Covariate Balancing Methods for Randomized Controlled Trials Are Not Adversarially Robust

Hossein Babaei, Sina Alemohammad, and Richard G. Baraniuk, *Fellow, IEEE*

***Abstract*— The first step toward investigating the effectiveness of a treatment via a randomized trial is to split the population into control and treatment groups then compare the average response of the treatment group receiving the treatment to the control group receiving the placebo. To ensure that the difference between the two groups is caused only by the treatment, it is crucial that the control and the treatment groups have similar statistics. Indeed, the validity and reliability of a trial are determined by the similarity of two groups' statistics. Covariate balancing methods increase the similarity between the distributions of the two groups' covariates. However, often in practice, there are not enough samples to accurately estimate the groups' covariate distributions. In this article, we empirically show that covariate balancing with the standardized means difference (SMD) covariate balancing measure, as well as Pocock and Simon's sequential treatment assignment method, are susceptible to worst case treatment assignments. Worst case treatment assignments are those admitted by the covariate balance measure, but result in highest possible ATE estimation errors. We developed an adversarial attack to find adversarial treatment assignment for any given trial. Then, we provide an index to measure how close the given trial is to the worst case. To this end, we provide an optimization-based algorithm, namely adversarial treatment assignment in treatment effect trials (ATASTREET), to find the adversarial treatment assignments.**

***Index Terms*— Adversarial analysis, causal effect, clinical trials, covariate balancing, econometric, experimental design, policy evaluation, randomized controlled trials (RCTs), sequential treatment assignment, treatment effect.**

I. Introduction

THE standard method to measure the causal relationship between two variables is the average treatment effect (ATE) [1]. The term ATE refers to the average outcome change that a certain intervention (which is called treatment) can make in a population in contrast to not making the intervention.

Manuscript received 23 October 2021; revised 27 August 2022 and 28 January 2023; accepted 20 February 2023. This work was supported in part by the National Science Foundation (NSF) under Grant 1842378 and Grant 1937134; in part by the Division of Computing and Communication Foundation Grant CCF-1911094; in part by the Division of Information and Intelligent Systems Grant IIS-1838177 and Grant IIS-1730574; in part by the Office of Naval Research (ONR) under Grant N00014-18-12571 and Grant N00014-20-1-2534; in part by the Multidisciplinary University Research Initiatives Grant MURI N00014-20-1-2787; in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-18-1-0478; and in part by the Vannevar Bush Faculty Fellowship, ONR, under Grant N00014-18-1-2047. *(Corresponding author: Hossein Babaei.)*

The authors are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: hb26@rice.edu; sa86@rice.edu; richb@rice.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3266429.

Digital Object Identifier 10.1109/TNNLS.2023.3266429

Randomized controlled trials (RCTs) are the gold standard for conducting quantitative experimental science [2], [3], [4], [5], [6]. RCT experimental design consists of recruiting a study population and splitting the participants into two groups: treatment and control.[1] If the treatment is assigned randomly, the difference between the average outcomes of the two groups is an unbiased estimator of the ATE [7]. Since the trial is only conducted once, it is of high importance to reduce the ATE estimation variance.

Covariate balancing methods (CBMs) are methods to measure and induce more similarity in the statistics of the two groups. To ensure that the difference between the two groups is caused only by the treatment, it is crucial that the control and the treatment groups have similar statistics. The similarity of the statistics is commonly used to evaluate the validity and reliability of the conclusions based on the estimated ATE in an RCT.

In this article, we perform worst case analysis of CBMs. We define the worst case treatment assignments of a given CBM in an RCT as the treatment assignments that would be evaluated as sufficiently balanced by the given CBM, but would result in the highest possible ATE estimation error. We provide quantitative definition of *sufficiently balanced* later in this article.

In this work, we perform worst case analysis on two commonly used CBMs, the standardized means difference (SMD) for nonsequential treatment assignments and the sequential assignment method of Pocock and Simon's (P&S) [8]. In both cases, we develop a method that finds the worst case treatment assignments in a given RCT that we dub the Adversarial Treatment ASsignment in TREatment Effect Trials (ATASTREET). ATASTREET reduces the combinatorially large space of possible treatment assignments to efficiently find the worst case treatment assignment.

To find worst case treatment assignments, ATASTREET work as an oracle method with the access to both potential outcomes. As an illustrative example, we use the semi-synthetic IHDP1000 [9], [10], [11] dataset, which provides both potential outcomes for each participant. IHDP is widely accepted as the standard benchmark dataset in heterogeneous treatment effect estimation. Naturally, some would criticize IHDP and argue that it is not a good reflection of a real-world RCT [12]. Nevertheless, IHDP is still considered as the dataset

[1] In this article, we use terminology associated with medical clinical trials. However, any argument about medical clinical trials can be generalized to wider applications.

that could provide the strongest evidence in heterogeneous treatment effect estimation literature.

We empirically demonstrate the worst case vulnerability of the investigated CBMs. The worst case treatment assignment can get selected for the trial as a result of CBM, either unluckily or by intentional deviations from a deceitful researcher. Since it results in maximally balanced groups, it encourages the confidence in the MATE with worst case ATE estimation error. Since the trial is conducted only once, it is important to ensure that the selected treatment assignments are not close to worst case treatment assignments.

We define *CBM deviation index* ρ to identify whether these worst cases of CBM happened in any given RCT. This index provides a measurement on how close the selected treatment assignment is to the worst case treatment assignments. For any given RCT, we use counterfactual estimation methods to estimate the unobserved potential outcomes. Then, ATASTREET finds the worst case assignments. The ρ-index can be measured afterward to identify the unlucky or deceitful deviations in the trial.

To further emphasis the importance of worst case analysis and such sanity check, we develop an adversarial attack to any given RCT that used the mentioned CBMs, and empirically evaluate our introduced adversarial attack on the IHDP dataset. An adversary can exploit the adversarial vulnerability and use adversarial treatment assignments to maximize (or minimize) the measured ATE in the trial while having maximally balanced treatment groups.

We summarize our contributions as follows. First, we propose an optimization-based algorithm (ATASTREET) to find worst case treatment assignments of SMD and P&S method as CBMs. We then empirically demonstrate worst case vulnerability of the mentioned CBMs. Second, we provide an index to identify if a given trial is close to the worst case assignment, Third, we introduce an adversarial treatment assignment method using ATASTREET. Finally, we demonstrate the adversarial vulnerability of SMD and P&S method and discuss some of the possible solutions to reduce the adversarial vulnerability.

II. Background

In this section, we first cover some of the basic definitions about ATE and RCTs, then discuss some recognized challenges. consequently, we cover how variance reduction techniques and CBMs are discussed in the literature.

A. Background on Randomized Clinical Trails

The ATE is defined using the potential outcome framework [1]. For each individual i in the population, we call the potential outcomes of that individual being assigned to the treatment Y_i^1 or the control group Y_i^0. A set of covariates for each subject is also recorded as $\vec{x}^{\,i}$. The ATE is defined as the average of the differences of the potential outcomes for all the individuals over the population

$$\text{ATE} = \frac{1}{N}\sum_i \left(Y_i^1 - Y_i^0\right) \tag{1}$$

where N is the population size.

In a trial to measure the ATE of a certain treatment(intervention), a *treatment assignment* $\mathcal{A} : \{1, 2, \ldots, N\} \rightarrow \{0, 1\}^N$ divides the population to either the treatment group or the control group. For each individual, the $Y_i^{(\text{obs})}$ is the observed outcome based on the selected treatment assignment

$$Y_i^{(\text{obs})} = \begin{cases} Y_i^1, & \mathcal{A}(i) = 1 \\ Y_i^0, & \mathcal{A}(i) = 0. \end{cases} \tag{2}$$

The "fundamental problem of causal inference" [13] is that each individual subject in the population can only be assigned to either the treatment or the control group. Therefore, the outcome of an individual subject given the treatment and that of the same individual not given the treatment cannot be observed in the same trial. As a result, half of the required data for estimating the ATE is unobservable.

Can this fundamental problem be solved? Dawid et al. [14] argued that estimating the unobserved potential outcomes can result in erroneous or metaphysical conclusions that are not substantiated by the data. Thus, solutions for the "fundamental problem of causal inference" are dubious and cannot be supported by evidence in the experiment. Pearl et al. [15] and Shpitser and Pearl [16] argued against this paradigm by providing a framework that, given some structural information about the causal relationships in the system, identifies cases where the unobserved potential outcomes can be discerned by observations. Their arguments support the claim that the estimation of unobserved potential outcomes is a mathematical, not metaphysical, question. Some works first learn a causal graph over the variables with methods such as [17]; then study the ATE identifiability problem in the presence of unobserved variables. They argue that from the causal graph and observational data, some ATEs are nonidentifiable due to the unmeasured confounders, and additional assumptions are required [18], [19], [20].

In the random treatment assignment method [7], the trial is conducted using a randomly selected treatment assignment $\mathcal{A}$. The measured average treatment effect (MATE) is then defined as

$$\text{MATE}(\mathcal{A}) = \frac{1}{N_1}\sum_{\mathcal{A}(i)=1} Y_i^{(\text{obs})} - \frac{1}{N_0}\sum_{\mathcal{A}(i)=0} Y_i^{(\text{obs})} \tag{3}$$

where N_0 and N_1 are the number of individuals assigned to the control and treatment groups, respectively.

Given the population, Athey and Imbens [7] demonstrated that the introduced MATE is an unbiased estimator of the ATE. It means that the expected value of MATE over the random treatment assignment $\mathcal{A}$ is equal to the true value of ATE.

The ATE estimation error for any given treatment assignment $\mathcal{A}$ is the error in the MATE when $\mathcal{A}$ is used as the treatment assignment

$$\tilde{\epsilon}(\mathcal{A}) = \text{MATE(A)} - \text{ATE}. \tag{4}$$

Generally, the goal is to reduce $|\tilde{\epsilon}(\mathcal{A})|$ as much as possible.

B. Challenges in Randomized Clinical Trials

The estimated ATE has some variance due to randomly selected treatment assignment. The mentioned *ATE variance*

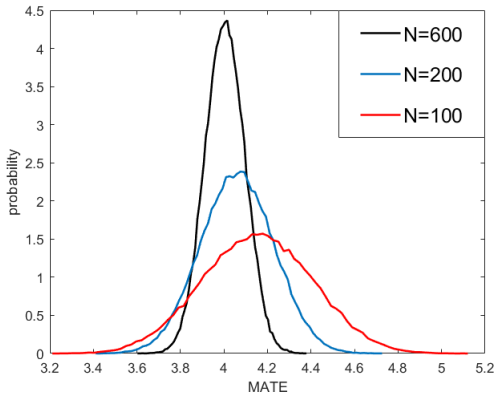

Fig. 1. Empirical probability distribution of the MATE in IHDP1000 for varying population sizes N. As the population size grows, the variance σ^2_{ATE} decreases.

is the variance of the ATE estimation when $\mathcal{A}$ is selected uniformly random

$$\sigma^2_{\text{ATE}} = \mathbb{E}_{\mathcal{A}\in\{0,1\}^N}\big[\tilde{\epsilon}^2\big]. \tag{5}$$

Although the MATE estimator is unbiased, it is a single observation estimate since the trial is typically conducted only once. As a result, there is uncertainty in the MATE. Another possible way to control the variance in MATE is to increase the population size used in the RCT. However, the variance can still be undesirably large for the affordable population size. To empirically show this issue, we measured the MATE for 10 000 different random treatment assignments in the IHDP dataset [9], [10], [11] for different sub-population sizes. Fig. 1 shows the empirical probability density distribution of the MATE. Clearly, the variance shrinks as the population size grows; however, variance might still be undesirable in sensitive tasks for the affordable population sizes (in this case $N < 600$).

Since a typical trial is conducted once, only one treatment assignment can be used for the trial. Thus, it is of high importance to ensure that one selected treatment assignment is selected properly [7]. Even in the case of proper randomization, it may be important to check whether the selected treatment assignment has imbalanced covariates by chance. Furthermore, it is common in practice that some participants dropout before the trial is finished. The drop-outs make the trial population different from the original population which was used in the randomization, which in turn might induce a selection bias. For all of the mentioned reasons, it is important to check for baseline imbalances.

Sometimes p-value based hypothesis testing is used to check whether the treatment assignment is selected properly. This usage has been recognized as illogical [21], [22], [23]. "Such significance tests assess the probability (i.e., P-value) that observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance" [24], [25]. As a result, p-values for baseline differences does not serve a useful purpose since it is not testing a useful scientific hypothesis [21], [23], [26], [27]. Later in this section balancing scores are discussed as tools that should be used to evaluate the baseline comparability.

C. MATE Variance Reduction

There have been numerous efforts to reduce the estimation variance of the MATE. Covariate adjustment and CBMs are two families of such efforts.

Covariate adjustment tools reduce the effects of baseline imbalances on the estimated ATE using different regression models.

Some believe that any dissimilarity in the statistics of the two groups can be compensated using covariate adjustment methods, such as ANCOVA [28], [29], [30], [31], [32], [33]. Thus, it is of no interest to test for similarity of statistics in the two groups, or try to use treatment assignments with more similar statistics [21].

Several authors have argued against this belief in four main arguments.

1) Covariate adjustment tools have complex statistical properties. Thus, unadjusted findings are preferred by authors and readers because such findings are simpler and have more clarity [27]. It explains why even when deployed, covariate adjusted findings are mostly used as the backup for the unadjusted findings [27].

2) It has been shown that different models can lead to various estimates and maybe even different clinical implications. Potential biased choices out of numerous different model families and parameter settings are one of the reasons of suspicion regarding the potential manipulations of covariate adjustment methods which ultimately make them less credible [27].

3) In some trials, covariate adjustment methods need more than affordable population size to adjust for all the covariates. As a result, those covariates that are expected to be more prognostic would be adjusted. In some trials, there is insufficient clinical agreement or there is lack of confidence on which covariates should be adjusted for [27].

4) Mokhtarian et al. [19], Pearl [34], Huang and Valtorta [35] have also studied the ATE identifiability problem and argued that in some cases, it is not possible to identify ATE in the presence of biases as they introduces some unmeasured confounders to the underlying causal graph.

Pocock et al. [27] have summarized these arguments as: "The scope for judgments in an ill-defined strategy, and biased (for example, most favorable) choices out of a multiplicity of possible analyses, means that covariate adjusted analyses may rightly be viewed with some suspicion, often leaving primary emphasis on the unadjusted analysis."

A common practice in trial reports is to devote "Table I" (also known as patient cohort) to compare the distributions of baseline variables among different treatment groups. In addition to the fact that it helps the reader to decide whether this study can be generalized to another population, there are two main goals in having separate columns for different treatment groups rather than just a single column for the whole population. First, it demonstrates that the randomization worked well, or it can identify any unlucky imbalances. Second, having balanced baseline variables adds credibility to the trial, especially encouraging confidence in the unadjusted analysis [27].

Can covariate adjustment substitute the need for baseline comparability? Although covariate adjustment tools have numerous benefits, following previous paragraphs, they cannot substitute the need for baseline comparability and balanced covariates.

D. Covariate Balancing Methods

CBMs are a family of methods in which treatment assignments with more similarity in the statistics of two groups have a higher chance to be selected for the trial. In CBMs, all of the variables that are expected to be related to the outcome are recorded for the population as the covariates. CBMs try to favor treatment assignments that have more similarity between the covariates' distributions in the two groups. Since the treatment and the control group are "similar" in such balanced treatment assignments, selection bias can thereby be reduced.

In practice, clinical researchers are required to leverage their expertise to ensure that all of the variables that can possibly have an effect on the outcome are recorded as covariates. Therefore, following this standard practice, we assume that there are not important unobserved covariates.

CBMs require a balancing score (also referred to as the covariate balance measure) that evaluates the similarity of the covariate distributions of the control and treatment groups.

The common motivation behind all of the CBMs is to promote similarity of the joint distribution of covariates between the two groups. In the mathematical language, if covariates of each subject are recorded as $\vec{x}^{\,i}$ then $P(\vec{x})$ for the treatment and the control groups should be similar. With the limited population size and high number of covariates, promoting and measuring this similarity becomes intractable in practice. That is where different CBMs relax the problem in different ways.

There are two main categories of RCTs: the nonsequential RCTs where covariates of the whole population are assumed to be accessible before the conductance of the trial and the sequential RCTs where subjects become available sequentially. Sequential and nonsequential CBMs are targeted toward the sequential and nonsequential RCTs, respectively.

1) Nonsequential CBMs: The first step of nonsequential CBM in RCTs includes recording the covariates for the population. Then, balanced treatment assignments are found by minimizing the covariate imbalance among the two groups. In the next stage, the trial is conducted according to the obtained balanced treatment assignment. The MATE, then, is calculated afterward.

There are different implementations for a given CBM. An initial treatment assignment can be selected randomly and then a greedy minimization modifies the treatment assignment until it reaches a desirable balancing score [36]. Alternatively, the whole randomization process can be repeated until a treatment assignment with a desired balance is reached [36]. Another option is that one exhaustively checks all possible treatment assignments to find the treatment assignment that is maximally balanced. Alternatively, one can find a set of acceptable treatment assignments, and then select one of them randomly.

One of the most commonly used CBMs is SMD, the difference of the means of each covariate between the treatment and the control group. To avoid scaling issues, this CBM standardizes the difference of the means of each covariate by the variance of that covariate [36], [37].

The balancing score for SMD is defined as

$$\mathcal{U}_p = \left\| \frac{1}{N_1} \sum_{\text{treatment}} \vec{x}^{\,i} - \frac{1}{N_0} \sum_{\text{control}} \vec{x}^{\,i} \right\|_p, \quad p \in \{1, \infty\} \tag{6}$$

where N_1 and N_0 are the size of the treatment and control group, respectively. And $\vec{x}^{\,i}$ is a vector containing the covariates of the ith subject. Both ℓ_1 and ℓ_∞ can be used for vector norms in cases with more than one covariate.

We assume that all of the covariates have the same variance without loss of generality. If that is not the case, one can simply normalize each covariate by its standard deviation.

Some other nonsequential CBMs has also been proposed. In [38], three different CBMs are proposed based on the propensity score as a scalar representation for the covariates of each individual. Using the propensity score concept, the three proposed CBMs are: 1) the difference of means of the propensity scores normalized to the variances; 2) the ratio of the variance of the propensity scores in the control and the treatment group; and finally, 3) the ratio of the variance of each covariate orthogonal to the propensity score in the treatment and the control group.

2) Sequential CBMs: Another recognized category of CBMs is sequential treatment assignment. In many of the trials, especially in the medical trials, the whole population is not accessible at once, and the population recruitment is performed sequentially. Even if the whole population is available at the beginning of the trial, there is always a possibility that some of them dropout from the trial or more subjects get added to the trial to increase quality of the results. The sequential treatment assignment can handle the mentioned situations.

One of the most popular sequential treatment assignment methods is proposed by Pocock and Simon [8]. We highly encourage the reader to study this method from the original source but we include a simplified executive summary of its binary version as Algorithm 2 in the Appendix for the ease of convenience.

Several other sequential treatment assignment methods have also been proposed to promote covariate balance [8], [39], [40]. P&S admits only categorical covariates. Alternative methods such as [41], [42] can be used in the presence of continuous covariates. Another alternative would be to use data-clustering methods such as K-means [43] to categorize continuous variables. A larger number of clusters will result in a finer categorization and thus less information loss. In this article, we assume that the continuous covariates are all categorized before any further analysis.

In this article, we investigate worst case vulnerability of one of the most used CBMs in each category of sequential and nonsequential treatment assignment. SMD is one of the most used nonsequential CBMs [7], [36], [44] [45], [46], [47], [48], [49], [50], [51], [52]. We also investigate P&S sequential assignment method as one of the well-known sequential CBMs.

SMD compares the means of the two joint distributions and forces covariates in different groups to have similar means. On the other hand, P&S sequential treatment assignment method promotes similarity in the marginal distributions of different covariates, which is a stronger similarity than the SMD. In the next sections, we provide arguments on the effects of promoting stronger similarity on the adversarial vulnerability.

III. Worst Case Treatment Assignments

In this article, for the first time, we empirically find worst case treatment assignments for the SMD and P&S sequential assignment method. Then we analyze the empirical results to study worst case behaviors of the given CBMs.

A. Definitions

To formally define worst case treatment assignments, some concepts should be defined beforehand.

The covariate balancing score $\mathcal{U}$ (also referred to as the balancing measure) is a scalar function that returns the amount of covariate imbalance of a given treatment assignment. It is noted that a higher covariate balancing score means that the treatment assignment is more imbalanced. *The expected imbalance* $\bar{\mathcal{U}}$ is the expected value of the covariate balance measurement $\mathcal{U}$ over all the possible treatment assignments in the trial. *The minimum imbalance* $\mathcal{U}_{\min}$ is the minimum value of $\mathcal{U}$ over all the possible treatment assignments in the trial.

The admissible treatment assignment set $\tilde{\mathcal{A}}$ is defined as the set of all the treatment assignments

$$\tilde{\mathcal{A}} = \left\{ \mathcal{A} \mid \forall \mathcal{A}' \,,\ \frac{\mathcal{U}(\mathcal{A}) - \mathcal{U}(\mathcal{A}')}{\bar{\mathcal{U}}} < \alpha_a \right\} \tag{7}$$

where $\alpha_a \ll 1$ is a parameter that controls the amount of balance induced by the CBM. Larger α_a relaxes the covariate balancing and allows for more treatment assignments to be admissible.

To quantify the vulnerability of a given RCT to worst case treatment assignments, we measure the maximum possible deviation of MATE in the admissible treatment assignments' set.

We define *worst case deviation factor* ξ as the range of the measured ATE by different admissible treatment assignments, normalized by the standard deviation of the measured ATE over random treatment assignments

$$\xi = \frac{\text{Range}\big[\text{MATE}(\tilde{\mathcal{A}})\big]}{2\sigma_{\text{ATE}}}. \tag{8}$$

B. Worst Case Assignments for SMD in Nonsequential Trials

We are interested in finding worst case treatment assignments of the SMD as CBM in the trial.

1) Worst Case Treatment Assignment for SMD: Assume that the potential outcomes of assigning each subject to the treatment or the control group are provided for a population size of N. The potential outcome for the subject i being assigned to the treatment group or the control group is y_i^1 and y_i^0, respectively. For each subject in the population, covariates are provided as an M-dimensional vector $\vec{x}^{\,i}$. The goal is to find the treatment assignment $\mathcal{A}^* : \{1, 2, \ldots, N\} \to \{0, 1\}^N$ dividing the population into two groups with equal sizes such that it maximizes the MATE and minimizes the covariate balancing score $\mathcal{U}_p$. We use Lagrange multipliers to formulate a combinatorial optimization problem over the space of all possible treatment assignments

$$\mathcal{A}_p^* = \text{argmax}_{\mathcal{A}} \big(\lambda\, \text{MATE}(\mathcal{A}) - \mathcal{U}_p(\mathcal{A})\big), \quad p = \{1, \infty\}. \tag{9}$$

The above problem is a combinatorial optimization problem over the space of all possible treatment assignments. ATASTREET converts the above problem to a constrained linear programming problem and solves it using mixed integer linear programming tools in an acceptable time [53], [54], [55], [56], [57], [58], [59]. More details are provided in the Appendix.

C. Worst Case Treatment Assignments for Sequential Trials

Finding worst case treatment assignments of the sequential CBMs is even more challenging since the treatment assignment of one subject affects the treatment assignment of the next subjects. We approach this challenge by providing a nonsequential balancing score

$$\mathcal{U}_{\text{P\&S}} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right| \tag{10}$$

where m is the number of covariates, N_i is the total number of categories for ith covariate,[2] and $N^i_{\text{control}}(j)$ is the total number of subjects in the control group with their ith covariate having the value of jth category, and $N^i_{\text{treatment}}(j)$ is the same for the treatment group. Then, we provide a theorem that tightly links our proposed balancing score to P&S sequential treatment assignment method (Algorithm 2).

Theorem 1: In P&S sequential treatment assignment method, using $\mathcal{U}_{\text{P\&S}}$ instead of G in P&S method (Algorithm 2) results in the same decision rule.

For the proof, see the Appendix.

The above theorem suggests that P&S sequential method is in fact a sequential greedy probabilistic minimization over a nonsequential CBM with $\mathcal{U}_{\text{P\&S}}$ as its balancing score. Putting the randomnesses aside, P&S sequential method favors treatment assignments with smaller $\mathcal{U}_{\text{P\&S}}$. The goal of our worst case analysis of P&S method would be to find treatment assignments that are favored by P&S method the most, and have maximum possible ATE estimation error.

Arguments in the previous paragraph motivate us to find the adversarial treatment assignments of the mentioned nonsequential CBM. Then, each of the resulting worst case treatment assignments should carefully be analyzed to see whether they are feasible to get selected by P&S sequential method.

1) Worst Case Treatment Assignment for P&S CBM: Assume that y_i^1, y_i^0, and $\vec{x}^{\,i}$ are given similar to worst case analysis for SMD. The goal is to find the treatment assignment $\mathcal{A}^* : \{1, 2, \ldots, N\} \to \{0, 1\}^N$ dividing the population into two groups with equal sizes such that it maximizes the

[2]Recall that the covariates are assumed to be categorical in P&S sequential assignment method in Algorithm 2.

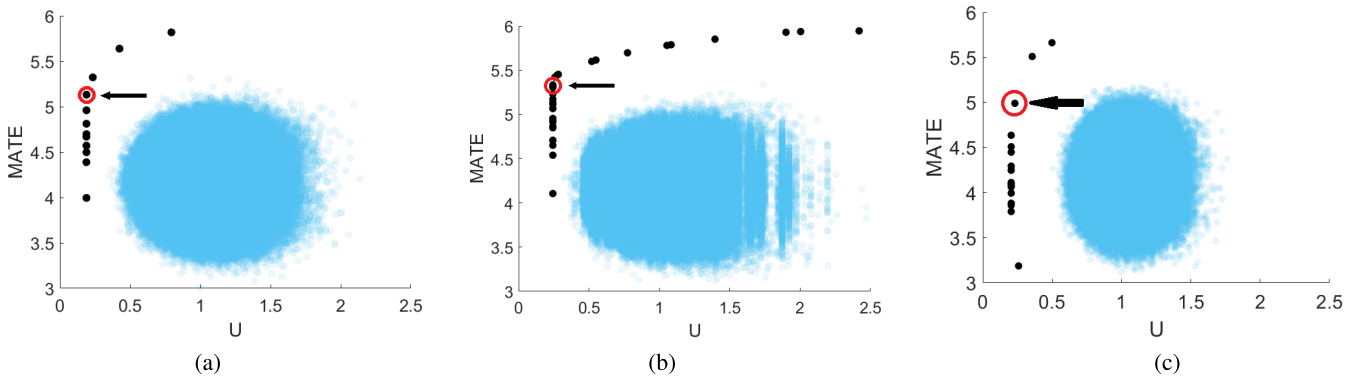

Fig. 2. We visualize ATASTREET results for different λs on the IHDP1000 dataset (black points). A reference set of randomly selected treatment assignments of IHDP is also visualized as blue dots. The $\mathcal{U}_p$ axis is normalized to $\bar{\mathcal{U}}$. This plot shows ATASTREET solutions in comparison to the reference set (blue points). The marked red circled points could have been selected in the IHDP trial. The perfect balance of covariates would encourage confidence in the trial with a large ATE estimation error. This plot shows vulnerability of the mentioned CBMs to worst case treatment assignment. (a) SMD with $\mathcal{U}$. (b) SMD with $\mathcal{U}$. (c) $\mathcal{U}$.

MATE and minimizes the covariate balancing score $\mathcal{U}_{\text{P\&S}}$. We use Lagrange multipliers to formulate a combinatorial optimization problem over the space of all possible treatment assignments

$$\mathcal{A}^*_{\text{P\&S}} = \text{argmax}_{\mathcal{A}} \ (\lambda \ \text{MATE}(\mathcal{A}) - \mathcal{U}_{\text{P\&S}}(\mathcal{A})). \qquad (11)$$

Similar to the previous case where we covered SMD for nonsequential RCTs, we obtain ATASTREET solution using mixed linear integer programming [53], [54], [55], [56], [57], [58], [59]. More details are provided in the Appendix.

D. Empirical Results of Worst Case Analysis

To provide a better understanding of worst case treatment assignments, we introduce a new visualization technique for different possible treatment assignments in the same trials. Each treatment assignment is visualized as a single point with its corresponding $\mathcal{U}$ as the horizontal coordinate, and its corresponding MATE as the vertical coordinate.

We used our introduced visualization technique to visualize ATASTREET's resulting treatment assignments for different parameter λ (Shown as black point in Fig. 2). A set of random treatment assignments with no CBM is also shown in each plot with blue points to act as a reference.

Several remarks follow from these results.

The CBMs in both our cases, the SMD for nonsequential case and P&S method for sequential case, are vulnerable against worst case treatment assignments. Analyzing the ATASTREET's resulting treatment assignments for different values of λ reveals some of the worst case treatment assignments (see Fig. 2). According to the results of our experiments, $\xi > 6$. In another language, it is possible to find admissible treatment assignments where groups are well-balanced, but the MATE has error higher than $6\sigma_{\text{ATE}}$.

Following the previous argument, both analyzed CBMs are vulnerable against worst case assignments. This vulnerability opens up unwanted potentials for deviations (intended or unintended) with considerable effects on the MATE. Restricting such potentials is very important in some applications like medical trials. In Fig. 2, the corresponding treatment assignment of the black point marked with the red circle is admissible with regards to having balanced covariates, yet yields a larger ATE estimation error than all the 10^6 random treatment assignment shown as blue points.

In the sequential case, it is not clear whether the worst case treatment assignments associated with $\mathcal{U}_{\text{P\&S}}$ are feasible to get selected by P&S sequential method. To demonstrate their feasibility, we considered different orders of subjects coming into the trial, and we set $P_0 = 1$ (Algorithm 2) to ensure that P&S sequential method would never go toward the unlikely path. We found several different subject ordering where the evolution path goes to any of the predetermined assignments in ATASTREET results (11). Although we do not provide any theoretical proof that ATASTREET solutions are always feasible for selection by P&S method with $P_0 = 1$, we have empirically provided several different paths for each of the resulting ATASTREET's assignments (Fig. 3). Furthermore, oftentimes, $P_0 < 1$ in practice. It means that any treatment assignment is now possible to get selected by P&S sequential method. Arguments regarding posterior probability of worst case assignments getting selected is out of the scope of this article. To summarize arguments in this section, we have empirically found treatment assignment evolution paths that P&S sequential assignment method ends up in each of the worst case assignments (Fig. 3).

Many RCT applications use an unequal allocation of the control and treatment groups. For simplicity of explanation, in this article, we have assumed equal population sizes for these groups. However, ATASTREET can easily handle arbitrary allocation ratios, as we detail in the Appendix. Our empirical results suggest no significant difference in the MATE error of adversarial assignments for different allocation ratios.

Our empirical results for different choices of $\mathcal{U}_1$ and $\mathcal{U}_\infty$ as different versions of SMD suggests that our arguments do not depend on the vector norm used in the SMD (6). We infer that the observed vulnerability is inherent in the SMD, and not the deployed vector norm.

P&S method is slightly better than SMD (Fig. 4). Even though P&S method has smaller worst case deviation factor ξ, it is still vulnerable and more CBMs should be investigated to

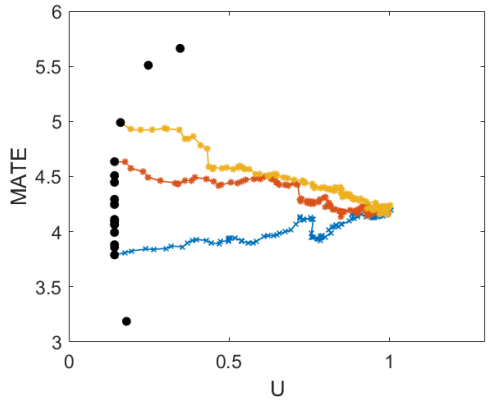

Fig. 3. Evolution of the treatment assignment sequence as new subjects are introduced to the trial. The $\mathcal{U}$ axis here represents the expected value of final $\mathcal{U}$ if the next subjects are to be assigned randomly (the expected value is approximated using Monte-Carlo method). This plot shows how P&S sequential assignment method reaches the worst case treatment assignments found using ATASTREET (11).

find CBMs with smaller ξs. One can modify ATASTREET for different CBMs to find their worst case treatment assignments and compare their worst case deviation factors ξ. Ultimately, the most worst case robust CBM could be identified. Such CBM is ideal in cases where the clinical implications of the RCT is important and large errors in ATE estimation would inflict intolerable losses to health or financial resources.

IV. How Close Is a Trial to Worse Case?

In Section III, we have empirically demonstrated that the two investigated CBMs are vulnerable to worst case assignments. It brings up an important question. *How to ensure a trial is not close to the worst case?* We answer this question by providing the *CBM deviation index* ρ.

A variety of ITE estimation tools can be used to assess the estimated ATE error for a given RCT [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]. Once the error interval is acquired, one can simply compare it to σ_{ATE} to interpret it as a unit-less number. In that case, the deployed treatment assignment is compared to random treatment assignments without CBM. To interpret the ATE estimation error in RCTs where CBM is used, we suggest comparing the ATE estimation error to the worst case error in the similar balancing scores.

We define the *CBM deviation index* ρ as the ratio of ATE estimation error to the worst case error in the similar balancing score. The actionable summary on how to measure ρ in any given trial without having access to unobserved counterfactual outcome is provided as Algorithm 1.

ITE estimation methods provide noisy imperfect estimates of the ITE as well as the unobserved potential outcomes for each subject. Using these methods to estimate the unobserved counterfactual outcomes compromises the efficiency of worst case assignments found by ATASTREET.

To investigate this, we formed a reconstructed version of IHDP by picking a realization of IHDP, then picked a treatment assignment at random and gave only the observed outcomes and deployed treatment assignment to GANITE [60]. Then, the estimated unobserved potential outcomes and the observed outcomes would form our reconstructed version of IHDP.

Algorithm 1 CBM Deviation Index ρ

1: Using the state-of-the-art ITE estimation method, estimate the unobserved counterfactual outcomes for all the subjects.
2: Approximate the ATE estimation error $|\tilde{\epsilon}(\mathcal{A})|$ in the RCT using the estimated potential outcomes.
3: Run ATASTREET on the resulting RCT table. Use a set of different parameter λ In ATASTREET.
4: Make a plot similar to Fig. 2.
5: Connect the resulting ATASTREET treatment assignments so that they form a continuous contour. For this purpose, the finer sweep for parameter λ results in a better approximation of the mentioned contour.
6: Using the deployed treatment assignment in the RCT, calculate the balancing score.
7: In ATASTREET contour, find a point with a balancing score equal to the balancing score of the deployed treatment assignment in the trial. This $\tilde{\epsilon}_{\max}$ is the maximum possible ATE estimation error in treatment assignments with similar balancing scores.
8: Report the CBM deviation index $\rho = \dfrac{|\tilde{\epsilon}(\mathcal{A})|}{\tilde{\epsilon}_{\max}}$.

To investigate the effect of using noisy estimates of unobserved potential outcomes, we took 15 random realizations of the reconstructed version of IHDP and found worst case treatment assignments using ATASTREET. Then we used ground truth from IHDP and measured the ground truth for ATE of the resulting assignments, Our results suggest that this imperfection resulted in estimating the *worst case deviation factor* ξ as five times smaller than it is true value. Indeed, using better ITE estimators results in better measurements of the worst case deviation factor ξ as well as the CBM deviation index ρ.

V. Toward Adversarial Attacks of Clinical Trials

In this section, we develop an adversarial attack to RCTs with mentioned CBMs. To do this, we provide an actionable summary of how to find adversarial treatment assignments for any given trial using ATASTREET.

In the previous sections, we empirically demonstrated that the mentioned CBMs are vulnerable to worst case treatment assignments. We then provided an index to check whether a given RCT is close to the worst case. To further emphasize the importance of such sanity check, we develop an adversarial attack to any given RCT and demonstrate that an adversary can use such attack to deceitfully deviate the MATE while having maximally balanced groups.

Can someone exploit this vulnerability and find adversarial treatment assignment in a given RCT? We uncovered the worst case assignments of the given CBMs using ATASTREET as an oracle method which has access to the ground truth values

TABLE I
P&S METHOD

	$\tilde{\epsilon}_{\text{adv}}$	ξ	ρ
Mean	1.50	7.87	0.20
Std	1.25	6.34	0.08
Max	5.33	24.05	0.37

TABLE II
SMD WITH ℓ_∞

	$\tilde{\epsilon}_{\text{adv}}$	ξ	ρ
Mean	1.51	5.93	0.21
Std	1.63	2.33	0.13
Max	6.25	11.89	0.52

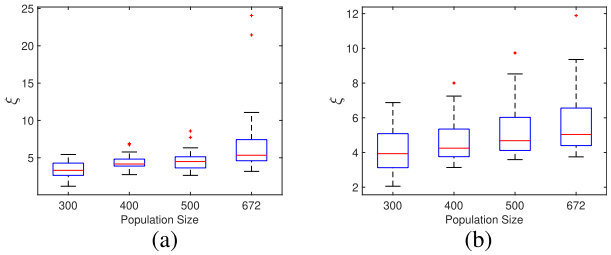

Fig. 4. Worst case deviation factor ξ for different population sizes measured for $\alpha_a = 0.02$ (7), (8). (a) P&S method. (b) SMD with $\mathcal{U}$.

of the unobserved counterfactual outcomes. In this section, we provide an actionable summary on how to find adversarial treatment assignments in any given RCT.

For any given RCT, pick the state-of-the-art ITE estimation method, and use the observed outcomes as well as the deployed treatment assignment to estimate the unobserved counterfactual outcomes for all the subjects, then form the reconstructed version of the given trial. We argue that the worst case assignments of the reconstructed version serve as adversarial assignments for the given RCT.

To empirically demonstrate this argument, we took 15 random realizations of IHDP1000, then formed the reconstructed version similar to the previous section by removing half of the observed potential outcomes and estimating them using GANITE. We found worst case assignments of the reconstructed version, and used the ground-truth values of potential outcomes in IHDP1000 to evaluate the resulting MATE of the adversarial treatment assignments. In Tables I and II, $\tilde{\epsilon}_{\text{adv}}$ is the resulting ATE estimation error of our adversarial attack normalized to σ_{ATE}, ξ is the worst case deviation factor in IHDP, and ρ is the efficiency of our introduced attack. As our result suggest, our introduced adversarial attack results in $\rho = 0.2$, which means that our introduced adversarial attack has the ATE estimation error five times smaller than the worst case assignment.

Using ITE estimators with better accuracy results in less estimation error in reconstruction of the RCTs. Counterfactual outcome estimation and ITE estimation are active research areas [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]. Introducing methods with higher accuracy results in adversarial treatment assignments closer to the worst case assignments (bigger ρ). Note that none of the ITE estimation methods outperforms all others in all settings. Therefore, depending on the application, data setting, and model assumptions, researchers and practitioners should carefully choose an appropriate ITE estimation method for their specific application.

Using ITE estimators with better accuracy results in less estimation error in reconstruction of the RCTs. Counterfactual outcome estimation and ITE estimation are active research areas and introducing methods with higher accuracy, results in adversarial treatment assignments closer to the worst case assignments (bigger ρ).

We investigated the effect of population size on the adversarial vulnerability of the analyzed CBMs. To do this, we randomly sub-sampled a population from the original population and found ATASTREET solutions, then plotted the resulting adversarial vulnerability factor ξ for different population sizes in Fig. 4. Unlike the MATE variance that shrinks with increasing population size, the MATE estimation error in adversarial cases would not shrink by increasing population size. As a result, the worst case deviation factor ξ increases with larger population sizes. Therefore, increasing the population size does not alleviate the adversarial vulnerability problem. It makes it even worse. However, increasing the population size is beneficial in another aspect and that is, matching the distributions of covariates in the control and the treatment group becomes more tractable, and higher quality CBMs can be used. It is still worth mentioning that increasing the population size would not alleviate the adversarial vulnerability in any of the given CBMs.

One might naturally think that by introducing randomness, or changing the stop criteria in the CBM procedure, the mentioned adversarial treatment assignments would be less likely to get selected. Examples of this would be to limit the number of iterations in SMD minimization in nonsequential cases, or to select a smaller p_0 in P&S method. However, it is rather running away from the problem instead of solving it. The gap between the black and blue points in the Fig. 2 is filled with other possible treatment assignments. Limiting the extent of using CBM would make it impossible for the current adversarial treatment assignments to be selected, but introduces even worse adversarial assignments. Note that the MATE of black points increases as more imbalance $\mathcal{U}$ is allowed.

VI. CONCLUSION

In this work, we have provided arguments to demonstrate that the SMD CBM and P&S sequential assignment method, two of the most used approaches to reduce selection bias in RCTs, are vulnerable to worst case treatment assignments (Fig. 2). To demonstrate these vulnerabilities, we proposed ATASTREET to find well-balanced treatment assignments where the studied CBMs fail in preventing large errors in the MATE. It uncovers a drawback for these CBMs and suggests

that these CBMs should not be used to evaluate reliability of the results in RCTs. The worst case vulnerability opens up opportunities for deceitful activities to exploit adversarial treatment assignments to deviate the MATE toward a desired ATE.

We provided an index to check whether a given RCT that used CBM is close to worst case treatment assignments. We also developed adversarial attacks to any given RCT to show that a deceitful researcher can take advantage of worst case vulnerability.

Our work suggests interesting future research directions. One direction is to assess the adversarial robustness of additional CBMs so that they could be ranked based on their potential for adversarial robustness. A complementary direction is identifying the CBM with the best adversarial robustness. Such a method would be highly desirable in cases where the nature of the trial has a high importance level that brings the need to use a method that is robust against any deceitful action (e.g., clinical trials during deadly pandemics).

APPENDIX

An executive summary of P&S sequential treatment assignment method is given as Algorithm 2.

A. P&S Sequential Assignment Method

Algorithm 2 P&S Sequential Binary Treatment Assignment.

1: For the few initial subjects, it does not matter how to assign them. Randomly assign the few first subjects to the treatment or the control group.
2: A new subject comes to the clinic and the goal is to assign them to one of the groups.
3: Assume the new subject is assigned to either of the groups, e.g. the treatment group.
In P&S method, covariates are assumed to be categorical. In case some of the covariates are continuous-valued, they should be discretized to different categories.
4: For each of the covariates, namely the i^{th} covariate, The new subject has the value of the j^{th} category for the i^{th} covariate. Count Number of subjects in the control and in the treatment group having the value of the j^{th} category for the i^{th} covariate. Define $d_i = d(N_{\text{treatment}}, N_{\text{control}})$ where $d(x, y)$ is a distance function. The most natural case for the binary treatment case is $d(x, y) = |x - y|$.
5: Define G as a (weighted) sum of d_is. $G = \sum_{i=1}^{m} \alpha_i d_i$. The weights α_i could be arbitrarily selected to emphasize balancedness in some of the covariates.
6: Go back to step 3 and this time, assign the new subject to the other group.
7: Sort two different resulting Gs for assigning the new subject to each of the groups. Flip a coin with the probability of being head equal to a pre-specified probability of P_0. If the coin was head, assign the subject to the group resulting in the smaller G; And if it was tail, assign it to the group resulting in the bigger G (Note that P_0 should be bigger than 0.5)
8: For the next subject, go back to step 2 and repeat the same procedure.

B. Proof of Theorem 1

Theorem 2: In P&S sequential treatment assignment method, the way a new subject is assigned to a group, minimizes $\mathcal{U}_{\text{P\&S}}$ with the probability of p_0. In other words, $\mathcal{U}_{\text{P\&S}}$ can be used instead of G in P&S method (Algorithm 2).

Proof: The goal is to prove $\mathcal{U}_{\text{P\&S}}$ with

$$\mathcal{U}_{\text{P\&S}} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right|$$

instead of G in Algorithm 2 results in same probability of assigning the subject to each of the treatment or control groups. Assume that the current subject has the value of c_i for the i^{th} covariate. Then immediately by the definition of G we have

$$G = \sum_{i=1}^{m} \alpha_i d_i = \sum_{i=1}^{m} \alpha_i \left| N^i_{\text{control}}(c_i) - N^i_{\text{treatment}}(c_i) \right|.$$

By adding and subtracting a term, we can write it as

$$\begin{aligned} &= \sum_{i=1}^{m} \sum_{j=1}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right| \\ &\quad - \sum_{i=1}^{m} \sum_{j=1, j\neq c_i}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right| \\ &= \mathcal{U}_{\text{P\&S}} - \sum_{i=1}^{m} \sum_{j=1, j\neq c_i}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right|. \end{aligned}$$

Now note that the second term is a positive number that would remain constant for different assignments of the current subject

$$G_2 - G_1 = \mathcal{U}_{\text{P\&S},2} - \mathcal{U}_{\text{P\&S},1}.$$

Thus, $\mathcal{U}_{\text{P\&S}}$ could be used instead of G in Algorithm 2 and result in the same decision.

We have introduced the adversarial attack to find adversarial treatment assignments in the manuscript, but did not provide details on how ATASTREET incorporates mixed linear programming to solve the given combinatorial optimization problems. Here, mathematical details for different versions of ATASTREET are provided.

C. ATASTREET for SMD With ℓ_1

To find adversarial attacks of the SMD with ℓ_1, one has to solve the optimization problem in (9)

$$\text{argmax}_{\mathcal{A}} \left(\lambda \, \text{MATE}(\mathcal{A}) - \frac{2}{N} \left\| \sum_{\text{treatment}} \vec{x}^{\,i} - \sum_{\text{control}} \vec{x}^{\,i} \right\|_1 \right)$$

$$\text{argmax}_{\mathcal{A}} \left(\lambda \sum_i \left(\mathcal{A}_i y_i^1 - (1 - \mathcal{A}_i) y_0^1 \right) \right.$$

$$- \left\| \sum_i (2\mathcal{A}_i - 1)\vec{x}^{\,i} \right\|_1 \Bigg).$$

Then, by throwing away a term that does not depend on the $\mathcal{A}$, we can write down the argmax problem as

$$\text{argmax}_{\mathcal{A}} \left(\lambda \sum_i \mathcal{A}_i \left(y_i^1 + y_0^1\right) - \left\| \sum_i (2\mathcal{A}_i - 1)\vec{x}^{\,i} \right\|_1 \right).$$

By introducing auxiliary variables t_j^+, t_j^-, this argmax problem can then be written as an argmin problem and then be solved using mixed integer linear programming tools

$$\text{argmin}_{\mathcal{A}, t^+, t^-} \left(-\lambda \sum_i \mathcal{A}_i \left(y_i^1 + y_0^1\right) + \sum_{j=1}^{m} \left(t_j^+ + t_j^-\right) \right)$$

$$\forall j, \quad t_j^+ - t_j^- = \sum_i \left(\mathcal{A}_i - \frac{1}{2}\right)\vec{x}_j^{\,i}$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{2} \right\rfloor$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^- , \quad \mathcal{A}_i \in \mathbb{N}.$$

As discussed in the article, ATASTREET can also handle unequal allocation ratios. Assuming that the ratio for treatment:control is $1 : \Psi$, and with some math, it can be shown that the general case of ATASTREET for unequal allocation ratio is

$$\text{argmin}_{\mathcal{A}, t^+, t^-} \left(-\lambda \sum_i \mathcal{A}_i \left(\Psi y_i^1 + y_0^1\right) + \sum_{j=1}^{m} \left(t_j^+ + t_j^-\right) \right)$$

$$\forall j, \quad t_j^+ - t_j^- = \sum_i \left(\mathcal{A}_i - \frac{1}{\Psi + 1}\right)\vec{x}_j^{\,i}$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{\Psi + 1} \right\rfloor$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^- , \quad \mathcal{A}_i \in \mathbb{N}.$$

D. ATASTREET for SMD With ℓ_∞

To find adversarial attacks of the SMD with ℓ_1, one has to solve the optimization problem in (9)

$$\text{argmax}_{\mathcal{A}} \left(\lambda \, \text{MATE}(\mathcal{A}) - \frac{2}{N} \left\| \sum_{\text{treatment}} \vec{x}^{\,i} - \sum_{\text{control}} \vec{x}^{\,i} \right\|_\infty \right)$$

$$\text{argmax}_{\mathcal{A}} \left(\lambda \sum_i \left(\mathcal{A}_i y_i^1 + (1 - \mathcal{A}_i) y_0^1\right) - \left\| \sum_i (2\mathcal{A}_i - 1)\vec{x}^{\,i} \right\|_\infty \right).$$

Then, by throwing away a term that doesn't depend on the $\mathcal{A}$, we can write down the argmax problem as

$$\text{argmax}_{\mathcal{A}} \left(\lambda \sum_i \mathcal{A}_i \left(y_i^1 + y_0^1\right) - \left\| \sum_i (2\mathcal{A}_i - 1)\vec{x}^{\,i} \right\|_\infty \right).$$

By introducing auxiliary variables t_j^+, t_j^-, T, this argmax problem can then be written as an argmin problem and then be solved using mixed integer linear programming tools

$$\text{argmin}_{\mathcal{A}, T, t_j^+, t_j^-} \left(-\lambda \sum_i \mathcal{A}_i \left(y_i^1 + y_0^1\right) + T \right)$$

$$\forall j, \quad t_j^+ - t_j^- = \sum_i \left(\mathcal{A}_i - \frac{1}{2}\right)\vec{x}_j^{\,i}$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{2} \right\rfloor$$

$$\forall j, \quad t_j^+ + t_j^- \le T$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^-, T , \quad \mathcal{A}_i \in \mathbb{N}.$$

As discussed in the article, ATASTREET can also handle unequal allocation ratios. Assuming that the ratio for treatment:control is $1 : \Psi$, and with some math, it can be shown that the general case of ATASTREET for unequal allocation ratio is

$$\text{argmin}_{\mathcal{A}, T, t_j^+, t_j^-} \left(-\lambda \sum_i \mathcal{A}_i \left(\Psi y_i^1 + y_0^1\right) + T \right)$$

$$\forall j, \quad t_j^+ - t_j^- = \sum_i \left(\mathcal{A}_i - \frac{1}{\Psi + 1}\right)\vec{x}_j^{\,i}$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{\Psi + 1} \right\rfloor$$

$$\forall j, \quad t_j^+ + t_j^- \le T$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^-, T , \quad \mathcal{A}_i \in \mathbb{N}.$$

E. ATASTREET for $\mathcal{U}_{\text{P\&S}}$

To find adversarial attacks of the P&S assignment method, one has to solve the optimization problem in (11)

$$\text{argmax}_{\mathcal{A}} \left(\lambda \, \text{MATE}(\mathcal{A}) - \sum_{i=1}^{m} \sum_{j=1}^{N_i} \alpha_i \left| N^i_{\text{control}}(j) - N^i_{\text{treatment}}(j) \right| \right).$$

To implement $\mathcal{U}_{\text{P\&S}}$ in a linear format, we write it as

$$\mathcal{U}_{\text{P\&S}} = \left\| \tilde{X}\left(2\vec{\mathcal{A}} - 1\right) \right\|_1$$

where $\tilde{X}$ is a matrix formed as below

$$\tilde{X} = \left[\begin{array}{cccc} d_1^1(1) & d_1^1(2) & \dots & d_1^1(N) \\ d_2^1(1) & d_2^1(2) & \dots & d_2^1(N) \\ \vdots & & \ddots & \vdots \\ d_{N_1}^1(1) & d_{N_1}^1(2) & \dots & d_{N_1}^1(N) \\ \hline d_1^2(1) & d_1^2(2) & \dots & d_1^2(N) \\ d_2^2(1) & d_2^2(2) & \dots & d_2^2(N) \\ \vdots & & \ddots & \vdots \\ d_{N_2}^2(1) & d_{N_2}^2(2) & \dots & d_{N_2}^2(N) \\ \hline \vdots & & \ddots & \vdots \\ \hline d_1^m(1) & d_1^m(2) & \dots & d_1^m(N) \\ d_2^m(1) & d_2^m(2) & \dots & d_2^m(N) \\ \vdots & & \ddots & \vdots \\ d_{N_m}^m(1) & d_{N_m}^m(2) & \dots & d_{N_m}^m(N) \end{array}\right] \tag{12}$$

where $d_j^i(k) = 1 \iff k$th subject has the jth category for $i$thcovariate.

Similar to previous section, by introducing auxiliary variables t_j^+, t_j^-, this argmax problem can then be written as an argmin problem and then be solved using mixed integer linear programming tools

$$\text{argmin}_{\mathcal{A}, t_j^+, t_j^-} \left(-\lambda \sum_i \mathcal{A}_i \left(y_i^1 + y_0^1\right) + \sum_{j=1}^{N_0+\dots+N_m} \left(t_j^+ + t_j^-\right) \right)$$

$$1 \le \forall j, \le N_0 + \dots + N_m \quad t_j^+ - t_j^- = \sum_i (2\mathcal{A}_i - 1)\tilde{X}(j,i)$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{2} \right\rfloor$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^- , \quad \mathcal{A}_i \in \mathbb{N}.$$

As discussed in the article, ATASTREET can also handle unequal allocation ratios. Assuming that the ratio for treatment:control is $1 : \Psi$, and with some math, it can be shown that the general case of ATASTREET for unequal allocation ratio is

$$\text{argmin}_{\mathcal{A}, t_j^+, t_j^-} \left(-\lambda \sum_i \mathcal{A}_i \left(\Psi y_i^1 + y_0^1\right) + \sum_{j=1}^{N_0+\dots+N_m} \left(t_j^+ + t_j^-\right) \right)$$

$$1 \le \forall j, \le N_0 + \dots + N_m \quad t_j^+ - t_j^-$$

$$= \sum_i \left(\mathcal{A}_i - \frac{1}{\Psi + 1} \right) \tilde{X}(j,i)$$

$$\sum_i \mathcal{A}_i = \left\lfloor \frac{N}{\Psi + 1} \right\rfloor$$

$$0 \le \mathcal{A}_i \le 1, \quad 0 \le t_j^+, t_j^- , \quad \mathcal{A}_i \in \mathbb{N}.$$

REFERENCES

[1] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, Oct. 1974.

[2] T. C. Chalmers et al., "A method for assessing the quality of a randomized control trial," *Controlled Clin. Trials*, vol. 2, no. 1, pp. 31–49, May 1981.

[3] E. Hariton and J. J. Locascio, "Randomised controlled trials—The gold standard for effectiveness research," *BJOG, Int. J. Obstetrics Gynaecol.*, vol. 125, no. 13, p. 1716, Dec. 2018.

[4] K. Benson and A. J. Hartz, "A comparison of observational studies and randomized, controlled trials," *New England J. Med.*, vol. 342, no. 25, pp. 1878–1886, Jun. 2000.

[5] J. Concato, N. Shah, and R. I. Horwitz, "Randomized, controlled trials, observational studies, and the hierarchy of research designs," *New England J. Med.*, vol. 342, no. 25, pp. 1887–1892, Jun. 2000.

[6] A. Deaton and N. Cartwright, "Understanding and misunderstanding randomized controlled trials," *Social Sci. Med.*, vol. 210, pp. 2–21, Aug. 2018.

[7] S. Athey and G. W. Imbens, "The econometrics of randomized experiments," in *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 2017, pp. 73–140.

[8] S. J. Pocock and R. Simon, "Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial," *Biometrics*, vol. 31, pp. 103–115, Mar. 1975.

[9] A. Multisite, "Enhancing the outcomes of low-birth-weight, premature infants: A multisite, randomized trial," *J. Amer. Med. Assoc.*, vol. 263, pp. 3035–3042, Jun. 1990.

[10] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *J. Comput. Graph. Statist.*, vol. 20, no. 1, pp. 217–240, Jan. 2011.

[11] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3076–3085.

[12] A. Curth, D. Svensson, J. Weatherall, and M. Van Der Schaar, "Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, 2021, pp. 1–14.

[13] P. W. Holland, "Statistics and causal inference," *J. Amer. Stat. Assoc.*, vol. 81, no. 396, pp. 945–960, 1986.

[14] A. P. Dawid, "Causal inference without counterfactuals," *J. Amer. Stat. Assoc.*, vol. 95, no. 450, pp. 407–424, 2000.

[15] J. Pearl, "Causal inference without counterfactuals: Comment," *J. Amer. Stat. Assoc.*, vol. 95, no. 450, pp. 428–431, 2000.

[16] I. Shpitser and J. Pearl, "What counterfactuals can be tested," in *Proc. 23rd Conf. Artif. Intell. (UAI)*, 2007, pp. 352–359.

[17] S. Akbari, E. Mokhtarian, A. Ghassami, and N. Kiyavash, "Recursive causal structure learning in the presence of latent variables and selection bias," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10119–10130.

[18] Y. Kivva, E. Mokhtarian, J. Etesami, and N. Kiyavash, "Revisiting the general identifiability problem," in *Proc. 38th Conf. Uncertainty Artif. Intell.*, 2022, pp. 1022–1030.

[19] E. Mokhtarian, F. Jamshidi, J. Etesami, and N. Kiyavash, "Causal effect identification with context-specific independence relations of control variables," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 11237–11246.

[20] I. Shpitser and J. Pearl, *Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models*. Los Angeles, CA, USA: Univ. California, 2006.

[21] S. Senn, "Testing for baseline balance in clinical trials," *Statist. Med.*, vol. 13, no. 17, pp. 1715–1726, Sep. 1994.

[22] L. A. Harvey et al., "Electrical stimulation plus progressive resistance training for leg strength in spinal cord injury: A randomized controlled trial," *Spinal Cord*, vol. 48, no. 7, pp. 570–575, Jul. 2010.

[23] D. G. Altman, "Comparability of randomised groups," *J. Roy. Stat. Soc., D, Statistician*, vol. 34, no. 1, pp. 125–136, 1985.

[24] K. F. Schulz, D. G. Altman, and D. Moher, "CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials," *J. Pharmacol. Pharmacotherapeutics*, vol. 1, no. 2, pp. 100–107, Dec. 2010.

[25] K. F. Schulz, D. G. Altman, and D. Moher, "Consort 2010 statement: Updated guidelines for reporting parallel group randomized trials," *Ann. Internal Med.*, vol. 152, no. 11, pp. 726–732, 2010.

[26] P. C. Austin, A. Manca, M. Zwarenstein, D. N. Juurlink, and M. B. Stanbrook, "A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: A review of trials published in leading medical journals," *J. Clin. Epidemiol.*, vol. 63, no. 2, pp. 142–153, Feb. 2010.

[27] S. J. Pocock, S. E. Assmann, L. E. Enos, and L. E. Kasten, "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practiceand problems," *Statist. Med.*, vol. 21, no. 19, pp. 2917–2930, 2002.

[28] H. J. Keselman et al., "Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses," *Rev. Educ. Res.*, vol. 68, no. 3, pp. 350–386, Sep. 1998.

[29] G. J. P. Van Breukelen, "ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies," *J. Clin. Epidemiol.*, vol. 59, no. 9, pp. 920–925, Sep. 2006.

[30] A. Rutherford, *ANOVA ANCOVA: A GLM Approach*. Hoboken, NJ, USA: Wiley, 2011.

[31] D. B. Wright, "Comparing groups in a before–after design: When t test and ANCOVA produce different results," *Brit. J. Educ. Psychol.*, vol. 76, no. 3, pp. 663–675, Sep. 2006.

[32] T. R. Johnson, "Violation of the homogeneity of regression slopes assumption in ANCOVA for two-group pre-post designs: Tutorial on a modified Johnson–Neyman procedure," *Quant. Methods Psychol.*, vol. 12, no. 3, pp. 253–263, Oct. 2016.

[33] J. Jamieson, "Analysis of covariance (ANCOVA) with difference scores," *Int. J. Psychophysiol.*, vol. 52, no. 3, pp. 277–283, May 2004.

[34] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[35] Y. Huang and M. Valtorta, "Identifiability in causal Bayesian networks: A sound and complete algorithm," in *Proc. AAAI*, 2006, pp. 1149–1154.

[36] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Stat. Sci.*, vol. 25, no. 1, pp. 1–21, Feb. 2010.

[37] P. R. Rosenbaum and D. B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *Amer. Statistician*, vol. 39, no. 1, pp. 33–38, Feb. 1985.

[38] D. B. Rubin, "Using propensity scores to help design observational studies: Application to the tobacco litigation," *Health Services Outcomes Res. Methodology*, vol. 2, no. 3, pp. 169–188, 2001.

[39] O. Atan, W. R. Zame, and M. van der Schaar, "Adaptive clinical trials: Exploiting sequential patient recruitment and allocation," 2018, *arXiv:1810.02876*.

[40] D. R. Taves, "Minimization: A new method of assigning patients to treatment and control groups," *Clin. Pharmacol. Therapeutics*, vol. 15, no. 5, pp. 443–453, May 1974.

[41] Z. Ma and F. Hu, "Balancing continuous covariates based on kernel densities," *Contemp. Clin. Trials*, vol. 34, no. 2, pp. 262–269, Mar. 2013.

[42] F. Hu, Y. Hu, Z. Ma, and W. F. Rosenberger, "Adaptive randomization for balancing over covariates," *WIREs Comput. Statist.*, vol. 6, no. 4, pp. 288–303, Jul. 2014.

[43] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[44] T.-L. Nguyen and L. Xie, "Incomparability of treatment groups is often blindly ignored in randomised controlled trials—A post hoc analysis of baseline characteristic tables," *J. Clin. Epidemiol.*, vol. 130, pp. 161–168, Feb. 2021.

[45] K. Imai, G. King, and E. A. Stuart, "Misunderstandings between experimentalists and observationalists about causal inference," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 171, no. 2, pp. 481–502, Apr. 2008.

[46] P. C. Austin, "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples," *Statist. Med.*, vol. 28, no. 25, pp. 3083–3107, Nov. 2009.

[47] M. S. Ali et al., "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review," *J. Clin. Epidemiol.*, vol. 68, no. 2, pp. 122–131, Feb. 2015.

[48] P. C. Austin, "Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement," *J. Thoracic Cardiovascular Surgery*, vol. 134, no. 5, pp. 1128–1135, Nov. 2007.

[49] P. C. Austin, "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003," *Statist. Med.*, vol. 27, no. 12, pp. 2037–2049, 2008.

[50] E. Gayat, R. Pirracchio, M. Resche-Rigon, A. Mebazaa, J.-Y. Mary, and R. Porcher, "Propensity scores in intensive care and anaesthesiology literature: A systematic review," *Intensive Care Med.*, vol. 36, no. 12, pp. 1993–2003, Dec. 2010.

[51] G. Lonjon, R. Porcher, P. Ergina, M. Fouet, and I. Boutron, "Potential pitfalls of reporting and bias in observational studies with propensity score analysis assessing a surgical procedure," *Ann. Surgery*, vol. 265, no. 5, pp. 901–909, 2017.

[52] W. G. Cochran, "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*, vol. 24, no. 2, pp. 295–313, 1968.

[53] L. A. Wolsey and G. L. Nemhauser, *Integer and Combinatorial Optimization*, vol. 55. Hoboken, NJ, USA: Wiley, 1999.

[54] E. Klotz and A. M. Newman, "Practical guidelines for solving difficult mixed integer linear programs," *Surveys Oper. Res. Manage. Sci.*, vol. 18, nos. 1–2, pp. 18–32, Oct. 2013.

[55] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, no. 9, pp. 917–922, Sep. 1977.

[56] D. S. Nau, V. Kumar, and L. Kanal, "General branch and bound, and its relation to A* and AO*," *Artif. Intell.*, vol. 23, no. 1, pp. 29–58, May 1984.

[57] J. Clausen, "Branch and bound algorithms-principles and examples," Dept. Comput. Sci., Univ. Copenhagen, Copenhagen, Denmark, 1999, pp. 1–30.

[58] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, Jul. 1960.

[59] A. H. Land and A. G. Doig, *An Automatic Method for Solving Discrete Programming Problems*. Springer, 2010.

[60] J. Yoon, J. Jordon, and M. Van Der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–22.

[61] A. Alaa and M. Van Der Schaar, "Validating causal inference models via influence functions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 191–201.

[62] Z. Qian, Y. Zhang, I. Bica, A. Wood, and M. Van Der Schaar, "SyncTwin: Treatment effect estimation with longitudinal outcomes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3178–3190.

[63] A. Curth and M. van der Schaar, "Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1810–1818.

[64] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 3020–3029.

[65] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Nonparametric tests for treatment effect heterogeneity," *Rev. Econ. Statist.*, vol. 90, no. 3, pp. 389–405, Aug. 2008.

[66] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Appl. Statist.*, vol. 4, no. 1, pp. 266–298, 2010.

[67] A. M. Alaa and M. Van Der Schaar, "Bayesian inference of individualized treatment effects using multi-task Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[68] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[69] S. Wager, S. Athey, S. Wager, and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, Jul. 2018.

[70] S. Tabib and D. Larocque, "Non-parametric individual treatment effect estimation for survival data with random forests," *Bioinformatics*, vol. 36, no. 2, pp. 629–636, Jan. 2020.

[71] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran, "Estimating individual treatment effect in observational data using random forest methods," *J. Comput. Graph. Statist.*, vol. 27, no. 1, pp. 209–219, 2017.

[72] J. Hoogland et al., "A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint," *Statist. Med.*, vol. 40, no. 26, pp. 5961–5981, Nov. 2021.

[73] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," *Rev. Econ. Statist.*, vol. 86, no. 1, pp. 4–29, Feb. 2004.

[74] Q. Ge et al., "Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection," *Frontiers Genet.*, vol. 11, Dec. 2020, Art. no. 585804.

[75] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[76] S. Powers et al., "Some methods for heterogeneous treatment effect estimation in high dimensions," *Statist. Med.*, vol. 37, no. 11, pp. 1767–1787, May 2018.

Hossein Babaei was born in Iran in 1996. He received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2019. He is currently pursuing the Ph.D. degree with Rice University, Houston, TX, USA.

He works on developing causal inference tools for learning systems under the supervision of Prof. Richard G. Baraniuk.

Sina Alemohammad received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2019. He is currently pursuing the Ph.D. degree with Rice University, Houston, TX, USA, under the supervision of Prof. Richard G. Baraniuk.

His research interests include deep learning theory, signal processing, and optimization.

Richard G. Baraniuk (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Manitoba, Winnipeg, MB, Canada, in 1987, the M.S. degree from the University of Wisconsin–Madison, Madison, WI, USA, in 1988, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992.

He is currently the Victor E. Cameron Professor of electrical and computer engineering with Rice University, Houston, TX, USA, and the Founding Director of OpenStax. He holds 35 U.S. and six foreign patents. His research interests include new theory, algorithms, and hardware for sensing, signal processing, and machine learning.

Dr. Baraniuk is a fellow of the American Academy of Arts and Sciences, the National Academy of Inventors, and the American Association for the Advancement of Science. He was a recipient of the DOD Vannevar Bush Faculty Fellow Award (National Security Science and Engineering Faculty Fellow), the IEEE James H. Mulligan, Jr. Education Medal, and the IEEE Signal Processing Society Technical Achievement, Education, Best Paper, Best Magazine Paper, and best column awards.