drugs prescribed to the patient changes in the EMR data. Each line of therapy was characterized by a unique drug regimen, defined to include all blood glucose regulation agents prescribed to the patient within the first 6 months after starting that line of therapy.

Regimens were defined as combinations of drugs from one or more drug classes. The drug classes we considered were metformin, insulin, and other blood glucose regulation agents; the other agents included sulfonylureas, thiazolidinediones, DPP-4 inhibitors, meglitinides, alpha-glucosidase inhibitors, GLP-1 agonists, and other antihyperglycemic agents. If a sufficient number of HbA1c observations existed during a period in which no drugs were prescribed, we defined the patient's line of therapy as "NoRx." We considered thirteen possible regimen types (Table 4.2). A combination of drug classes was included as a regimen type if it was observed in a sufficient number of patient visits.

Table 4.2: Pharmacological regimens.

Observed standard of care regimen (Abbreviation)	**Number of patient visits**
No regimen prescribed, new patient (NEWPT)	5,449
No regimen prescribed, existing patient (NORX)	2,137
Metformin monotherapy (MET0)	9,649
Insulin monotherapy (INS0)	7,539
Other blood glucose regulation agent monotherapy (OTHER0)	4,671
Metformin combined with one other non-insulin agent (MET1)	6,959
Metformin combined with insulin (METINS0)	3,977
Insulin combined with one non-metformin oral agent (INS1)	2,139
Combination of two non-metformin, non-insulin agents (OTHER1)	1,047
Metformin combined with two other oral agents (MET2)	1,749
Metformin combined with insulin and one other agent (METINS1)	2,005
Insulin combined with two non-metformin agents (INS2)	249
All other multi-drug (3+) combinations (MULTI)	570
Total	48,140

Patient Visits. Within each line of therapy, we considered patient visits occurring every 100 days, beginning with the visit at which that regimen was initiated and continuing until no later than 80 days prior to the start of the subsequent regimen. There were 48,140 unique patient visits in our dataset (Table 4.2). At each visit, we defined a set of visit-specific patient characteristics, including the current line of therapy (i.e. therapy given during the 100 days immediately preceding the current

visit) and recent HbA1c and BMI history. The outcome was measured as average HbA1c 75 to 200 days after the visit. This effect period was chosen to allow for a complete red blood cell life cycle to elapse before measuring the effect of a drug therapy.

We defined the standard of care for each visit as the drug regimen which was administered. For 16.3% of visits, the provider prescribed an adjustment to the current line of therapy; in the other 83.7%, the provider's prescription was to continue the current regimen.

4.2.4 Prescriptive algorithm

Our novel prescriptive algorithm considers a menu of available treatment options, including the patient's current treatment; uses k-nearest neighbor regression models to predict potential outcomes under each option; rejects any non-current treatment option with predicted outcome above a pre-specified HbA1c threshold; and chooses the remaining option with best predicted outcome. The menu of options for a given patient could be determined by the provider, accounting for contraindications and other preferences, such as not using intensive control for elderly patients or patients with a history of severe hypoglycemia.

For the purposes of this analysis, the menu of options for each patient was chosen relative to the intensity and composition of the patient's current treatment regimen. Specifically, the algorithm considered only regimens that represented an incremental addition or subtraction of a drug, or substitution of a drug of comparable intensity; metformin and insulin were considered to be of the lowest and highest intensity, respectively. Patients with serum creatinine levels, greater than 1.5 mg/dL [Lipska et al., 2011], a sign of kidney disease, were not offered metformin-based regimens. The menu options used in our analysis, differentiated by current treatment, are depicted in Figure 4-2; by definition, the algorithm never recommended metformin-based therapies for patients with the contraindication described above.

For each patient visit, the outcomes predicted by kNN under each treatment were compared. Our algorithm selected the treatment with the best predicted HbA1c

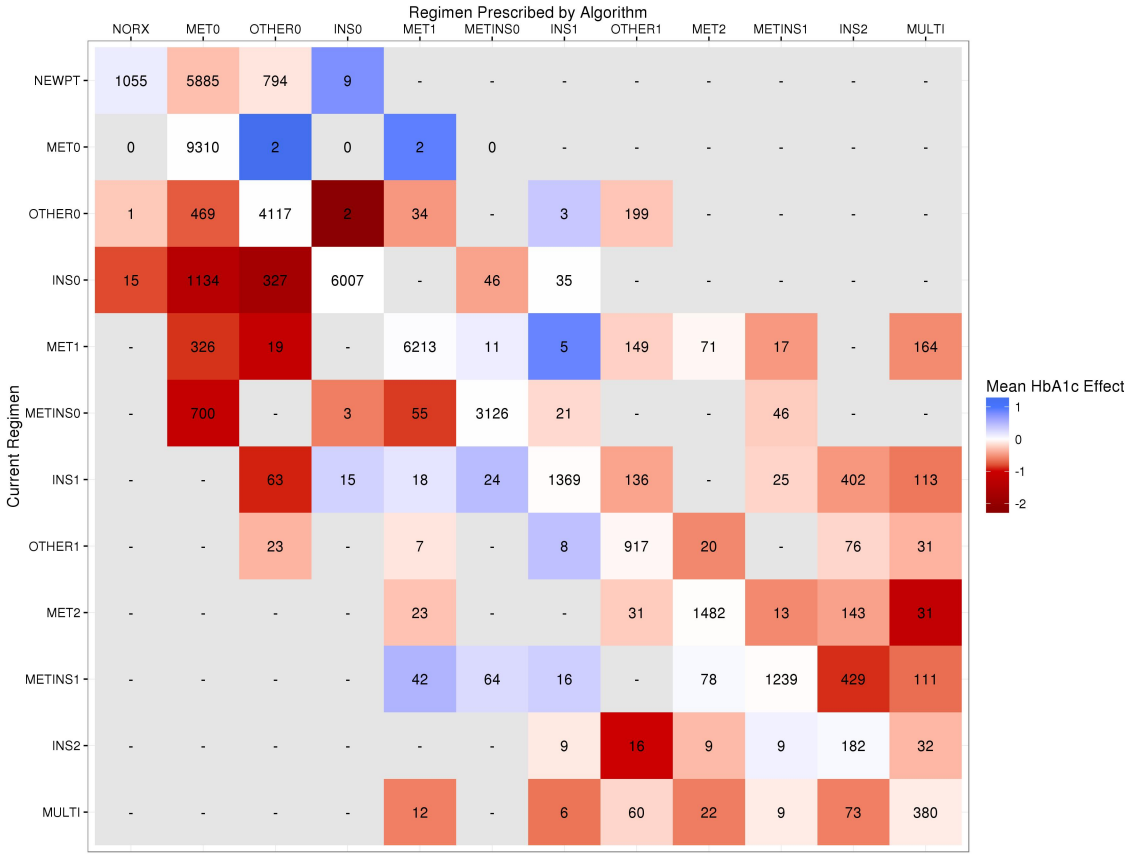


Figure 4-2: HbA1c benefit of prescriptive algorithm for patients switching regimens.

Each cell in the figure represents patients for whom the prescriptive algorithm recommended switching from the regimen on the vertical axis to the regimen on the horizontal axis. The color in each cell indicates the mean HbA1c benefit (%) of the prescriptive algorithm for patients in that cell, with red indicating benefits of the algorithm and blue indicating worsening relative to standard of care. Each cell is labeled with the number of patients who made that switch; cells labeled with a dash were not on the menu of options provided to patients currently on a given regimen. Patients with serum creatinine levels greater than 1.5 mg/dL were not considered for metformin-based regimens, and therefore are never assigned by the algorithm to columns with metformin-based regimens.

outcome subject to the condition that this best predicted outcome improve upon the predicted outcome under the patient's current treatment by at least some threshold δ. We chose the optimal threshold value of 0.8% by testing the algorithm on a single test set, using values of δ ranging from 0% to 1.5%. Increasing the threshold δ causes the algorithm to recommend switching for fewer patients, but the mean benefit among those who switch increases. Above a certain threshold, the recommendation fits to noise in the training data and does not provide better mean benefits in the testing set. The optimal threshold balances these concerns.

kNN regression is a non-parametric, instance-based algorithm that makes predictions by averaging the outcomes for the subset of observations most similar to the target as defined by some distance metric [Cover and Hart, 1967]. To predict potential outcomes under each regimen, we used a kNN regression based on a treatment-specific weighted Euclidean distance across normalized patient and visit-specific factors. The weights were derived by training a separate ordinary least squares linear regression model for each treatment regimen and using the magnitudes of the regression coefficients (Figure 4-3). This weighted distance improves upon classical kNN by selecting neighbors based on the factors most predictive of HbA1c outcome, rather than weighting all factors equally.

We considered factors from the following categories: demographic information, medical history, and treatment history. Specifically, the demographic factors used in the model were age, sex, and race. The medical history factors were days since first diabetes diagnosis; the patient's average serum creatinine level in the previous year; the patient's past two HbA1c and most recent BMI observations up to and including the current visit; the patient's average, median, 25th percentile, and 75th percentile HbA1c and BMI in the 1000-day period up to and including the current visit; and the patient's frequency of HbA1c measurements. The treatment history factors were the number of regimens the patient had tried; the number of visits since starting the current regimen; whether or not the patient had been previously prescribed metformin; and the patient's current regimen.

The prediction step of our algorithm is best illustrated through an example.

Feature	NEWPT	NORX	MET0	ORAL0	INS0	MET1	METINS0	INS1	ORAL1	MET2	METINS1	INS2	MULTI
Male sex	0.01	0.01	<0.01	0.01	0.01	<0.01	0.01	0.01	0.03	0.01	<0.01	<0.01	0.01
Black race/ethnicity	0.01	0.01	<0.01	<0.01	0.01	0.01	0.01	0.02	0.01	<0.01	0.03	0.10	<0.01
Hispanic race/ethnicity	0.01	0.01	<0.01	0.01	0.01	0.01	0.01	0.02	<0.01	<0.01	0.02	0.07	<0.01
White race/ethnicity	<0.01	0.01	<0.01	0.02	0.02	<0.01	0.01	0.01	<0.01	0.01	0.04	0.08	0.02
Age	<0.01	<0.01	0.01	0.01	0.03	0.04	0.02	0.01	0.03	0.01	0.02	0.04	<0.01
Time since first diabetes treatment	0.07	0.03	0.03	0.04	0.01	0.02	0.03	0.03	0.03	0.01	0.02	0.02	<0.01
Most recent HbA1c	0.21	0.31	0.22	0.24	0.25	0.17	0.31	0.15	0.11	0.09	0.14	0.05	0.07
Frequency of HbA1c measurements	0.02	0.01	0.01	0.02	0.02	0.02	0.03	0.02	0.01	<0.01	0.02	0.01	0.01
Second most recent HbA1c	0.01	0.07	0.01	0.01	0.04	0.01	0.02	0.01	0.02	0.04	<0.01	0.01	0.01
Median HbA1c (last 1000 days)	0.05	0.11	0.02	0.07	0.01	0.08	<0.01	0.04	0.08	0.09	0.04	0.08	0.10
75th percentile HbA1c (last 1000 days)	<0.01	0.12	0.10	0.09	0.07	0.07	0.07	0.02	0.13	0.13	0.06	0.02	0.05
25th percentile HbA1c (last 1000 days)	0.06	0.09	0.04	0.02	0.03	0.03	0.08	<0.01	0.18	0.14	0.07	0.01	0.06
Mean HbA1c (last 1000 days)	0.18	0.04	0.06	0.04	0.07	0.07	0.01	0.09	0.08	0.22	0.03	0.04	0.15
Most recent BMI	0.03	0.07	0.05	0.04	0.01	0.05	<0.01	0.01	0.03	0.03	0.01	0.02	0.03
Median BMI (last 1000 days)	0.04	<0.01	0.07	0.02	0.04	0.03	0.04	0.03	<0.01	0.04	0.04	0.04	0.08
75th percentile BMI (last 1000 days)	0.07	0.03	0.04	0.01	0.05	0.05	0.02	0.08	<0.01	0.01	0.13	0.03	0.01
25th percentile BMI (last 1000 days)	0.09	0.01	0.03	0.04	0.03	0.05	0.03	0.14	0.07	0.02	0.05	0.06	0.09
Mean BMI (last 1000 days)	0.14	0.03	0.11	0.07	0.01	0.08	0.01	0.17	0.11	0.05	0.21	0.14	0.20
Ever prescribed metformin previously	<0.01	<0.01	<0.01	0.02	0.01	<0.01	0.01	0.02	<0.01	<0.01	0.01	0.02	0.01
Serum creatinine level >= 1.5 mg/dL	<0.01	0.01	0.01	0.01	0.02	<0.01	0.01	0.01	<0.01	<0.01	0.02	0.01	0.01
Number of lines of therapy tried	<0.01	0.01	0.01	0.02	0.03	0.03	0.06	0.01	0.01	0.02	0.01	0.06	0.02
Number of visits since starting current line of therapy	<0.01	<0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.03	0.03	0.05	0.05
Patient currently prescribed metformin	<0.01	0.02	0.04	0.05	0.12	0.04	0.03	0.03	0.01	0.01	0.01	0.01	0.01
Patient currently prescribed insulin	<0.01	<0.01	0.11	0.01	0.05	0.06	0.05	0.01	0.05	0.03	0.01	0.04	<0.01
Number of unique drugs in current regimen	<0.01	0.01	0.01	0.10	0.04	0.07	0.11	0.05	<0.01	<0.01	0.01	0.01	0.01

Figure 4-3: Feature weights used to calculate similarity between patient visits.

Darker shading indicates larger values.

Suppose we would like to estimate a patient's potential outcome under metformin monotherapy. To identify the importance of each factor in predicting outcomes, we used patient visits in which metformin monotherapy was prescribed to train an ordinary least squares regression on normalized values of each patient factor listed above. The most predictive factors were: the patient's most recent HbA1c measurement (regression coefficient magnitude 0.22), whether the patient was currently prescribed insulin (0.11), the patient's mean BMI over the past 1000 days (0.11), and several other HbA1c and BMI measurements (coefficient magnitudes ranging from 0.03 to 0.10); see Figure 4-3 for full details. To estimate the patient's potential outcome, we used the coefficient magnitudes to weight the Euclidean distance between this patient visit and each patient visit in which metformin monotherapy was prescribed. Thus, for any choice of k, we could rank the k closest neighbors from this treatment group. This procedure was repeated for each therapy in the patient's menu of treatment options.

Intuitively, the number of neighbors k used to estimate post-treatment HbA1c levels should increase with the size of the dataset. For each treatment t, we found the value k_t^* that minimized the root-mean-square error of the kNN predictions on a subset of the data not used to evaluate the algorithm. We regressed k_t^* on $\sqrt{n_t}$, and thus derived the dependence function $k_t^* = 0.34 \cdot \sqrt{n_t}$, which was used to select k in the prescriptive algorithm.

To verify the accuracy of the kNN HbA1c predictions, we evaluated the R^2 metric. Positive values of R^2 suggest patient characteristics are predictive of future HbA1c. For comparison, we evaluated the predictive accuracy of LASSO regression [Tibshirani, 1996] and random forest [Breiman, 2001], two state-of-the-art machine-learning methods used widely due to their high prediction accuracy. We used the predictions from these models in two alternative prescriptive algorithms.

4.2.5 Model evaluation

To evaluate the performance of the kNN-based prescriptive model, we tested the algorithm's recommendations on a set of patient data that had not been used when

training the models.

Because counterfactual treatment effects are not observable, we used the weighted matching approach embedded in the kNN regression to impute potential outcomes, an approach commonly used for causal inference in observational studies when randomization is unavailable [Imbens and Rubin, 2015]. For each visit, we applied our prescriptive algorithm to recommend a therapy. If that recommendation matched the prescribed standard of care therapy, we observed the true effect from the therapy. Otherwise, the outcome was imputed by averaging the outcomes of the most similar patient visits at which the recommended therapy was administered; these similar visits were chosen from a test set not used for training, and the number of neighbors k_t^* was selected to fit the size of the test set. This estimated outcome was compared to the true outcome under standard of care at the given patient visit.

Our hypothesis was that the average predicted HbA1c outcome after applying our prescriptive algorithm would be less than that observed from administering standard of care, resulting in a net average improvement in outcomes.

4.2.6 Sensitivity analysis

To ensure the evaluation of our algorithm was not sensitive to the particular random split of the database into training and test data, we evaluated the effectiveness of our algorithm (with fixed threshold $\delta = 0.8$) under additional random splittings of the data.

4.2.7 Software

All analyses were performed in R 3.3.0 [R Core Team, 2016].

4.3 Results

The R^2 of the kNN predictions on unseen data ranged from 0.20 to 0.54 depending on the regimen (Table 4.3). The strongest models were for insulin monotherapy,

metformin monotherapy, metformin plus insulin, and multi-drug (3+) therapies. The R^2 values from the LASSO and random forest models ranged from 0.24 to 0.53. The predictive power was similar across the three methods.

Table 4.3: Out-of-sample R^2 under various predictive methods.

Regimen	**kNN**	**LASSO**	**Random Forest**
NEWPT	0.38	0.33	0.41
MET0	0.46	0.42	0.48
INS0	0.54	0.53	0.53
OTHER0	0.40	0.39	0.40
MET1	0.42	0.39	0.42
METINS0	0.46	0.46	0.47
INS1	0.44	0.43	0.43
OTHER1	0.34	0.35	0.35
MET2	0.32	0.32	0.33
METINS1	0.41	0.42	0.45
INS2	0.20	0.31	0.24
MULTI	0.46	0.36	0.46

The performance of the prescriptive algorithm is summarized in Table 4.4. The mean HbA1c outcome after treatment was 0.14% lower under the prescriptive algorithm than under the standard of care treatment, with standard error (SE) 0.01% and significance level $p \ll 0.001$. Of the 48,140 patient visits in our dataset, the algorithm differed from the standard of care for 15,323 visits, 31.8% of all visits. For this subset of visits, the mean HbA1c outcome under the algorithm was lower by $0.44 \pm 0.03\%$ compared with standard of care, with $p \ll 0.001$, a reduction from 8.37% under the standard of care to 7.93% under our algorithm. The median outcome for these visits was 0.21% lower under the prescriptive algorithm compared with standard of care. For comparison, the median difference for all visits was zero because, for 68.2% of visits, there was no difference between the algorithm's recommendation and the standard of care.

In our analysis, the mean difference in HbA1c was more negative than the median due to a left-skewed distribution. Some patients received particularly large benefits from using the prescriptive algorithm, which had an outsize effect on the mean but

Table 4.4: Performance of prescriptive algorithms.

All patient visits ($N = 48,140$)			
	k**NN**	**LASSO**	**Random Forest**
Mean HbA1c benefit relative to standard of care, % (SE)	-0.14 (0.01)*	-0.13 (0.01)*	-0.07 (0.01)*

Visits for which algorithm's recommendation differed from observed standard of care			
	k**NN**	**LASSO**	**Random Forest**
Number of visits (%)	15,323 (31.8%)	12,684 (26.3%)	14,302 (29.7%)
Mean HbA1c benefit relative to standard of care, % (SE)	-0.44 (0.03)*	-0.45 (0.03)*	-0.26 (0.03)*

* $p \ll 0.001$.

did not affect the median.

Figure 4-2 depicts the number of patients for whom the prescriptive algorithm recommended switching from a given current line of therapy to a given new line of therapy, along with the mean reduction in HbA1c for patient visits in each category. Among trajectories with at least 300 patients, the largest benefit of the algorithm was achieved through personalized recommendations for 7,564 patients currently on insulin monotherapy to switch to monotherapy with metformin or another blood glucose regulation agent. However, for the vast majority of patients currently on insulin-based regimens, the algorithm recommends that those patients continue with that therapy. Among the 7,564 patient visits, those who were recommended to switch from insulin were on average younger (mean age 52.9 years versus 61.4 years) and had substantially higher average HbA1c (11.0% versus 8.0%).

The performance of the prescriptive algorithm in specific patient subgroups is summarized in Tables 4.5 and 4.6. The overall mean HbA1c outcome using the prescriptive algorithm was 0.14% lower than standard of care for both male and

female patients. The benefit of using the algorithm was 0.14% for black patients (29,120 visits), 0.09% for white patients (7,444 visits), 0.22% for Hispanic patients (6,732 visits), and 0.11% for all other patients (4,844 visits). The benefit of the algorithm was 0.20% for patients under the age of 60 and 0.08% for patients aged 60 or older. The benefit was 0.20% for patients with poor glycemic control, i.e. current HbA1c greater than 7.0% as compared with 0.05% for those with good glycemic control.

Table 4.5: Performance of algorithm in study subgroups; all patient visits.

Subgroup	**Number of visits***	**Mean HbA1c benefit relative to standard of care, % (SE)†**
Male	20,231	-0.14 (0.01)
Female	27,909	-0.14 (0.01)
Black	29,120	-0.14 (0.01)
White	7,444	-0.09 (0.01)
Hispanic	6,732	-0.22 (0.01)
Other	4,844	-0.11 (0.01)
Age <60	23,705	-0.20 (0.01)
Age 60+	24,435	-0.08 (0.00)
Good glycemic control (HbA1c ≤ 7)	18,156	-0.05 (0.01)
Poor glycemic control (HbA1c > 7)	29,984	-0.20 (0.01)

* $N = 48,140$
† $p \ll 0.001$ for all instances.

Our methodology motivates a provider dashboard that would report information on the demographics, medical history, and response to treatment for patients similar to an index patient. A prototype dashboard visualization for one sample patient visit is shown in Figure 4-4. The dashboard would include the patient's demographic and health information along with visualizations of the patient's treatment history and

Table 4.6: Performance of algorithm in study subgroups; visits for which algorithm's recommendation differed from standard of care.

Subgroup	**Number of visits**	**Percent of all visits in subgroup**	**Mean HbA1c benefit relative to standard of care, % (SE)***
Male	6,363	31.5%	-0.44 (0.02)
Female	8,960	32.1%	-0.44 (0.02)
Black	9,103	31.3%	-0.45 (0.02)
White	2,309	31.0%	-0.29 (0.03)
Hispanic	2,400	35.7%	-0.61 (0.03)
Other	4,844	31.2%	-0.34 (0.04)
Age <60	8,783	37.1%	-0.55 (0.02)
Age 60+	6,540	26.8%	-0.30 (0.02)
Good glycemic control (HbA1c ≤ 7)	4,438	24.4%	-0.20 (0.02)
Poor glycemic control (HbA1c > 7)	10,885	36.3%	-0.54 (0.02)

* $p \ll 0.001$ for all instances.

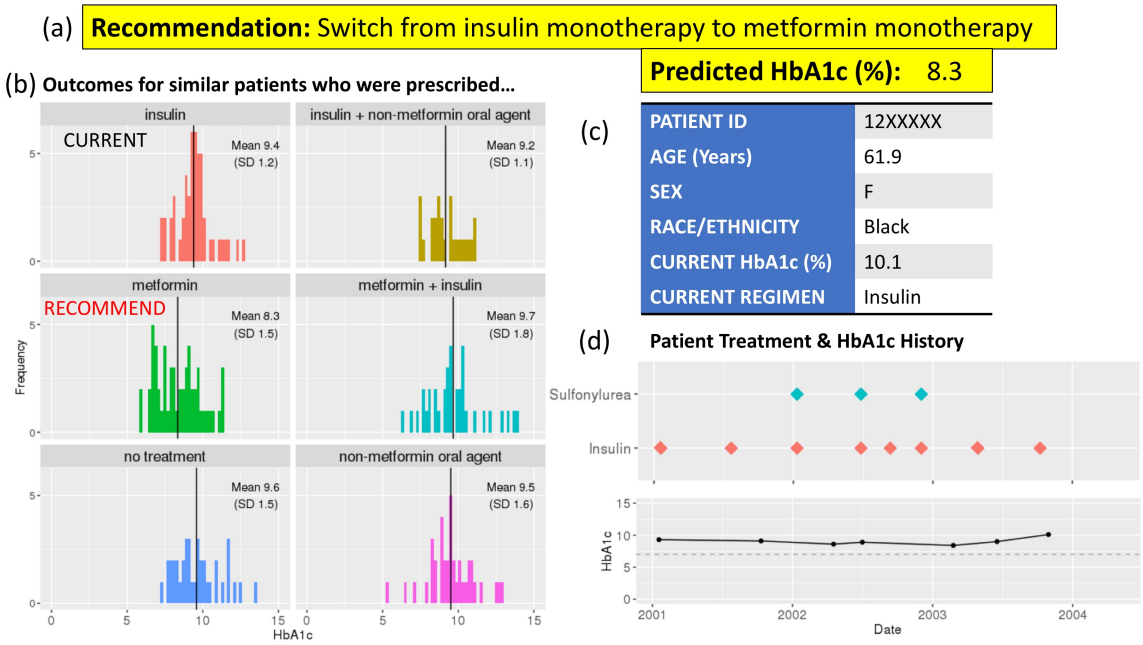


Figure 4-4: Visualization of prescriptive algorithm: Provider dashboard prototype.

This figure visualizes how the prescriptive algorithm can be used by providers for a single patient.

- Panel (a) displays the algorithm's treatment recommendation along with the predicted post-treatment HbA1c under that treatment.
- Panel (b) depicts the mean, standard deviation, and full distribution of post-treatment HbA1c outcomes for the k_t^* most similar patient visits in the data set, for each of the six regimens on this patient's menu of options. In each sub-panel, the post-treatment HbA1c level is on the horizontal axis and the number of visits is on the vertical axis.
- The table in panel (c) presents basic information about the patient's demographic and medical history.
- Panel (d) depicts the history of diabetes progression and treatment for the patient, with date along the horizontal axis. The vertical axis of the upper sub-panel indicates various drug classes; the vertical axis of the lower sub-panel depicts HbA1c %.

HbA1c progression. In addition, the dashboard would display the mean, standard deviation, and full distribution of HbA1c outcomes among the k_t^* nearest neighbors who received each treatment in the menu of options. Based on this evidence, the dashboard would display a treatment recommendation. The provider would have the ability to override this recommendation given any special management needs of the patient. For instance, if the patient is elderly and the distribution of HbA1c outcomes indicates that the recommended therapy has an elevated risk of hypoglycemia, the provider may opt for an alternative treatment.

The overall mean HbA1c outcome using the LASSO-based prescriptive algorithm was lower by $0.13 \pm 0.01\%$ ($p \ll 0.001$) compared with the mean standard of care outcome. The benefit from using the random-forest-based prescriptive algorithm relative to standard of care was $0.07 \pm 0.01\%$ ($p \ll 0.001$).

In the sensitivity analyses, under three alternate random splittings of the dataset, the overall mean benefit of using the prescriptive algorithm compared with standard of care ranged from 0.11% to 0.15% ($p \ll 0.001$ in all instances).

4.4 Discussion

To our knowledge, we present the first prescriptive method for personalized type 2 diabetes care. Using historical data from a large EMR database, this novel prescriptive method resulted in an average HbA1c benefit of 0.44% at each doctor's visit for which the algorithm's recommendation differed from standard of care.

Our method incorporates patient-specific demographic and medical history data to determine the best course of treatment. Compared to other machine learning methods considered, the kNN prescriptive approach is highly interpretable and flexible in clinical applications. The novelty of our approach is in personalizing the decision-making process by incorporating patient-specific factors. This method can easily accommodate alternative disease management approaches within specific subpopulations, such as patients with chronic kidney disease and elderly patients. We believe this personalization is the primary driver of benefit relative to standard of care.

In practice, the algorithm can be integrated into existing EMR systems to dynamically suggest personalized treatment paths for each patient based on historical records. The algorithm ingests and analyzes EMR data and generates recommendations. An intuitive, interactive dashboard summarizes the evidence for the recommendation, including the expected distribution of outcomes under alternative treatments (Figure 4-4).

Due to the nature of retrospective data from existing EMR, this study has several limitations. Patients were not randomized into treatment groups. While our matching methodology controls for several confounding factors that could explain differences in treatment effects, we can only estimate counterfactual outcomes. EMR data do not include socio-economic factors or patient preferences that may be important in treatment decisions. Due to lack of sufficient data, GLP-1 agonists were not considered as a separate drug class. If more data were available, we could further differentiate regimen types beyond the thirteen we include in this analysis. In addition, the study population from BMC may not be representative of the US population as a whole.

With EMR medication order data alone, we cannot be certain whether a prescribed medication was filled or taken, and cannot know precisely when the medication was stopped. Although this data quality issue could hamper attempts to make drug efficacy comparisons, our analysis aims to address the question of which drugs to prescribe under real-worlds scenarios. We optimize for an outcome that takes into account unobserved factors such as non-adherence. For instance, if non-adherence is more prevalent among patients prescribed insulin than other regimens, this perspective may explain why, in our study population, the algorithm recommends insulin less often than it is prescribed in clinical practice.

Our method can be extended to be more flexible and comprehensive. Currently the prescriptive algorithm does not support individualized glycemic targets; we assume that a lower glycemic level is always preferred. The study currently optimizes only for a single health outcome; a more comprehensive algorithm would consider adverse event outcomes as well.

Despite these limitations, the study establishes strong evidence of the benefit of

individualizing diabetes care. The success of this data-driven approach invites further testing using datasets from other hospital and care settings. Testing the prescriptive algorithm in a clinical trial setting would provide even stronger evidence of clinical effectiveness. As large-scale genomic data becomes more widely available, the algorithm could readily incorporate such data to reach the full potential of personalized medicine in type 2 diabetes.

In this study, we developed a novel data-driven prescriptive algorithm for type 2 diabetes that improves significantly on the standard of care when tested on patient-level EMR data from a large medical center. Our work is a key step toward a fully patient-centered approach to diabetes management.

Chapter 5

Conclusion

Decision-making in healthcare presents a striking challenge and opportunity for the field of analytics and operations research. By applying and combining advances in the areas of optimization, statistics, and machine learning, with new and interesting sources of data, we have the potential to make substantive improvements in the approval of novel and effective treatments and the outcomes of the patients receiving them.

In this thesis, we have demonstrated some of the ways this impact can be realized. In Chapters 2 and 3, we showed how mixed-integer optimization can be used to transform the design and analysis of clinical trials, yielding the potential for new insights and added value. Our covariate-adaptive optimization algorithm for sequential clinical trials (Chapter 2) achieves a desired level of statistical power at equal or smaller sample sizes compared to state-of-the-art methods, without sacrificing computational tractability or protection from experimental bias. These practical gains were made possible by a choice of uncertainty set that allows us to find a closed-form solution to the inner robust formulation. Our optimization approach to identifying exceptional responders from randomized trials (Chapter 3) has the potential to add value in a number of practical applications - from pinpointing promising subpopulations in failed clinical trials, to detecting adverse events experienced by certain subgroups. We used simple transformations to model the problem as a tractable MIO formulation. In Chapter 4, we considered the problem of personalized medicine in the setting

of type 2 diabetes management. Our study was the first to use machine learning to make personalized treatment recommendations using electronic medical records. We showed the algorithm could improve outcomes for diabetes patients, while giving providers evidence for these recommendations via an intuitive visual dashboard.

Through this work, we have made valuable contributions in the area of data-driven decision-making in healthcare. Perhaps more importantly, we have demonstrated the potential for impact when analytics and operations research are applied to this important and challenging domain, where there is much more work to be done.

Bibliography

A. B. Antognini and M. Zagoraiou. The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika*, 98(3): 519–535, 2011.

A. C. Atkinson. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69(1):61–67, 1982.

C. Bandi and D. Bertsimas. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming*, 134(1):23–70, 2012.

A. Ben-Tal, A. Nemirovski, and C. Roos. Robust solutions of uncertain quadratic and conic-quadratic problems. *SIAM Journal on Optimization*, 13(2):535–560, 2002.

E. R. Berndt and I. M. Cockburn. Price indexes for clinical trial research: a feasibility study. Technical report, National Bureau of Economic Research, 2013.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, 2014.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific, Belmont, MA, 1997.

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

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

D. Bertsimas, A. K. O'Hair, and W. Pulleyblank. *The Analytics Edge*. Dynamic Ideas LLC, 2016.

D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, 40:210–217, 2017.

J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

N. Bhat, V. F. Farias, and C. C. Moallemi. Optimal A-B testing. *Working paper*, 2015.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC Press, 1984.

D. Byar and S. Green. The choice of treatment for cancer patients based on covariate information. *Bulletin du cancer*, 67(4):477–490, 1979.

Centers for Disease Control and Prevention. Basics about diabetes. `http://www.cdc.gov/diabetes/basics/diabetes.html`, 2015. Accessed: 2015-04-16.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*, volume 169. John Wiley & Sons, 1991. URL `https://www.umass.edu/statdata/statdata/data/pbc.txt`.

J. C. Foster, J. M. G. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

R. S. Franco. Measurement of red cell lifespan and aging. *Transfusion Medicine and Hemotherapy*, 39(5):302–307, 2012.

J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.

R. Greevy, B. Lu, J. H. Silber, and P. Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004.

Gurobi Optimization, Inc. Gurobi Optimizer Reference Manual, 2016. URL `http://www.gurobi.com`.

D. S. Hardin, R. D. Rohwer, B. H. Curtis, A. Zagar, L. Chen, K. S. Boye, H. H. Jiang, and I. A. Lipkovich. Understanding heterogeneity in response to antidiabetes treatment: a post hoc analysis using sides, a subgroup identification algorithm. *Journal of Diabetes Science and Technology*, 7(2):420–430, 2013.

M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51, 2014.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

S. E. Inzucchi, R. M. Bergenstal, J. B. Buse, M. Diamant, E. Ferrannini, M. Nauck, A. L. Peters, A. Tsapas, R. Wender, and D. R. Matthews. Management of hyperglycemia in type 2 diabetes: A patient-centered approach position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, 35(6):1364–1379, 2012.

F. Ismail-Beigi, E. Moghissi, M. Tiktin, I. B. Hirsch, S. E. Inzucchi, and S. Genuth. Individualizing glycemic targets in type 2 diabetes mellitus: Implications of recent clinical trials. *Annals of Internal Medicine*, 154(8):554–559, 2011.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

A. Kapelner and A. Krieger. Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70(2):378–388, 2014.

V. Kehl and K. Ulm. Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis*, 50(5):1338–1355, 2006.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

K. J. Lipska, C. J. Bailey, and S. E. Inzucchi. Use of metformin in the setting of mild-to-moderate renal insufficiency. *Diabetes Care*, 34(6):1431–1437, 2011.

M. Lubin and I. Dunning. Computing in operations research using Julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

C. Morris. A finite selection model for experimental design of the health insurance study. *Journal of Econometrics*, 11(1):43–61, 1979.

National Center for Chronic Disease Prevention and Health Promotion. National Diabetes Statistics Report. Technical report, Centers for Disease Control and Prevention, 2014.

S. J. Pocock and R. Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103–115, 1975.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016, 2016.

H. W. Rodbard, P. S. Jellinger, J. A. Davidson, D. Einhorn, A. J. Garber, G. Grunberger, Y. Handelsman, E. S. Horton, H. Lebovitz, P. Levy, et al. Statement by an American Association of Clinical Endocrinologists/American College of Endocrinology consensus panel on type 2 diabetes mellitus: An algorithm for glycemic control. *Endocrine Practice*, 15(6):540–559, 2009.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

W. F. Rosenberger and O. Sverdlov. Handling covariates in the design of clinical trials. *Statistical Science*, 23(3):404–419, 2008.

M. Schemper. Non-parametric analysis of treatment-covariate interaction in the presence of censoring. *Statistics in Medicine*, 7(12):1257–1266, 1988.

X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

S. Subramanian and I. B. Hirsch. Personalized diabetes management: Moving from algorithmic to individualized therapy. *Diabetes Spectrum*, 27(2):87–91, 2014.

D. R. Taves. Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, 15(5):443, 1974.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

U.S. Food and Drug Administration. *The Drug Development Process: Clinical Research*, 2015. http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm.

C. Zhang and R. Zhang. More effective glycaemic control by metformin in African Americans than in Whites in the prediabetic population. *Diabetes & Metabolism*, 41(2):173–175, 2015.