When Is Heterogeneity Actionable for Targeting?[*]

Anya Shchetkina
The Wharton School
University of Pennsylvania

Ron Berman
The Wharton School
University of Pennsylvania

August 2024

Abstract

We apply five popular personalization approaches to two large-scale field experiments with many interventions (aka megastudies) aimed at increasing vaccination rates (the Walmart study of Milkman et al. (2022) and the Penn-Geisinger study of Milkman et al. (2021) and Patel et al. (2023)). We find limited value of targeting in the Walmart experiment and a four times higher value of targeting in the Penn-Geisinger experiment. We seek to explain the difference in the gains from personalization between the two studies and show that the presence of heterogeneity alone is not sufficient to predict whether a targeting exercise will be successful. Instead, a specific form of heterogeneity, which we call "actionable" heterogeneity, determines the value of targeting. We demonstrate how the amount of actionable heterogeneity depends on three forces: (1) within- and (2) cross-treatment heterogeneity, as well as (3) cross-treatment correlation. For studies with many interventions, such as the ones we analyze, determining the magnitude of actionable heterogeneity can be challenging. To aid this task, we develop a model that estimates the value of personalized policies compared to the best untargeted intervention using three simple summary statistics of the data. We find that the value of actionable heterogeneity of the Penn-Geisinger study is higher than that of the Walmart study, which can explain the difference in the observed values of targeting. Our model also illustrates conditions when adding more treatments to an experiment may hurt the value of targeting even in infinite samples.

[*] Research reported in this paper was made possible by Flu Lab, the Social Science Research Council's Mercury Project (with funding from the Rockefeller Foundation, the Robert Wood Johnson Foundation, Craig Newmark Philanthropies, and the Alfred P. Sloan Foundation), the Wharton Dean's Research Fund, and the United States-Israel Binational Science Foundation [Grant 2020022]. We thank the Behavior Change for Good Initiative for making this research possible. We also thank Wendy De La Rosa, Ryan Dew, Katherine Milkman, Christophe Van den Bulte, participants at CDSM'23, CODE@MIT'23, Marketing Modelers Meeting, and the 4 School Conference, as well as three anonymous EC'24 reviewers for their feedback.

1 Introduction

A/B tests (online randomized control trials) are a popular strategy to find effective interventions in economics (e.g., Blake et al., 2015; Azevedo et al., 2020), marketing (e.g., Feit and Berman, 2019; Fong and Hunter, 2022), operations (e.g., Zhang et al., 2020; Ye et al., 2023), information systems (e.g., Burtch et al., 2015; Bauman and Tuzhilin, 2018), and data science (e.g., Johari et al., 2017; Jamieson and Jain, 2018). Different policies and interventions are tested against each other in real-world contexts and the best-performing intervention is deployed. The randomized experimental design ensures internal validity, and the field setting makes findings realistic.

A/B tests can be leveraged to further improve desired outcomes (such as revenue, policy outreach, etc.) using targeting and personalization policies.[1] In addition to treatment assignment and the outcome of interest, experimenters can collect covariates (e.g., demographics, purchase history, or location) to compute Heterogeneous Treatment Effects (HTEs), which are the effects of treatment on specific subgroups, rather than the entire population. This enables analysts to determine the optimal intervention for each subgroup instead of applying the same policy to all. For instance, analysis might reveal that shorter promotional messages are more effective for younger individuals, while longer ones work better for older individuals. Targeted interventions of this nature have the potential to significantly enhance outcomes, and therefore it is important to understand under what conditions HTEs are effective for targeting.

There are multiple ways one can approach a targeting task. Ascarza (2018) and Athey et al. (2023) caution against relying on popular heuristics, such as targeting high-risk individuals, and instead advise to explicitly model heterogeneity. Typically, such modeling involves three stages: training, prediction, and optimization.[2] (1) Initially, a flexible machine learning model is trained on experimental data linking each treatment arm and individual covariates to observed outcomes. (2) In the prediction stage, the trained model forecasts counterfactual outcomes for each individual

[1] Throughout this paper, we will use the terms "targeting" and "personalization" interchangeably, meaning an assignment of an intervention (including, possibly, no intervention) to an individual based on their observable characteristics.

[2] Similar stages are involved in other applications of machine learning tools to real-world decisions. Recent literature emphasized that this approach may not be optimal (see, e.g., Elmachtoub and Grigas, 2022; Chung et al., 2022).

under all possible treatment assignments. (3) Finally, the treatment with the highest predicted outcome for that individual becomes the recommended personalized policy. However, there are also other methods outside this paradigm that were developed for targeting. The causal tree method (Athey and Imbens, 2016) directly searches for heterogeneous treatment effects, combining steps (1) and (2). The causal forest (Wager and Athey, 2018) generalizes this method and is now extensively used for targeting (see, e.g., Davis and Heller, 2017; Luo et al., 2019; Bonander and Svensson, 2021). Similarly, Hitsch et al. (2023) develop a non-parametric approach to estimate HTEs. Another approach is to combine all three stages and directly optimize the targeting policy. Examples include outcome-weighted learning (Zhao et al., 2012) and Policy DNN (Zhang, 2023). The majority of these methods focus on experiments with binary treatments but are not easily extendable to experiments with many interventions.

Empirically, past research has found mixed evidence regarding the effectiveness of personalization. A few examples reporting effective targeting policies are identifying geographical regions for targeted lockdowns during the COVID-19 pandemic (Acemoglu et al., 2021), refugee placement (Ahani et al., 2021), teacher-to-classroom assignment (Graham et al., 2022), cancer outreach interventions (Chen et al., 2020), advertising in mobile apps (Rafieian and Yoganarasimhan, 2021), and promotion of household energy conservation (Knittel and Stolper, 2019). Across this variety of contexts, researchers have identified substantial benefits from targeting that sometimes exceeded 100% increase in outcome level relative to uniform (untargeted) policies.

However, in other contexts, particularly in experiments with many interventions, researchers did not find such large gains from personalization. Yoganarasimhan et al. (2023) find limited benefits from personalizing free trial lengths: the uniform policy assigning the shortest trial length to everyone outperformed causal forest-based targeting methods. Dubé and Misra (2023) report advantages of personalized pricing, yet the added value of personalization does not significantly surpass the confidence interval of the best uniform policy. Smith et al. (2023) find that machine learning targeting methods yield effects ranging from -31% to +15% compared to the best uniform policy, depending on available data inputs. Perdomo et al. (2023) show that individual school dropout risk scores do not provide targeting opportunities that go beyond the information contained in environmental variables.

What makes personalization effective in some cases and not effective in others? One explanation might be that the studies that show little value of targeting do not have enough observable heterogeneity: if everyone reacts similarly to an intervention, it does not make sense to personalize it, and the best policy is to deploy the best-performing intervention uniformly. Another issue that might arise in experiments with multiple interventions is correlation in responses to treatments across individuals. For example, two versions of a landing page that only have a minor difference may appeal to the same people, making targeting ineffective. This paper shows that while heterogeneity and lack of correlation among interventions are important, neither alone suffices to indicate the value of targeting over the best-performing uniform policy. Instead, we demonstrate that only a specific form of heterogeneity can create value from targeting. Namely, to be "actionable" for targeting, heterogeneity that moderates the treatment effects within different subgroups is not sufficient. It also needs to have a few subgroups for which the most successful intervention is not the same. Visually, this appears as a crossover between one intervention and another, if the individuals are ranked by their treatment effects. We show that the magnitude of crossovers with the best-performing intervention determines what portion of heterogeneity is "actionable" for targeting.

To illustrate the concept of actionable heterogeneity, we analyze two large-scale field experiments with many interventions: the Walmart flu shots study Milkman et al. (2022) and the Penn-Geisinger flu shots study Milkman et al. (2021); Patel et al. (2023). In these studies, 22 and 19 behavioral nudges informed by psychological theory were tested concurrently to improve flu vaccination rates. We evaluate several popular targeting approaches and find a relatively small gain from personalization (3% relative improvement over the best uniformly applied treatment) in the Walmart study and a more substantial value of targeting (13% relative improvement) in the Penn-Geisinger study.

We develop a statistical model of the value from targeting and show that the magnitude of crossovers is affected by three forces: (1) within-treatment heterogeneity (the variation of individual responses for the same intervention), (2) cross-treatment heterogeneity (the variation of average responses across interventions), and (3) cross-treatment correlation (how independent are responses to different interventions for the same individual). We describe how this model can be used to gauge the potential from personalization *before* running an experiment if a researcher has prior expectations for the amounts of within- and cross-heterogeneity and cross-correlation. We find that, surprisingly, sometimes having more interventions can hurt the potential gain from personalization. In addition, we show how the model can be applied *after* running an experiment, taking into account the amount of prediction error in estimation of counterfactual outcomes. This method can point a researcher who is concerned with lackluster returns to personalization into the right direction: whether to find more precise estimation methods, or to experiment with other interventions and collect more data. When we apply the model to both megastudies we analyze (Walmart and Penn-Geisinger), we find that the Walmart study has a lower amount of actionable heterogeneity, which can explain the difference in the targeting values we find.

To summarize, this paper offers three contributions. Section 2 makes a substantive contribution by estimating the value of personalization in two large-scale field experiments with many interventions. Section 3 makes a theoretical contribution: we show that heterogeneity alone is not a sufficient indicator of the potential value from targeting. Instead, the magnitude of crossovers with the best-performing intervention determines the value of targeting, and the magnitude of crossovers is in turn influenced by three moments of the data: within-treatment heterogeneity, cross-treatment heterogeneity, and cross-treatment correlation. Section 4 applies this model and shows how to gauge the personalization potential prior to running an experiment, compare the targeting potential across different studies, and distinguish between inefficiencies of targeting methodologies and a lack of actionable heterogeneity.

2 The Value of Targeting in Two Large-Scale Experiments

2.1 Description of the Experiments

We assess the value from targeting using data from two large-scale field experiments. Both these experiments were conducted in a "one-shot", non-adaptive setting, meaning that the treatment assignment was done once and all people received at most one intervention. Throughout our analysis, we will focus on this setting, leaving beyond the scope of this paper personalization in online adaptive experiments (see, e.g., Schmit and Johari, 2018; Liao et al., 2020; Goldenberg et al., 2021; Rafieian, 2023; Ghosh et al., 2024).

The first study (Milkman et al., 2022, the "Walmart study") analyzed the impact of low-cost behavioral nudges on vaccination rates. Independent teams of behavioral researchers designed 22 text reminders informed by psychological theory to encourage people to get their seasonal flu shot at Walmart. On average, these text reminders increased vaccination rates by 2.0 percentage points compared to the business-as-usual control group. The dataset includes covariates such as gender, age, insurance type, health information, race, and zipcode-level variables such as median income and ethnic composition. Summary statistics are provided in Table 1. Figure 1 displays the average vaccination rates for each intervention on a training sample (70% of the population). Interventions are ordered by decreasing response rates.

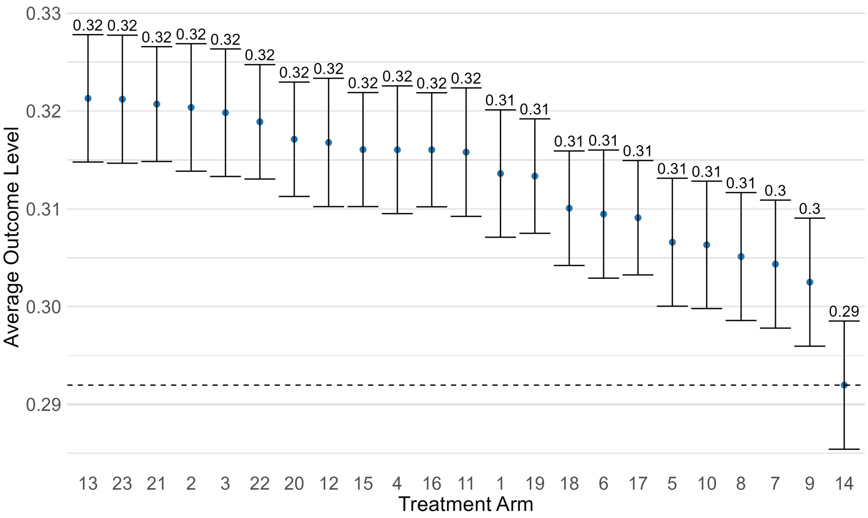
The second study (Milkman et al., 2021; Patel et al., 2023, the "Penn-Geisinger study") also investigated the impact of text nudges on flu shot uptake, but in a different context. Behavioral researchers developed 19 text messages to be sent to individuals with upcoming appointments at Penn Medicine or Geisinger Health, two large health systems in the Northeastern United States. On average, the text nudges led to a 1.8 percentage point increase in vaccination rates. The dataset includes a wide array of covariates, such as health system, gender, age, insurance type, health information, smoking status, weight, marital status, race, whether the patient received flu shots in 2015–2019, message sending date, and zipcode-level median income. Figure 2 summarizes the average vaccination rates for each intervention on a training sample (70% of the population). Interventions are ordered by decreasing response rates.

Table 1: Summary Statistics of Megastudies

Dataset	# Observations	# Arms	#Covariates		
			Total	Continuous	Discrete
Walmart	689,693	23	12	5	7
Penn-Geisinger	74,811	20	22	7	15

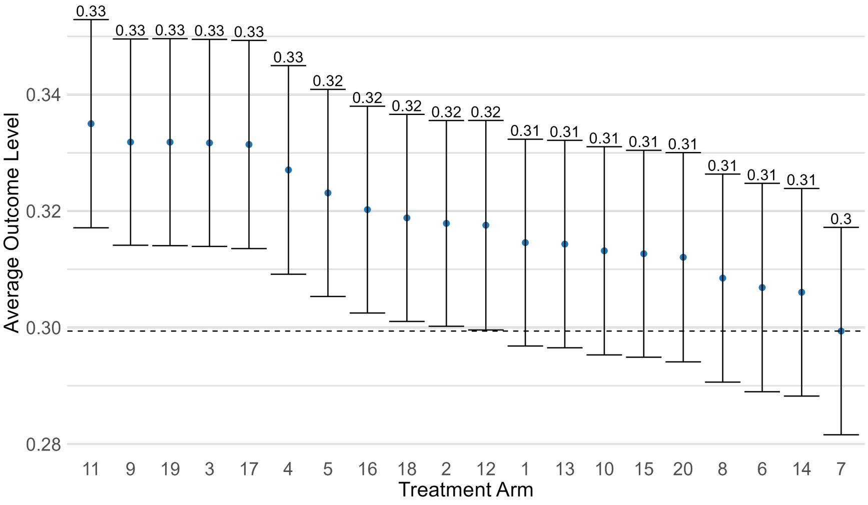
Note: Arms denote the interventions (22 and 19 text messages respectively), as well as the business-as-usual control.

Figure 1: Interventions of the Walmart Study



The average flu vaccination rates across 23 interventions on a training sample (70%) ordered by outcome levels. Intervention 14 is the control (no text reminders).

Figure 2: Interventions of the Penn-Geisinger Study



The average flu vaccination rates across 20 interventions on a training sample (70%) ordered by outcome levels. Intervention 7 is the control (no text reminders).

2.2 Description of the Targeting Methods

For both experiments, we estimate and evaluate five popular targeting methods: OLS, S-Learner XGBoost, T-Learner XGBoost (Chen and Guestrin, 2016), Causal Forest (Athey et al., 2019), and Policy Tree (Zhou et al., 2023), which are described in detail below. We chose a sample of methods to demonstrate various common approaches to targeting: (i) a simple linear targeting method (OLS), (ii) targeting methods based on a general multi-purpose machine learning model (XGBoost), (iii) a targeting approach based on a specialized machine learning model for uncovering HTEs (Causal Forest), and finally (iv) a machine learning model fully specialized for targeting (Policy Tree). A summary comparing the characteristics of these methods is provided in Table 2. We allocate 70% of the dataset of each experiment for training the personalization policy, while the remaining 30% is reserved for evaluation.

Table 2: Targeting Methods Properties

	# Models	Nonlinearities	Objective
OLS	1	No	Prediction
S-XGB	1	Yes	Prediction
T-XGB	# Arms	Yes	Prediction
MACF	1	Yes	Treatment effects
PT	1	Yes	Optimal policy

We will use the following notation. An experiment involves n participants, where $i = 1, 2, \ldots, n$. Each participant is characterized by covariates X_i and is randomly assigned to an intervention A_i, with $A_i \in \mathcal{A} = \{1, 2, \ldots, m\}$ being a categorical variable. The observed outcome of individual i is denoted by Y_i. A targeting policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from covariates to interventions, i.e. a targeting policy is a rule that selects a recommended intervention based on a person's observed characteristics. This definition highlights the role of covariates: a targeting policy can only be effective if the available covariates capture meaningful differences between people (Rossi et al., 1996; Smith et al., 2023). We do not focus on how to select the best covariates or which covariates are best for targeting, but rather on estimating the best policy given all available covariates.

We implement five popular targeting policies. To maintain parsimony, when we write A_i, we mean the indicator for being exposed to intervention A_i.

OLS. In this targeting approach, the outcome variable is modeled by a single linear regression involving all covariates, all treatments, and all two-way interactions between treatments and covariates:

$$Y_i = \beta X_i + \gamma_a A_i + \delta_a X_i \times A_i + \varepsilon_i \tag{1}$$

The model is then used to predict the outcome variable for a given individual i for each treatment assignment a:

$$\widehat{Y_i^a} = \hat{\beta} X_i + \hat{\gamma_a} a + \hat{\delta_a} X_i \times a \tag{2}$$

The intervention with the highest predicted outcome is selected as the targeting policy:

$$\pi_{OLS}(X_i) = \arg\max_a \widehat{Y_i^a} \tag{3}$$

S-Learner XGBoost. If treatment effects are nonlinear in X_i, OLS may be suboptimal. To account for potential nonlinearities, we employ XGBoost (Chen and Guestrin, 2016), which is a gradient tree boosting algorithm. In the S-Learner version, the intervention indicator is treated as a regular feature fed into the algorithm, and a single model f is trained for all observations:

$$Y_i = f(X_i, A_i) + \varepsilon_i \tag{4}$$

Similarly to OLS, we predict the outcome for each individual i and treatment assignment a, and choose the intervention yielding the highest prediction.

$$\pi_{S-XGB}(X_i) = \arg\max_a \hat{f}(X_i, a) \tag{5}$$

T-Learner XGBoost. In contrast to S-Learners, which consist of a single model, T-Learners employ a model for each intervention. The overall sample is divided into subsamples, one for each arm (for arm a, the subsample consists of all individuals i such that $A_i = a$). This ensures that the interventions are incorporated into the modeling process, even if their predictive strength is relatively low compared to the covariates (Hu, 2023). We evaluate a T-Learner variant of the

XGBoost algorithm, where separate XGBoost models are trained for each subsample:

$$Y_i = f_a(X_i) + \varepsilon_i \tag{6}$$

where f_a is trained on the portion of data with $A_i = a$. The predictions from all models are then compared, and the intervention corresponding to the model with the highest predicted outcome is chosen:

$$\pi_{T-XGB}(X_i) = \arg\max_a \hat{f}_a(X_i) \tag{7}$$

Multi-arm Causal Forest. The methods above can be characterized as "predict-then-optimize" (Elmachtoub and Grigas, 2022; Chung et al., 2022). That is, targeting is done in two steps: training a model to predict outcomes for all possible treatment assignments, followed by selecting the highest predicted outcome. Since the objective of these methods is a prediction of outcome levels, they are not necessarily optimal for uncovering heterogeneity (Athey and Imbens, 2016), which is necessary for targeting. To address this, we estimate a multi-arm causal forest, as implemented in the R grf package (Athey et al., 2019; Wager and Athey, 2018; Nie and Wager, 2021). Multi-arm causal forests extend the standard causal forest to more than one intervention. The standard causal forest is designed to identify sub populations with the largest treatment effect heterogeneity.

A multi-arm causal forest outputs $\widehat{\tau^a_{X_i}}$ — an estimated individual-level treatment effect of arm a relative to the baseline arm a_0. To construct a targeting policy, we select the treatment arm with the highest estimated treatment effect (we set $\widehat{\tau^{a_0}_{X_i}} = 0$):

$$\pi_{MACF}(X_i) = \arg\max_a \widehat{\tau^a_{X_i}} \tag{8}$$

Policy tree. The final method we consider is the policy tree (Zhou et al., 2023), which takes the results of the causal forest as input and seeks the optimal targeting policy in the form of a decision tree with a specific depth. We estimate a depth-2 policy tree (since a depth of 3 is not computationally feasible for our datasets) and directly utilize its output as the targeting policy. Because splits at each node are binary, a depth of 2 implies that no more than 4 interventions will be used in a targeting policy derived via a policy tree.

2.3 Evaluation of Policies

We evaluate the policies on the 30% sample holdout using Inverse Probability Weighting (IPW, e.g., Rafieian and Yoganarasimhan, 2023; Simester et al., 2020). Using a holdout sample for policy evaluation is crucial to avoid the winner's curse and ensure an unbiased estimate (Andrews et al., 2024). To evaluate a policy π, we assign to every individual in the holdout sample the treatment prescribed by the policy, which is denoted by $\pi(X_i)$. Subsequently, we identify individuals whose actual experimental treatment assignment corresponds to the one prescribed by the policy, that is, the individuals for whom $A_i = \pi(X_i)$. Finally, we reweight the outcomes of these people according to the propensity scores of receiving the treatment $A_i = \pi(X_i)$:

$$\widehat{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}\{\pi(X_i) = A_i\} Y_i}{\hat{e}(A_i|X_i)} \tag{9}$$

where $\hat{e}(a|X_i)$ is the propensity score of treatment a (the estimated probability of receiving treatment a given the covariates X_i). As both datasets come from randomized experiments, the propensity scores $\hat{e}(a|X_i) = e(a)$ are known.

Table 3 shows the value of targeting from each method. The IPW score of each targeting policy is compared to the uniform benchmark, which identifies the best-performing intervention on the training (70%) sample and reports the mean response in the test (30%) sample.

Table 3: Targeting Results

	Best Uniform	OLS	S-XGB	T-XGB	MACF	PT-CF
Walmart	31.2	31.5	31.4	31.7	32.2	32.2
Bootstrap SE	(0.5)	(0.6)	(0.6)	(0.6)	(0.6)	(0.6)
Improvement		1%	0%	1%	3%	3%
Penn-Geisinger	31.5	32.5	34.9	34.8	35.6	34.1
Bootstrap SE	(1.4)	(1.7)	(1.8)	(1.7)	(1.8)	(1.6)
Improvement		3%	11%	10%	13%	8%

The table presents the comparison of five targeting policies. The benchmark for relative performance is the best uniform policy identified on the training sample and estimated on the test sample. The standard errors are derived by bootstrapping from the test data.

Both datasets exhibit targeting potential (all targeting policies perform better out-of-sample compared to the best uniform policy). However, in the Penn-Geisinger study, all machine learning targeting methods achieve at least 10% relative improvement over the uniform benchmark, while in the Walmart study, the best-performing targeting method (policy tree) only achieves 3% relative improvement over the benchmark. Appendix A.1 provides the details of the best targeting policy in the Penn-Geisinger study.

To summarize, we evaluated five common targeting methods on two large-scale field experiments and found that the Penn-Geisinger study shows a value of targeting that is four times higher than the value of targeting in the Walmart study. We now turn to exploring this difference.

3 What Affects the Value of Targeting?

In this section, we provide a theory for why intuitive factors such as the sample size and the number of covariates are not sufficient to predict the value of targeting. We develop the concept of actionable heterogeneity, which depends on the levels of within-treatment heterogeneity, cross-treatment heterogeneity, and cross-treatment correlation.

3.1 Potential Factors Influencing the Value from Targeting

Intuitively, a difference in the value of targeting might come from differences in statistical power (sample size) or the available information about individuals (number of covariates). Table 4 shows that the Penn-Geisinger experiment has more covariates, while the Walmart experiment has a much larger sample size. That is, these two intuitive factors point in different directions regarding which experiment might have a higher value of targeting.

Another factor that may influence the value of targeting is heterogeneity in outcomes within one treatment. Heterogeneity is a necessary precursor for targeting: all else being equal, the more variation there is in the outcomes of people with different characteristics, the higher value personalization should have. However, if covariates are not predictive of differences in outcomes, targeting is futile. We estimate the amount of within-treatment heterogeneity in the two datasets by using an out-of-sample version of T-Learner XGBoost:

1. For each intervention, estimate an outcome model among the people who received this intervention
2. On the holdout sample, predict the outcome for each person under each intervention
3. Within each intervention, group people into 10 quantiles based on the predicted outcome level and compute the mean outcome level within each quantile
4. Calculate the standard deviation of quantile-level mean outcomes for each intervention and average across interventions.

When applying this procedure to the Penn-Geisinger and the Walmart studies, we estimate the amount of heterogeneity in the Penn-Geisinger dataset at 0.263, and in the Walmart study at 0.074, favoring the Penn-Geisinger study as having potentially larger benefits from personalization.

However, there is another factor that can influence the value from targeting, particularly in studies with many interventions: the level of correlation in responses across treatments. If different treatments work on the same group of people (e.g., young females respond well to treatments A and B while other people do not), choosing a personalized action will not generate high returns even if there is a lot of variation in responses within each treatment. On the other hand, if interventions are independent, or even negatively correlated, each additional intervention may affect a group of people unaffected by other interventions, and thereby increase the targeting potential. To estimate the correlation of treatments given the observed covariates we also use an out-of-sample version of T-Learner XGBoost:

1. For each intervention, estimate an outcome model among the people who received this intervention
2. On the holdout sample, predict the outcome for each person under each intervention
3. Within each intervention, group people into 10 quantiles based on the predicted outcome level and compute the mean outcome level within each quantile
4. Assign the mean outcome level within the person's quantile as a predicted response for a given treatment

5. Compute the correlation matrix of the resulting predictions across interventions

6. Calculate the mean of the off-diagonal elements of the correlation matrix.

The average correlation between interventions is equal to 0.65 in the Walmart study and to 0.81 in the Penn-Geisinger study. That is, the interventions in the Walmart study are more independent, favoring the Walmart study as having potentially larger benefits from personalization.

Table 4: Factors That Can Influence the Value from Targeting

	Walmart	Penn-Geisinger
Number of covariates	12	22
More is better		✓
Sample size	689,693	74,811
More is better	✓	
Within-treatment heterogeneity	0.074	0.263
More is better		✓
Cross-treatment correlation	0.65	0.81
Less is better	✓	
Cross-treatment heterogeneity	0.07	0.07
Less is better	✓	✓

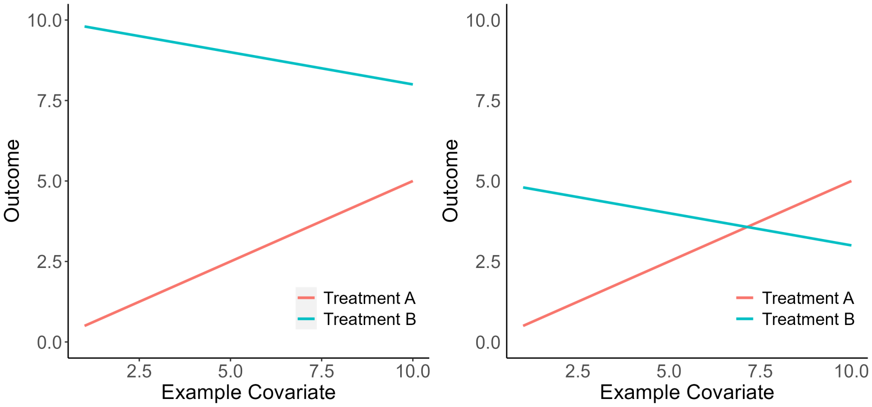
As can be seen from Table 4, the factors that can influence the value from targeting point in different directions regarding which experiment might have a higher potential value from personalization. Next, we present a framework of actionable and useless heterogeneity that highlights the interplay between these factors. This framework allows us to arrive at a composite measure of targeting potential that quantifies the difference between the two studies we analyze and also offers a generalizable tool applicable to a wide range of experiments. By identifying and quantifying actionable heterogeneity, researchers can assess the potential benefits of personalization in their specific contexts.

3.2 Actionable Heterogeneity

To provide intuition, we will start with an example of an experiment where only one covariate is available and the treatment indicator takes two values (e.g., an A/B test). Even in this simple setting, heterogeneity does not inherently facilitate targeting. Figure 3 presents two panels each with a possible result of this hypothetical experiment. Each panel contains two lines, which represent the expected outcome under treatment A or treatment B for each value of the covariate. Both panels feature an interaction of the treatment with the covariate (i.e., heterogeneity).

In the left panel, despite treatment B's effectiveness being significantly moderated by the covariate, there is no "crossover": treatment B performs better for all individuals, regardless of their covariate value. Consequently, the uniform policy assigning treatment B will consistently outperform any targeting policy involving both treatments. On the other hand, in the right panel, the targeting policy that allocates treatment A to individuals with high covariate values and treatment B to those with low values will outperform both uniform policies.[3]

Figure 3: Example — Crossover Interactions



An example of two A/B tests, in both of which the treatment effect is moderated by a covariate. However, in the left panel, targeting is not possible, while it is effective in the right panel.

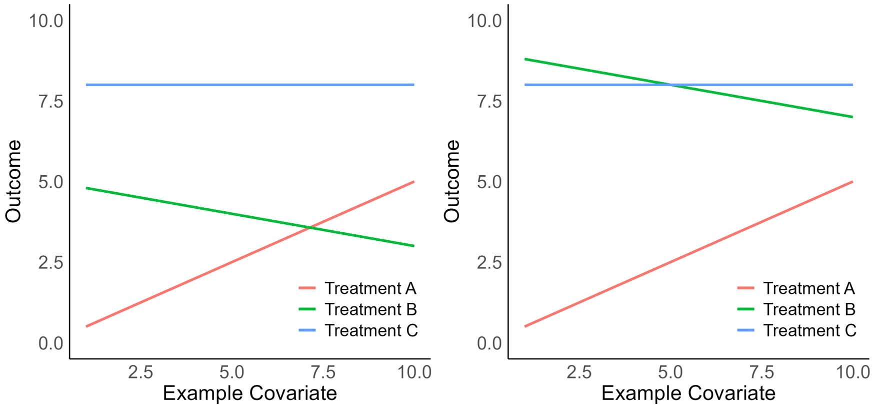
We now generalize to an experiment with more than two treatments. In this case, the presence of a crossover is no longer sufficient. Figure 4 presents two panels each with a possible result of a hypothetical experiment with three interventions. As in the previous example, the figure depicts outcome levels for three treatments conditional on a covariate. Both panels exhibit a crossover interaction.

In the left panel, despite the crossover interaction between treatments A and B, targeting is not effective because treatment C outperforms both across all covariate values. In the right panel, while treatment C remains the average best performer, treatment B outperforms it for low values of the covariate, indicating potential returns to targeting. In other words, in studies with multiple

[3] This discussion assumes that if the costs of treatments A and B are different, the respective costs are subtracted from the outcome before plotting.

interventions, the targeting potential is determined by the presence and magnitude of crossover interactions with the best uniform treatment.

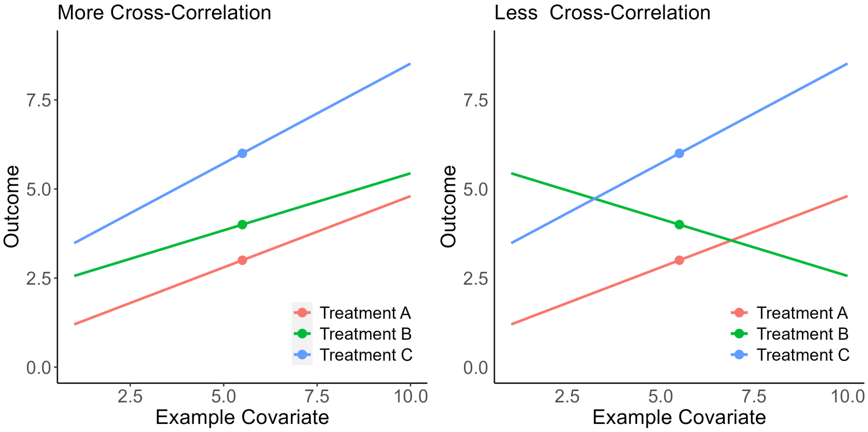
Figure 4: Example — Dominating Arm



An example of two three-arm tests, in both of which there is a crossover interaction. However, in the left panel, targeting is not possible, while it is effective in the right panel.

To summarize, heterogeneity is actionable for targeting when it takes the form of a crossover with the best uniform intervention. Using this framework, we can illustrate how crossovers are affected by certain summary statistics of a data-generating process, namely, within- and cross-treatment heterogeneity and cross-treatment correlation. Figure 5 depicts two scenarios of a three-treatment experiment. The cross-treatment heterogeneity (variation in average treatment outcomes) and the within-treatment heterogeneity (variation in individual outcomes within one condition) are the same in both scenarios, however, in the right panel this heterogeneity is actionable, while in the left panel, it is not actionable because the cross-treatment correlation matrices are different. Figure 6 shows two scenarios of a three-treatment experiment with the same cross-treatment heterogeneity and cross-treatment correlation matrices but different within-treatment heterogeneity. In the right panel, the heterogeneity is actionable for targeting, while in the left panel, it is not. Finally, Figure 7 shows two scenarios of a three-treatment experiment with the same within-treatment heterogeneity and cross-treatment correlation matrices but different cross-treatment heterogeneity. A lot of variation in the average outcomes means that it is harder to outperform the best uniform treatment, and so in the right panel, when this variation is high, the heterogeneity is not actionable for targeting.

Figure 5: The Effect of Cross-Treatment Correlation



The figure depicts two panels that were generated with the same average responses to treatments and the same level of within-treatment heterogeneity but different levels of cross-treatment correlation (more correlation in the left panel). More cross-treatment correlation results in fewer crossovers with the best uniform treatment.

3.3 A Model of the Value from Targeting

Given the intuition provided, we develop a statistical model to analyze the potential for targeting given three factors: (i) the heterogeneity in average outcomes across interventions, (ii) the heterogeneity in individual outcomes within an interventions, and (iii) the correlation of individual outcomes across interventions for the same person.

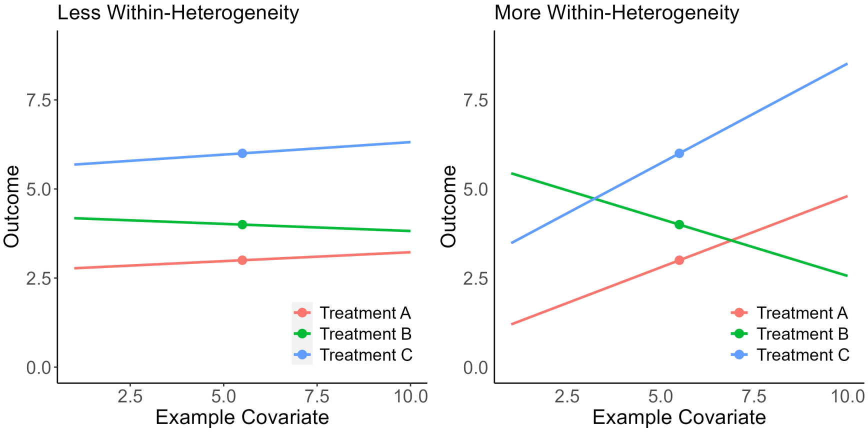
To simplify exposition, we assume that there are two arms, 0 and 1 (control and treatment). The potential outcome for person i exposed to arm $a \in \{0,1\}$ is denoted at Y_i^a. The potential outcomes are drawn from a multivariate normal distribution where the expected value of the outcomes for arm a is μ_a, the variance within each arm is σ^2, and the correlation of potential outcomes across arms for the same individual is ρ. With two arms we can write:

$$\begin{pmatrix} Y_i^0 \\ Y_i^1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \tag{10}$$

Without loss of generality, we will assume $\mu_1 > \mu_0$. In this case, the expected value from targeting is:

$$\mathbb{E}[\mathbb{I}(Y_i^1 > Y_i^0)Y_i^1 + \mathbb{I}(Y_i^1 \leq Y_i^0)Y_i^0] - \mu_1 \tag{11}$$

Figure 6: The Effect of Within-Treatment Heterogeneity



The figure depicts two panels that were generated with the same average responses to treatments and the same cross-treatment correlation matrix but different levels of within treatment heterogeneity (more heterogeneity in the right panel). More within-treatment heterogeneity results in more crossovers with the best uniform treatment.

which can be written as

$$\mathbb{E}[(1-\mathbb{I}(Y_i^1 \leq Y_i^0))Y_i^1 + \mathbb{I}(Y_i^1 \leq Y_i^0)Y_i^0] - \mu_1 \tag{12}$$

and thus simplifies to

$$\mathbb{E}[\mathbb{I}(Y_i^0 - Y_i^1 \geq 0)(Y_i^0 - Y_i^1)] \tag{13}$$

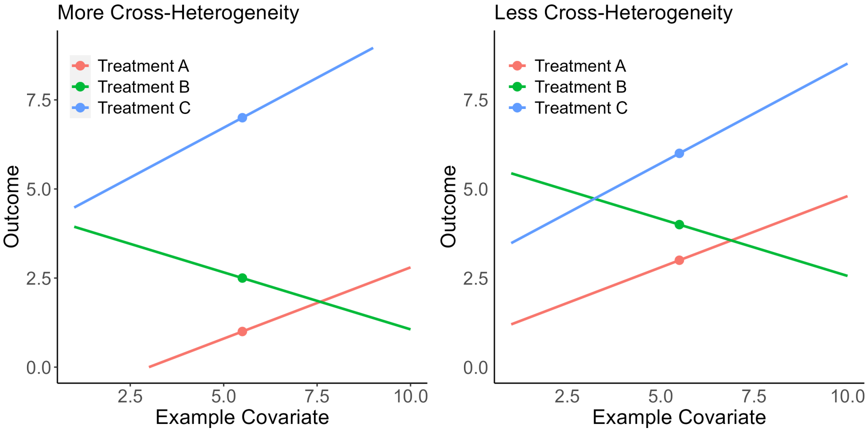
This is the expected value of a rectified normal distribution with mean $\mu_0 - \mu_1$ and variance $2\sigma^2(1-\rho)$ and is equal to:

$$(\mu_0 - \mu_1)\left[1 - \Phi\left(\frac{\mu_1 - \mu_0}{\sigma\sqrt{2(1-\rho)}}\right)\right] + \sigma\sqrt{2(1-\rho)}\phi\left(\frac{\mu_1 - \mu_0}{\sigma\sqrt{2(1-\rho)}}\right) \tag{14}$$

We compute partial derivatives of the value from targeting with respect to σ, ρ and $\mu_1 - \mu_0$ (see Appendix A.2) and find that the value from targeting is increasing in σ (within-treatment heterogeneity) and decreasing in ρ (cross-treatment correlation) as illustrated above.

Finally, if we also assume that the expected outcomes of arm a, μ_a, are drawn i.i.d from a normal distribution $\mathcal{N}(M, s^2)$, we can explore the effect of the variance of this distribution (s^2, cross-treatment heterogeneity) on the value from targeting. Let us denote the value of Equation 14 by $V(d)$, where $d = \mu_1 - \mu_0$. When computing this value, we assumed that $\mu_1 > \mu_0$, i.e., $d > 0$.

Figure 7: The Effect of Cross-Treatment Heterogeneity



The figure depicts two panels that were generated with the same average responses to treatments and the same cross-treatment correlation matrix but different levels of within treatment heterogeneity (more heterogeneity in the right panel). More within-treatment heterogeneity results in more crossovers with the best uniform treatment.

Since μ_1, μ_0 are i.i.d. draws, the expectation of the value from targeting T over this distribution can be written using the law of total expectation:

$$\mathbb{E}_{\mu_0,\mu_1}[T(\mu_0,\mu_1)] = \mathbb{E}[T(\mu_0,\mu_1)|\mu_1 \geq \mu_0] \cdot P(\mu_1 \geq \mu_0) + \mathbb{E}[T(\mu_0,\mu_1)|\mu_1 < \mu_0] \cdot P(\mu_1 < \mu_0) \quad (15)$$

By symmetry, $P(\mu_1 > \mu_0) = \frac{1}{2}$, and $T(\mu_0,\mu_1)|(\mu_1 \geq \mu_0) = T(\mu_0,\mu_1)|(\mu_1 < \mu_0)$. Therefore,

$$\mathbb{E}_{\mu_0,\mu_1}[T(\mu_0,\mu_1)] = \mathbb{E}_d[V(d)|d \geq 0], d = \mu_1 - \mu_0 \quad (16)$$

Since μ_0, μ_1 are i.i.d draws from a normal distribution $\mathcal{N}(M, s^2)$, $d \sim \mathcal{N}(0, 2s^2)$, and therefore $V(d)$ is evaluated over a half-normal distribution. Let us compare two cases: $\mu_0, \mu_1 \sim \mathcal{N}(M, s^2)$ and $\mu_0', \mu_1' \sim \mathcal{N}(M, s'^2)$ with $s' > s$. In Appendix A.2, we show that the conditional distribution of $d' = \mu_1' - \mu_0', d' \geq 0$ stochastically dominates the conditional distribution of $d = \mu_1 - \mu_0, d \geq 0$. Therefore, by the first-order stochastic dominance theorem, since $V(d)$ is decreasing in d,

$$\mathbb{E}_d[V(d)|d \geq 0] \geq \mathbb{E}_{d'}[V(d')|d' \geq 0]$$

In other words, the value of targeting is decreasing in the variance of the distribution of mean outcomes μ_1, μ_0.

In summary, the value of targeting is determined by the presence and size of crossovers with the best uniform treatment, which in turn is affected by within- and cross-treatment heterogeneity and cross-treatment correlation. Drawing upon this insight and our statistical model, we propose a moment-based simulation procedure that aims to predict and measure the potential for targeting of a given study. We find that the Penn-Geisinger study has a larger potential for targeting compared to the Walmart study, which can explain why targeting methods provide more incremental value in the former.

4 Predicting the Potential for Targeting

To extend the analysis from the two arm case and normal distribution of the average responses, in this section we numerically analyze the effects of the three forces we identified (within-treatment heterogeneity, cross-treatment heterogeneity, and cross-treatment correlation), using the same modeling approach as in Section 3.3. Algorithm 1 describes the data generating and analysis process.

Algorithm 1: Computation of the potential from targeting

```
Input: distribution of average responses $F$, within-treatment heterogeneity $\sigma$,
       cross-treatment correlation $\rho$, number of treatments $m$
From distribution $F$ draw $\mu$ — an $m$-vector of average responses for each treatment
Construct a $m \times m$ individual responses covariance matrix $\Sigma$ with $\sigma^2$ on the diagonal and
  $\rho\sigma^2$ off-diagonal
Draw potential outcomes $Y_i(a)$ from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$
The value of the optimal targeting policy corresponds to the maximum of potential
  outcomes for each individual averaged over individuals: $V_t = \frac{1}{n}\sum_{i=1}^{n} \max_a Y_i^a$
The value of the best uniform policy corresponds to the maximum of average responses
  $V_u = \max_a \frac{1}{n}\sum_{i=1}^{n} Y_i^a$
The potential from targeting is the difference between the two: $V_t - V_u$
```

4.1 When Adding More Treatments Can Hurt Personalization?

We are particularly interested to understand if there are situations when more treatments in an experiment lead to a reduction in the value from targeting. Specifically, for peaked distributions of

μ_a we expect a tradeoff. On one hand, more arms are close to one another in average responses. On the other hand, the best uniform intervention might be more of an outlier and harder to "beat". Interestingly, the role of heavy-tailed distributions in A/B tests was also noted in Azevedo et al. (2020). However, in our case the mechanism that affects the results is different — a heavier-tailed distribution affects the benchmark that the targeting needs to "beat".

To explore this tradeoff, we generate values for the average potential outcomes per arm μ_a from the following distribution $F(\cdot)$:

$$\mu_a \sim \begin{cases} M & \text{with prob. } \pi \\ \mathcal{N}(M, s) & \text{with prob. } 1-\pi \end{cases} \tag{17}$$

This distribution is a spike-and-slab mixture, where the spike provides value M, and the slab is drawn from a normal distribution centered around M. [4] The variance of the resulting distribution is $(1-\pi)s^2$, and in our analysis we hold this value constant while changing π for ease of comparison. We analyze four cases:

(a) $\pi = 0$ (normal), $(1-\pi)s^2 = 10$ (low variance);

(b) $\pi = 0$ (normal), $(1-\pi)s^2 = 50$ (high variance);

(c) $\pi = 0.9$ (spike-and-slab), $(1-\pi)s^2 = 10$ (low variance);

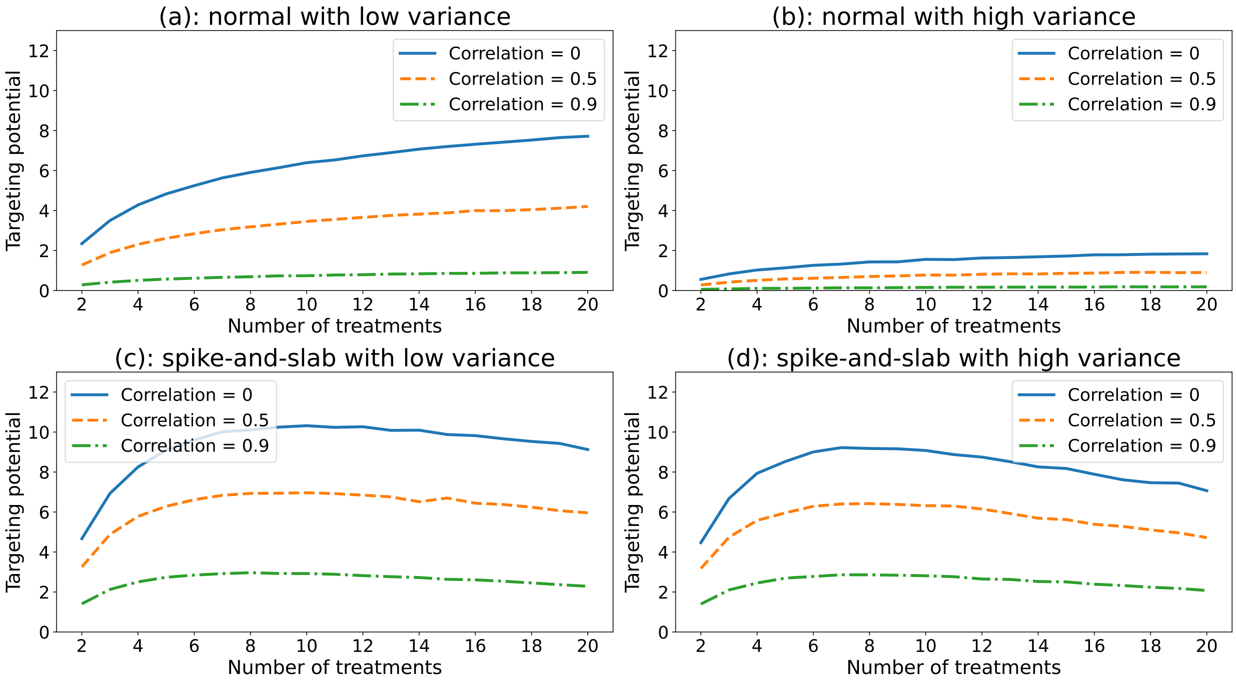
(d) $\pi = 0.9$ (spike-and-slab), $(1-\pi)s^2 = 50$ (high variance).

Figure 8 shows the results of the numeric computation. We normalize the within-treatment heterogeneity (σ) to 10 and vary the cross-treatment correlations $\rho = \{0, 0.5, 0.9\}$. The four subplots show the targeting potential as a function of the number of treatments for the cases described above.

Figure 8(a) shows that the results from the two-arm model extend to more arms — higher correlation across potential outcomes of arms is detrimental to targeting, but more arms generate potential for better outcomes from targeting. In comparison, figure 8(b) illustrates that when the variance of the average potential outcomes s is higher, the targeting potential suffers dramatically.

[4] The value of M does not affect the analysis, as it shifts both the targeting policy and the uniform benchmark by the same value

Figure 8: The Effect of Data Moments on Targeting Potential



The figure shows the simulated effects of the distribution of average responses to treatments on targeting potential for different levels of cross-treatment correlation. Within-treatment heterogeneity (σ) is normalized to 10 in all plots.

This is because the value of the best uniform benchmark is more likely to be higher, and thus harder to beat, lowering the benefit from personalizing treatments.

Continuing to Figure 8(c), we suddenly see that increasing the number of arms can hurt the value from targeting, and this effect is even more pronounced in Figure 8(d). In these cases, the spike-and-slab distribution of the average potential outcomes causes two effects, leading to an inverse U-shape. First, when the number of arms is large, there is a high probability that at least one of the interventions will be drawn from the slab component, making its value potentially high and hard to beat, thereby lowering the targeting potential. However, when the number of arms is small, there is a high probability that all arms are drawn from the spike component, leading to low cross-treatment heterogeneity and thus increasing the targeting potential. As a result, the value of personalization can be higher for a smaller number of arms compared to a higher number of arms. This effect is particularly noticeable if we compare Figures 8(b) and 8(d). For 8(d), while the number of arms is small, all of them are likely to be from the spike, and the cross-treatment heterogeneity is very low, so 8(d) looks very similar to 8(c) at first. However, as the number of arms increases, the underlying slab component with a high variance comes into play, the cross-treatment heterogeneity increases, and the targeting potential deteriorates. In contrast, in 8(b), the effect of high cross-treatment heterogeneity is uniformly applied to any number of arms, and so the targeting potential is low from the beginning and is slowly increasing with the number of arms.

This analysis shows that sometimes having more treatments can hurt the value of personalization even under perfect knowledge of potential outcomes (i.e., with no finite-sample restrictions). Of course, in reality, samples are finite, and having more treatments might dilute per-treatment sample sizes, potentially resulting in poorer performance of personalized policies due to large estimation errors.

4.2 Potential from Targeting: Applications

In addition to deriving insights into how various factors affect the value of personalization, Algorithm 1 can be used to gauge the amount of maximum targeting potential by researchers before they run an experiment. If an analyst has priors (either from past studies or a pilot) on values of within- and cross- treatment heterogeneity and cross-treatment correlation, these values can be imputed in Algorithm 1 to estimate the expected benefit of personalization.[5] However, Algorithm 1 assumes direct access to potential outcomes for every individual under all possible treatment assignments. Thus, it provides an accurate estimate of the value of personalization only when these potential outcomes are either observed or estimated very precisely. In other cases, the result of this algorithm presents an upper bound.

In reality, potential outcomes for all possible treatment assignments are rarely observed, and usually there is significant noise in predictions. For these cases, we developed a sensitivity analysis that explains how the targeting potential will change with added prediction noise, and calibrated the analysis to the noise estimates from our data. Algorithm 2 adds noise to the true potential outcomes, and uses the noisy predictions for arm assignment during personalization. However, as in our empirical application, the value from this assignment is estimated using the true potential

[5] For proper Bayesian priors, Algorithm 1 can be applied to samples from prior distributions and thus generate not only a point estimate but also a credible interval.

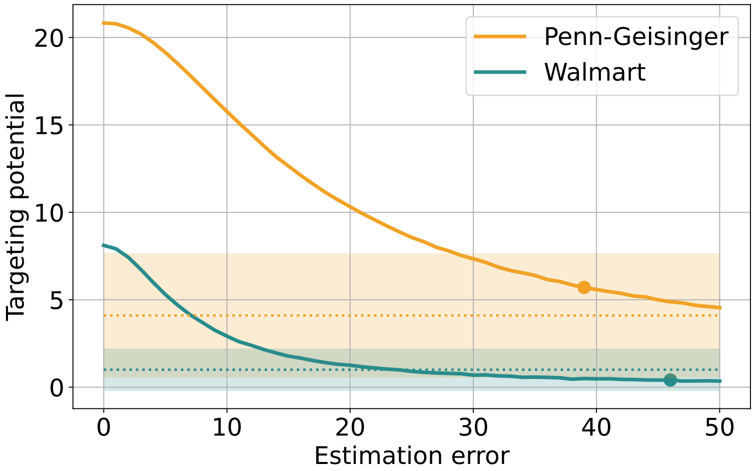
outcomes similarly to IPW.

```
Algorithm 2: Computation of the potential from targeting under prediction error
Input: distribution of average responses $F$, within-treatment heterogeneity $\sigma$,
       cross-treatment correlation $\rho$, number of treatments $m$, standard deviation of
       prediction errors $\sigma_\varepsilon$
From distribution $F$ draw $\mu$ — an $m$-vector of average responses for each treatment
Construct a $m \times m$ individual responses covariance matrix $\Sigma$ with $\sigma^2$ on the diagonal and
  $\rho\sigma^2$ off-diagonal
Draw potential outcomes $Y_i^a$ from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$
Estimated potential outcomes $\hat{Y}_i^a$ are equal to true potential outcomes $Y_i(a)$ plus noise
  coming from estimation: $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$
The value of the optimal targeting policy corresponds to the argmax of estimated potential
  outcomes $\hat{Y}_i^a$ for each individual evaluated on true potential outcomes $Y_i^a$ and averaged
  over individuals: $V_t = \frac{1}{n}\sum_{i=1}^n Y_i^{a^*(i)}, a^*(i) = \arg\max_a \hat{Y}_i^a$
The value of the best uniform policy corresponds to the argmax of estimated average
  responses evaluated on true potential outcomes
  $V_u = \frac{1}{n}\sum_{i=1}^n Y_i^{a^*}, a^* = \arg\max_a \frac{1}{n}\sum_{i=1}^n \hat{Y}_i^a$
The potential from targeting is the difference between the two: $V_t - V_u$
```

We applied Algorithm 2 to our data using the moments estimated using T-Learner XGBoost (Table 4) for the Penn-Geisinger and the Walmart experiments. Figure 9 presents the results. When the estimation error is zero on the x-axis, we receive the same outcome as in the previous analysis (Algorithm 1). When we increase the estimation error (σ_ε in Algorithm 2), the value of personalization goes down as expected. The dots indicate the numerical predicted value of targeting using Algorithm 2 with the estimated values of noise from our XGBoost analysis. The horizontal dotted lines indicate the estimated value from targeting (Section 2, Table 3). As we can see, the values are relatively consistent, providing credibility for Algorithm 2 as a tool for computing the potential from targeting using summary statistics of within- and cross-treatment heterogeneity and cross-treatment correlation. [6]

[6] Our Algorithm 2 complements the RATE algorithm proposed by Yadlowsky et al. (2021). RATE AUTOC focuses on comparing and evaluating targeting *policies*, while our algorithm aims to estimate the quality of *data* for targeting, staying agnostic of a specific policy.

Figure 9: Simulated Targeting Potential for Two Experiments



The figure shows the estimated targeting potential for different amounts of estimation noise (standard deviation, in percentage points) for the Penn-Geisinger and the Walmart experiments. The dot depicts the prediction error of T-XGBoost out-of-sample (the same method used to calculate moments) and the corresponding predicted targeting potential. The dotted lines and the shaded area around them correspond to the IPW estimates of the value from targeting in Table 3 of Section 2.

5 Conclusion

In this paper, we evaluated five popular targeting methods on two large-scale field experiments and found that in one study, targeting achieves a relative improvement of 3% over the top-performing intervention, and in the other the relative improvement is a more substantial 13%. When we quantified the value of within-treatment heterogeneity and cross-treatment correlation in the two datasets, we found that they suggest different directions regarding which dataset might have a higher potential for targeting. We showed that neither of these measures on its own is sufficient to explain the differences in targeting value, and instead a compound measure of three forces — the magnitude of crossovers with the best uniform treatment — captures the potential for targeting.

To provide a theoretical explanation for the observed difference in the value from personalization, we developed a statistical model of targeting potential. We found that when within-treatment heterogeneity increases, we expect a higher value from targeting because the benefit of switching to a different treatment has a higher variance and therefore might be higher. However, when potential outcomes for an individual are correlated across treatments, the benefit of switching gets eroded. A third force that also lowers the value from targeting is high variance in average outcomes because every targeting policy needs to beat the best uniform, the value of which increases with higher variance. Under some conditions, this force can countervail the effect of adding more treatments, resulting in an inverse U-shape relationship between the number of treatments and the value from targeting.

Finally, when we calibrated this model to the moments we estimated from both megastudies and took into account the prediction error, we found that the predictions from the model are consistent with our empirical results from Section 2.

Future work may address questions beyond the scope of this paper, such as variable selection. Heterogeneity is inherently linked to the covariates available to a researcher. As emphasized by Rossi et al. (1996), targeting can only be as good as the covariates. For the same experiment, collecting one set of covariates may generate a lot of actionable heterogeneity, which would enable targeting — while another set of covariates may not exhibit any heterogeneity at all. By applying the model to different experiments and different sets of covariates (e.g., Smith et al., 2023), we might be able to get insights into which covariates tend to offer heterogeneity that is useful for targeting and make recommendations on which variables to collect. Future research may also use the model to understand which contexts tend to have higher targeting potentials and possibly extend the setting to adaptive personalization.

Our paper has several implications. For practitioners, it provides a fast and easy-to-implement tool to determine the amount of actionable heterogeneity in the data either before or after running an experiment — and decide whether it is worth exploring different targeting policies or if there is an opportunity to improve the existing ones. From a substantive perspective, it sheds light on the determinants of the value of targeting and highlights the futility of expecting a certain benefit from personalization just because an experiment has many interventions. Our case study comparing the Walmart and the Penn-Geisinger field experiments illustrates the unpredictable nature of the value from personalization. The two studies have remarkably similar contexts, and yet the gains from personalization in one are four times higher than in the other. Our research uncovered that in that case, the difference is mostly driven by the levels of within-treatment heterogeneity. Finally, our paper provides an interesting perspective on analysis of heterogeneity. As it turns out, not all kinds of heterogeneity are actionable for targeting, and personalization opportunities might be limited even when outcomes are heterogeneous.

References

Acemoglu, D., Chernozhukov, V., Werning, I., and Whinston, M. D. (2021). Optimal targeted lockdowns in a multigroup SIR model. *American Economic Review: Insights*, 3(4):487–502.

Ahani, N., Andersson, T., Martinello, A., Teytelboym, A., and Trapp, A. C. (2021). Placement optimization in refugee resettlement. *Operations Research*, 69(5):1468–1486.

Andrews, I., Kitagawa, T., and McCloskey, A. (2024). Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of marketing Research*, 55(1):80–98.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Keleher, N., and Spiess, J. (2023). Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal. *arXiv preprint arXiv:2310.08672*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., and Weyl, E. G. (2020). A/B testing with fat tails. *Journal of Political Economy*, 128(12):4614–000.

Bauman, K. and Tuzhilin, A. (2018). Recommending remedial learning materials to students by filling their knowledge gaps. *MIS quarterly : management information systems.*, 42(1).

Blake, T., Nosko, C., and Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174.

Bonander, C. and Svensson, M. (2021). Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Economics*, 30(8):1818–1832.

Burtch, G., Ghose, A., and Wattal, S. (2015). The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Science*, 61(5):949–962.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, Y., Lee, J.-Y., Sridhar, S., Mittal, V., McCallister, K., and Singal, A. G. (2020). Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment. *Journal of Marketing*, 84(3):1–27.

Chung, T.-H., Rostami, V., Bastani, H., and Bastani, O. (2022). Decision-aware learning for optimizing health supply chains. *arXiv preprint arXiv:2211.08507*.

Davis, J. M. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550.

Dubé, J.-P. and Misra, S. (2023). Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189.

Elmachtoub, A. N. and Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, 68(1):9–26.

Feit, E. M. and Berman, R. (2019). Test & roll: Profit-maximizing A/B tests. *Marketing Science*, 38(6):1038–1058.

Fong, J. and Hunter, M. (2022). Can facing the truth improve outcomes? effects of information in consumer finance. *Marketing Science*, 41(1):33–50.

Ghosh, S., Kim, R., Chhabria, P., Dwivedi, R., Klasnja, P., Liao, P., Zhang, K., and Murphy, S. (2024). Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling. *Machine Learning*, pages 1–37.

Goldenberg, D., Kofman, K., Albert, J., Mizrachi, S., Horowitz, A., and Teinemaa, I. (2021). Personalization in practice: Methods and applications. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1123–1126.

Graham, B. S., Ridder, G., Thiemann, P., and Zamarro, G. (2022). Teacher-to-classroom assignment and student achievement. *Journal of Business & Economic Statistics*, pages 1–27.

Hitsch, G. J., Misra, S., and Zhang, W. (2023). Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*.

Hu, A. (2023). Heterogeneous treatment effects analysis for social scientists: A review. *Social Science Research*, 109:102810.

Jamieson, K. G. and Jain, L. (2018). A bandit approach to sequential experimental design with false discovery control. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525.

Knittel, C. R. and Stolper, S. (2019). Using machine learning to target treatment: The case of household energy use. Technical report, National Bureau of Economic Research.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Luo, X., Lu, X., and Li, J. (2019). When and how to leverage e-commerce cart targeting: The relative and moderated effects of scarcity and price incentives with a two-stage field experiment and causal forest optimization. *Information Systems Research*, 30(4):1203–1227.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., et al. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6):e2115126119.

Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., et al. (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20):e2101165118.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Patel, M. S., Milkman, K. L., Gandhi, L., Graci, H. N., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Akinola, M., et al. (2023). A randomized trial of behavioral nudges delivered through text messages to increase influenza vaccination among patients with an upcoming primary care visit. *American Journal of Health Promotion*, 37(3):324–332.

Perdomo, J. C., Britton, T., Hardt, M., and Abebe, R. (2023). Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205.*

Rafieian, O. (2023). Optimizing user engagement through adaptive ad sequencing. *Marketing Science*, 42(5):910–933.

Rafieian, O. and Yoganarasimhan, H. (2021). Targeting and privacy in mobile advertising. *Marketing Science*, 40(2):193–218.

Rafieian, O. and Yoganarasimhan, H. (2023). AI and personalization. *Artificial Intelligence in Marketing*, pages 77–102.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340.

Schmit, S. and Johari, R. (2018). Learning with abandonment. In *International Conference on Machine Learning*, pages 4509–4517. PMLR.

Simester, D., Timoshenko, A., and Zoumpoulis, S. I. (2020). Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science*, 66(8):3412–3424.

Smith, A. N., Seiler, S., and Aggarwal, I. (2023). Optimal price targeting. *Marketing Science*, 42(3):476–499.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966.*

Ye, Z., Zhang, Z., Zhang, D., Zhang, H., and Zhang, R. P. (2023). Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Available at SSRN 4375327.*

Yoganarasimhan, H., Barzegary, E., and Pani, A. (2023). Design and evaluation of optimal free trials. *Management Science*, 69(6):3220–3240.

Zhang, D. J., Dai, H., Dong, L., Qi, F., Zhang, N., Liu, X., Liu, Z., and Yang, J. (2020). The long-term and spillover effects of price promotions on retailing platforms: Evidence from a large randomized experiment on Alibaba. *Management Science*, 66(6):2589–2609.

Zhang, W. W. (2023). Optimal comprehensible targeting.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

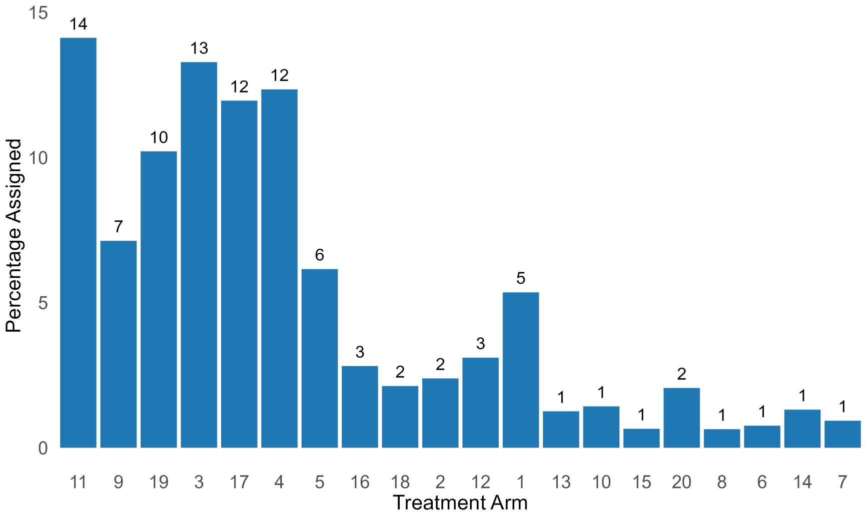
Zhou, Z., Athey, S., and Wager, S. (2023). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183.

A Supplementary Materials

A.1 Description of the Best Performing Targeting Policy

Figure 10 illustrates the treatment assignment according to the best targeting policy (multi-arm causal forest) in the Penn-Geisinger study. The interventions are ranked based on the average vaccination rates in the training sample (same as in Figure 2). The graph indicates that all twenty interventions are assigned to non-empty subpopulations, with four treatments encompassing over 50% of the total population. Although these four interventions overlap with the set of four best-performing uniform treatments, these sets do not align exactly. The multi-arm causal forest is a black box and non-interpretable model, making it challenging to summarize the groups of people assigned to each arm. However, insights into important covariates for targeting can be gained by counting the number of splits involving a particular variable across all trees in the causal forest. The top five covariates, based on this criterion in descending frequency, are zipcode-level median income, BMI, weight, age, and the date of the intervention.

Figure 10: Arm Assignment; Best Targeting Policy in Penn-Geisinger



Percentages of population assigned to each intervention under the best targeting policy for Penn-Geisinger dataset (multi-arm causal forest). The interventions are ranked with respect to the decreasing average outcome levels in the training sample (same as Figure 2).

A.2 Effects of Three Forces on Targeting Potential Derivation

A.2.1 Within-treatment heterogeneity

We will compute the derivative of Equation 14 with respect to σ. Let us denote $d = \mu_1 - \mu_0$ (we assume $d \geq 0$). The partial derivative is equal to:

$$-\frac{d^2}{\sigma^2\sqrt{2(1-\rho)}}\phi\left(\frac{d}{\sigma\sqrt{2(1-\rho)}}\right) + \sqrt{2(1-\rho)}\phi\left(\frac{d}{\sigma\sqrt{2(1-\rho)}}\right) + \frac{d^2}{\sigma^2\sqrt{2(1-\rho)}}\phi\left(\frac{d}{\sigma\sqrt{2(1-\rho)}}\right)$$
$$= \sqrt{2(1-\rho)}\phi\left(\frac{d}{\sigma\sqrt{2(1-\rho)}}\right) > 0$$

That is, the value of targeting is increasing in σ.

A.2.2 Cross-treatment correlation

For simplicity, we will first replace $t = \sqrt{2(1-\rho)}$, t is decreasing with ρ.

Equation 14 becomes:

$$-d\left[1-\Phi\left(\frac{d}{\sigma t}\right)\right] + \sigma t \phi\left(\frac{d}{\sigma t}\right)$$

Taking a partial derivative with respect to t,

$$-\frac{d^2}{\sigma t^2}\phi\left(\frac{d}{\sigma t}\right) + \sigma\phi\left(\frac{d}{\sigma t}\right) + \frac{d^2}{\sigma t^2}\phi\left(\frac{d}{\sigma t}\right) = \sigma\phi\left(\frac{d}{\sigma t}\right) > 0$$

Since t is decreasing in ρ, the value of targeting is also decreasing in ρ.

A.2.3 Difference in means

We will now take a partial derivative with respect to $d = \mu_1 - \mu_0$. For simplicity, we will let $v = \sigma\sqrt{2(1-\rho)}$.

Equation 14 becomes:

$$-d\left[1-\Phi\left(\frac{d}{v}\right)\right] + v\phi\left(\frac{d}{v}\right)$$

Taking a partial derivative with respect to d,

$$-1 + \Phi\left(\frac{d}{v}\right) + \frac{d}{v}\phi\left(\frac{d}{v}\right) - \frac{d}{v}\phi\left(\frac{d}{v}\right) = \Phi\left(\frac{d}{v}\right) - 1 < 0$$

Therefore, the targeting value is decreasing in $\mu_1 - \mu_0$.

A.2.4 Half-normal stochastic dominance

Here we will show that if $\mu_1, \mu_0 \sim N(M, s^2), \mu_1', \mu_0' \sim N(M, s'^2)$ then the conditional distribution of $d' = \mu_1' - \mu_0', d' \geq 0$ stochastically dominates the conditional distribution of $d = \mu_1 - \mu_0, d \geq 0$ if $s' > s$.

One distribution stochastically dominates another if:

$$F_{d'}(x) \leq F_d(x) \text{ for all } x$$

with a strict inequality for at least one x.

For $x > 0$, the two cdfs are given by:

$$\operatorname{erf}\left(\frac{x}{(\sqrt{2}s')\sqrt{2}}\right) < \operatorname{erf}\left(\frac{x}{(\sqrt{2}s)\sqrt{2}}\right)$$

and the inequality holds because $s' > s$ and erf is an increasing function.