

When Is Heterogeneity Actionable for Personalization?*

Anya Shchetkina

Ron Berman

The Wharton School

The Wharton School

University of Pennsylvania

University of Pennsylvania

November 22, 2024

*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). Ron Berman acknowledges support from the Wharton Dean's Research Fund, The Mack Institute for Innovation Management, Wharton's AI & Analytics Initiative, 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, Zhenling Jiang, Katherine Milkman, Christophe Van den Bulte, Walter Zhang, Oded Netzer and 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.

When Is Heterogeneity Actionable for Personalization?

Abstract

Targeting and personalization policies can be used to improve outcomes beyond the uniform policy that assigns the best performing treatment in an A/B test to everyone. Personalization relies on the presence of heterogeneity of treatment effects, yet, as we show in this paper, heterogeneity alone is not sufficient for personalization to be successful. We develop a statistical model to quantify “actionable heterogeneity,” or the conditions when personalization is likely to outperform the best uniform policy. We show that actionable heterogeneity can be visualized as crossover interactions in outcomes across treatments and depends on three population-level parameters: within-treatment heterogeneity, cross-treatment correlation, and the variation in average responses. Our model can be used to predict the expected gain from personalization prior to running an experiment and also allows for sensitivity analysis, providing guidance on how changing treatments can affect the personalization gain. To validate our model, we apply five common personalization approaches to two large-scale field experiments with many interventions that encouraged flu vaccination. We find an 18% gain from personalization in one and a more modest 4% gain in the other, which is consistent with our model. Counterfactual analysis shows that this difference in the gains from personalization is driven by a drastic difference in within-treatment heterogeneity. However, reducing cross-treatment correlation holds a larger potential to further increase personalization gains. Our findings provide a framework for assessing the potential from personalization and offer practical recommendations for improving gains from targeting in multi-intervention settings.

Keywords: heterogeneous treatment effects, targeting, personalization, A/B tests, experimental design, machine learning.

1 Introduction

A/B tests (online randomized control trials) are a popular strategy to find effective interventions in business applications including marketing (e.g., Sahni et al., 2018; Fong and Hunter, 2022; Goli et al., 2024a), economics (e.g., Blake et al., 2015; Simonov et al., 2018; Mosleh et al., 2024), operations (e.g., Zhang et al., 2020; Boutilier et al., 2024), information systems (e.g., Burtch et al., 2015; Bauman and Tuzhilin, 2018), and data science (e.g., Johari et al., 2017; Ye et al., 2023). 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 targeted and personalized 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 personalization.

There are multiple ways one can approach a personalization task. One way is to use heuristics, such as targeting high-risk individuals. However, Ascarza (2018) and Athey et al. (2023) caution against relying on popular heuristics because empirical results show little efficacy of this approach. Instead they advise to explicitly model heterogeneity. Typically, such modeling involves three stages: training, prediction, and optimization.² This predict-then-optimize approach can be

¹Throughout 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.

²Similar stages are involved in other predict-then-optimize applications of machine learning tools to real-world decisions. Although common, this indirect approach can sometimes be improved by direct optimization (see, e.g., Elmachoub and Grigas, 2022; Chung et al., 2022).

applied on the outcomes from each treatment, or directly on treatment effects. (1) Initially, a flexible machine learning model is trained on experimental data linking each treatment arm and individual covariates to observed outcomes or treatment effects (compared to a control). (2) In the prediction stage, the trained model forecasts counterfactual outcomes or treatment effects for each individual under all possible treatment assignments. (3) The treatment with the highest predicted outcome or treatment effect for that individual becomes the recommended personalization policy. For predicting outcomes, general purpose machine learning algorithms such as XGBoost (Chen and Guestrin, 2016) are commonly used. Simester et al. (2020b) provides a comparison of different algorithms in this context under common data challenges. For treatment effects, the causal tree method (Athey and Imbens, 2016), and its generalization the causal forest (Wager and Athey, 2018), are now extensively used for targeting (see, e.g., Davis and Heller, 2017; Luo et al., 2019; Bonander and Svensson, 2021). Similarly, Hitsch et al. (2024) develop a non-parametric approach to estimate HTEs that can be used for personalization.

Another approach is to combine all three stages and directly optimize the personalization 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 compared to deploying the best uniform policy. A few examples reporting effective targeting policies include 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), pricing in B2B settings (Zhang et al., 2014), ad loads on streaming platforms (Goli et al., 2024b), 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. This inconsistency raises a question: what makes personalization effective in some cases and not effective in others?

The goal of our paper is to answer this question. To address it, we develop a statistical model to quantify when and how much heterogeneity in treatment effects is “actionable” for personalization. We show that heterogeneity alone is insufficient: actionable heterogeneity requires that the most effective intervention differs from one subgroup to another. Visually, this appears as a crossover interaction between interventions if the individuals are ranked by their treatment effects. We further decompose the presence and magnitude of such crossovers into the effects of several parameters of the data generating process: (1) within-treatment heterogeneity (the variation of individual responses for the same intervention), (2) cross-treatment correlation (how similar are responses to different interventions for the same individual), and (3) variation in average outcomes across interventions. Factor (1) is the classic dimension most researchers and experimenters have been focusing on to increase gains from personalization. Factor (3) is more influential when there are more than two treatments, and has been considered in the literature on A/B testing (Feit and Berman, 2019; Azevedo et al., 2020; Berman and Van den Bulte, 2022). One of our contributions in this paper is to carefully consider factor (2), the correlation between potential outcomes within individuals, which has a dramatic impact on the gains from personalization. We also develop a method to estimate this correlation.

We describe how this model can be used to gauge the potential gain from personalization *before* running an experiment if a researcher has prior expectations about the levels of within-treatment

correlation, cross-treatment heterogeneity and variation in average outcomes. We find that, surprisingly, sometimes having more interventions can hurt the potential gain from personalization even in the absence of finite-sample effects. 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 to the right direction: whether to search for more precise estimation methods, to experiment with other interventions, or to collect more data.

To illustrate the concept of actionable heterogeneity and the applicability of our model, we analyze two large-scale field experiments with many interventions: the Walmart flu shots study of Milkman et al. (2022) and the Penn-Geisinger flu shots study of Milkman et al. (2021) and Patel et al. (2023). In these studies, 22 and 19 behavioral nudges informed by psychological theory were tested concurrently to encourage flu vaccination.³ We benchmark our model’s prediction against several popular personalization approaches and find a relatively small gain from personalization (4% relative improvement over the best uniformly applied treatment) in the Walmart study and a more substantial gain from personalization (18% relative improvement) in the Penn-Geisinger study. The empirical estimates are consistent with the predictions from our model. We then use our model to run a counterfactual analysis and find that the difference in gains from personalization between the two studies is mostly driven by a drastic difference in within-treatment heterogeneity. This difference can be attributed to a large degree to the availability of additional relevant covariates in the Penn-Geisinger study and also to a smaller degree to differences in sample composition. However, if a marketer could improve one of the parameters of the studies by 1%, we find that improving within-treatment heterogeneity does not bring the highest return in terms of an additional gain from personalization. Instead, a reduction in cross-treatment correlation would be the most valuable for both studies.

To summarize, this paper offers three contributions. Our first contribution is to show that het-

³Throughout our analysis, we focus on this one-shot setting. There is also literature on personalization in online adaptive experiments, which some of our insights might apply to (see, e.g., Schmit and Johari, 2018; Liao et al., 2020; Goldenberg et al., 2021; Rafeian, 2023; Ghosh et al., 2024). In addition, past literature has shown that data from multiple experiments can be combined to improve on targeting policies (Timoshenko et al., 2020; Huang and Ascarza, 2024; Yang et al., 2024).

erogeneity alone is not a sufficient indicator of the potential gain from personalization. Instead, the magnitude of crossovers with the best-performing intervention determines this gain, and the magnitude of crossovers is in turn influenced by three population-level parameters: within-treatment heterogeneity, cross-treatment correlation, and variation in average outcomes. We demonstrate the effects of these forces on the value of personalization and investigate the effects of having many treatments in an experiment. One of our interesting findings is that more treatments might lead to lower gains from personalization.

Second, from a methodological perspective, we develop a statistical model of gain from personalization that can be applied to gauge the personalization potential prior to running an experiment, compare the personalization potential across different studies, and find areas to improve personalization potential for future experiments. We also develop a method to estimate cross-treatment correlation, the new important factor that affects this gain.

Third, we also estimated the value of personalization in two large-scale field experiments with many interventions using state-of-the-art machine learning methods. We found that our model is consistent with the predictions of these methods, and we are able to provide insights into why personalization has a larger gain in one experiment, and smaller in another.

The next section provides a motivating example for a personalization exercise, introduces notation and the concept of actionable heterogeneity. Section 3 then introduces the formal model and illustrates the effects of the three factors. Section 4 analyzes the effects of many treatments, and Sections 5 and 6 illustrate our method and findings using two large scale field experiments.

2 Motivating Example

A company ran an A/B test sending two versions of text messages (A and B) to customers and measured the spending of each customer afterwards. For example, message A could be “sale ends today”, and message B “there are only a few items left in stock”. The customers were assigned at random to receiving one of the messages such that n_A customers received A and n_B customers received B. We denote the message (treatment) customer i received as T_i , e.g., customer i assigned message A will have $T_i = A$. The outcome (spending) is denoted as Y_i .

Each customer has two potential outcomes (Y_i^A, Y_i^B) , reflecting the outcomes of receiving message A or B, respectively. In the data only one of these outcomes is observed depending on T_i . Each customer i is also characterized by covariates x_i , such as their age, gender, and income.

The goal of the company is to learn about the most profitable way to send messages to future customers. They have three possible strategies they can choose from going forward: (i) send A to everyone, (ii) send B to everyone, and (iii) personalize such that some customers receive A and some receive B. For this simplified example, we assume that the cost of sending each message is the same (and small) regardless of its content. We also assume that not sending any message is less profitable than any of the three options above.

To decide on the optimal strategy, the company evaluates the following metrics:

1. The average response to message A: $\bar{Y}^A = \frac{1}{n_A} \sum_{i: T_i=A} Y_i^A$,
2. The average response to message B: $\bar{Y}^B = \frac{1}{n_B} \sum_{i: T_i=B} Y_i^B$,
3. Heterogeneous treatment effects for each x :

$$\hat{\tau}(x) = \frac{1}{n_B(x)} \sum_{i: T_i=B, x_i=x} Y_i^B - \frac{1}{n_A(x)} \sum_{i: T_i=A, x_i=x} Y_i^A,$$

where $n_B(x)$ is the number of individuals with covariates $x_i = x$ assigned to treatment B, and similarly for $n_A(x)$.

To implement strategy (iii) from above, the company can use the heterogeneous treatment effects and design a *personalization policy* $\pi(x)$ that assigns to each individual a treatment based on their covariates:

$$\pi(x) = \begin{cases} A & \text{if } \hat{\tau}(x) \leq 0 \\ B & \text{if } \hat{\tau}(x) > 0 \end{cases} \quad (1)$$

Because there are only two treatments A and B, this policy assigns to each person the treatment that yields the highest predicted sales, which is the optimal personalization policy in expectation.

If this personalization policy is deployed to the experimental sample, it is expected to yield an average response of:

$$\bar{Y}^\pi = \frac{1}{n_A + n_B} \sum_i [Y_i^A \mathbb{I}(\pi(x_i) = A) + Y_i^B \mathbb{I}(\pi(x_i) = B)] \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

A personalization policy is expected to outperform a uniform strategy if there is sufficient variation in individual treatment effects across the population. However, such variation is not always sufficient to guarantee a gain over assigning a uniform treatment as the following example shows.

Table 1 and Figure 1 present results for the same set of two hypothetical experiments.⁴ The table presents the average responses from applying the three policies, and the figure presents a common style of plots used to assess the distribution of treatment effects in each experiment. The best uniform policy (without personalization) is to send message B to everyone. Both experiments exhibit heterogeneity with respect to the covariate x , which in this example is the income percentile of individuals. An analysis of these data would show an interaction between the treatment and the covariate, indicating that there is heterogeneity in treatment effects. However, only in the first experiment this heterogeneity is “actionable” for personalization. In Experiment 1 a personalized policy provides a gain of 0.4 over sending message B to everyone, while in Experiment 2 a personalized policy provides no benefit over sending B to everyone.

Table 1: Results of Experiments 1 and 2

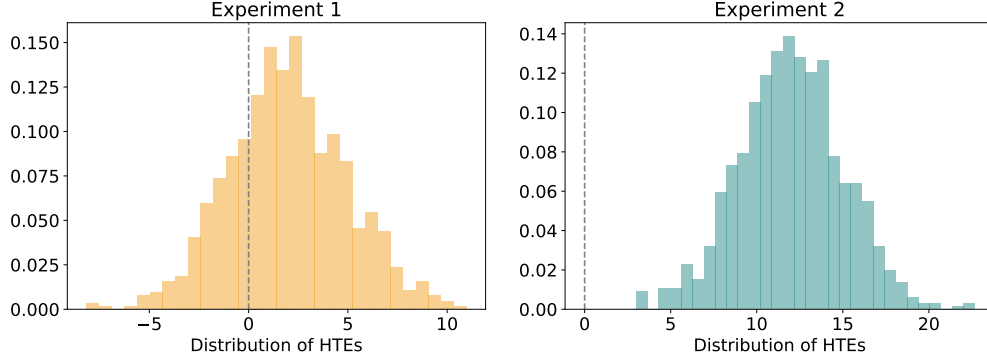
	\bar{Y}^A	\bar{Y}^B	\bar{Y}^π
Experiment 1	24.5	26.5	26.9
Experiment 2	19.5	31.5	31.5

Average responses using the three strategies in two hypothetical experiments.

The reason is that in experiment 2, message B’s average response is so much higher than message A’s, that the amount of variation in individual treatment effects relative to this difference is not high

⁴In experiment 1, $Y_i^A = 22 + 0.5x_i$, $Y_i^B = 34 - 1.5x_i$. In experiment 2, $Y_i^A = 17 + 0.5x_i$, $Y_i^B = 39 - 1.5x_i$. In both experiments, $x_i \sim N(5, 1.5)$.

Figure 1: Heterogeneous Treatment Effects in Experiments 1 and 2

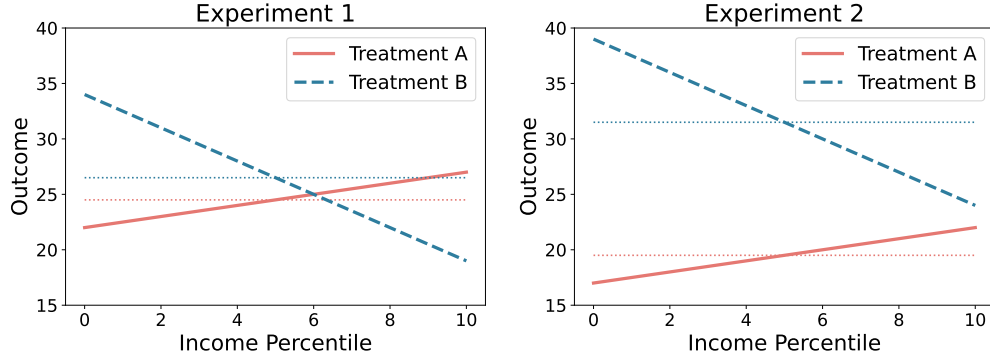


The figure shows hypothetical distributions of HTEs $\tau(x) = Y^B(x) - Y^A(x)$ in Experiments 1 and 2. The vertical dashed line denotes $\tau(x) = 0$. Experiment 1 exhibits a gain from personalization while experiment 2 does not.

enough to provide additional gains. In experiment 1, the average responses are close, making the individual variation relatively larger, and this provides opportunities for gain from personalization.

A useful approach to illustrate why some types of heterogeneity are actionable while others are not is to disentangle the HTE distribution plots of Figure 1 in terms of the outcomes for each treatment. Figure 2 presents such outcome illustrations for both experiments. In each panel the red solid line represents the outcomes under treatment A, and the blue dashed line under treatment B, for each value of the covariate x .⁵ The dotted lines indicate the average outcomes.

Figure 2: Actionable vs. non-Actionable Interactions



The plot displays two A/B tests in which the treatment effect is moderated by a covariate. In the left panel, heterogeneity is actionable and personalization is effective, while in the right panel they are not.

In experiment 1 (left panel), the personalization policy that allocates B to individuals with lower

⁵When there are multiple covariates, this plot becomes multi-dimensional but the same principle applies.

income and A to those with higher income outperforms both uniform policies. In comparison, in experiment 2 (right panel), despite the covariate moderating treatment B’s effectiveness, there is no “crossover” of the outcomes: treatment B performs better than A for all individuals regardless of their income. Consequently the uniform policy that assigns B to all individuals will outperform any personalization policy that uses a mix of both treatments.

This motivating example shows that heterogeneity in treatment effects is necessary but not sufficient to gain from personalization. In the next section we formally define the concept of gains from personalization, and develop a theory to explain the multiple factors that affect such gains.

3 Actionable Heterogeneity

We say that heterogeneity in treatment effects is actionable if the gain from personalization is large, i.e., it meaningfully improves outcomes beyond a uniform strategy. In this section we develop a statistical model that quantifies how a set of population level parameters influence the gain from personalization. We also extend the visual illustrations of the motivating example and provide a framework that explains how these factors affect the amount of actionable heterogeneity.

3.1 A Model of the Gain from Personalization

The gain from personalization would depend on the properties of the sample used in the experiment (e.g., the sample size), the quality of the algorithm used to estimate HTEs (i.e., how precise the estimates are), as well as population-level factors. Here, we develop a population-level model to analyze the expected gain from personalization, while in our empirical application we will also consider finite sample and algorithm quality effects.

We identify three factors that influence the gain from personalization: (i) the heterogeneity in individual outcomes within an intervention, (ii) the correlation of potential outcomes across interventions for the same person, and (iii) the variation in average outcomes across interventions.⁶

To simplify exposition, we use the notation and setting of the motivating example (two treatments, A and B with potential outcomes Y_i^A and Y_i^B). We assume that the potential outcomes are

⁶The motivating example only illustrated the influence of the third factor.

drawn from a multivariate normal distribution as follows:

$$\begin{pmatrix} Y_i^A \\ Y_i^B \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right) \quad (3)$$

where μ_A and μ_B are the population-level average outcomes, which are drawn i.i.d from a normal distribution $\mathcal{N}(M, s^2)$, σ^2 is the population-level variance of potential outcomes within a treatment, and ρ is the correlation of potential outcomes across treatments for the same individual.

Factor (i), which we call within-treatment heterogeneity, is captured by σ^2 , factor (ii), cross-treatment correlation, is captured by ρ , and factor (iii), variation in average outcomes, is captured by s^2 .

Connecting to the motivating example, within-treatment heterogeneity will be high if when exposed to "sales ends today", some people buy a lot and some buy little. Cross-treatment correlation will be high if the group of people that buy a lot under "sale ends today" is the same group that would also buy a lot under "there are only a few items left". The variation in average responses will be high if the average sales are very different under the two messages, such as in Experiment 2 in the motivating example.

Without loss of generality, we assume that $\mu_B > \mu_A$. This makes treatment B better on average than treatment A, and we call it the best uniform treatment. To have a positive gain from personalization, the personalization policy will need to improve over the average outcomes of treatment B. Hence, we compute the gain of a personalization policy in comparison to the best uniform treatment. The expression for this gain is:

$$\mathbb{E}[\mathbb{I}(Y_i^B > Y_i^A)Y_i^B + \mathbb{I}(Y_i^B \leq Y_i^A)Y_i^A] - \mu_B \quad (4)$$

which can be written as

$$\mathbb{E}[(1 - \mathbb{I}(Y_i^B \leq Y_i^A))Y_i^B + \mathbb{I}(Y_i^B \leq Y_i^A)Y_i^A] - \mu_B \quad (5)$$

and simplified to

$$\mathbb{E}[\mathbb{I}(Y_i^A - Y_i^B \geq 0)(Y_i^A - Y_i^B)] \quad (6)$$

This is the expected value of a rectified⁷ normal distribution with mean $\mu_A - \mu_B$ and variance $2\sigma^2(1 - \rho)$ and is equal to:

$$gain(\mu_A, \mu_B) = (\mu_A - \mu_B) \left[1 - \Phi \left(\frac{\mu_B - \mu_A}{\sigma \sqrt{2(1 - \rho)}} \right) \right] + \sigma \sqrt{2(1 - \rho)} \phi \left(\frac{\mu_B - \mu_A}{\sigma \sqrt{2(1 - \rho)}} \right) \quad (7)$$

Here, $\phi(\cdot)$ and $\Phi(\cdot)$ represent the probability density and cumulative distribution functions of the standard normal distribution, respectively.

3.2 What Affects the Gain from Personalization?

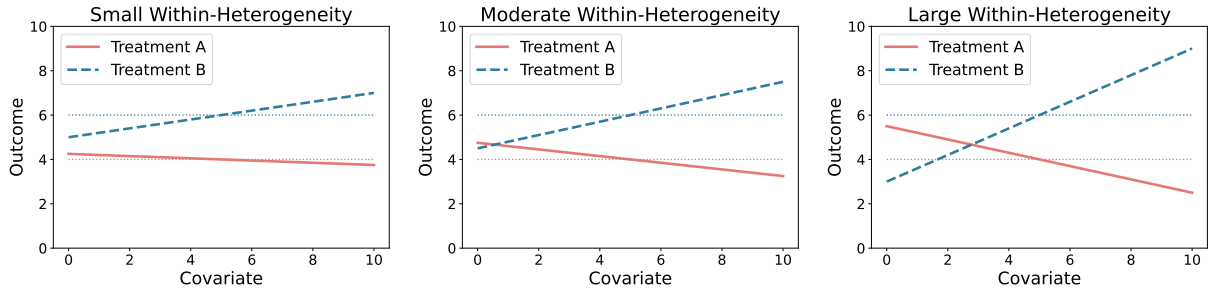
We use Equation (7) to analyze the effects of the three factors σ , ρ and $Var(\mu_B - \mu_A) = 2s^2$ on the gain from personalization.

Within-treatment heterogeneity (σ): Following the motivating example, if everyone reacts to a given text message in a similar way, there should be little to gain from personalization. Figure 3 presents the impact of σ in two ways. First (left panel), heterogeneity might not be sufficient to generate a crossover in potential outcomes within the range of the observed covariates. Even if theoretically there is a crossover in some range of the data (left of $x = 0$), there is no opportunity for personalization within the observed range. Second, even if there is a crossover within the observed data range (middle panel), the difference in potential outcomes for the same person might not be large enough to create a substantial gain over the best uniform policy. Finally, the right panel shows a case where there is large enough heterogeneity and the crossover is within the actionable data range. That is, gains from personalization are higher for higher σ , and we confirm this intuition in Proposition 1 in the Appendix, where we show that the partial derivative of Equation (7) with respect to σ is positive. We note that in practical applications, this heterogeneity must be *observable*, i.e., either captured by available covariates or by repeated observations of the same

⁷In a rectified distribution, the values below zero are replaced with zero and not just truncated.

individual. If the heterogeneity is truly unobserved, then it cannot be used to inform personalization policies, as it provides no guidance on what treatment to assign to an individual. Therefore, the empirical gain from personalization will always depend on the quality of available covariates (Rossi et al., 1996; Smith et al., 2023).

Figure 3: Effect of within-treatment heterogeneity



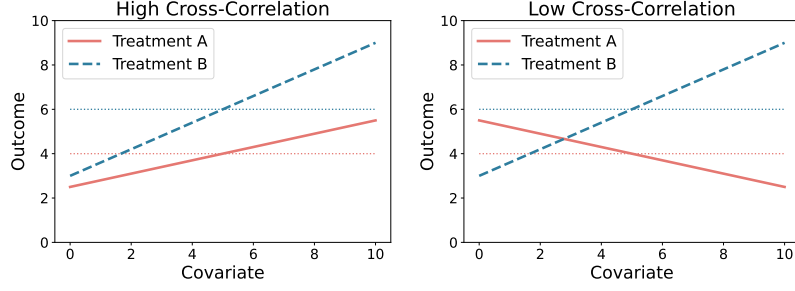
The plots display three A/B tests with identical cross-treatment correlation and average responses but varying within-treatment heterogeneity. Higher within-treatment heterogeneity (from left to right) increases the probability and magnitude of crossovers, and generates higher gain from personalization.

Cross-treatment correlation (ρ): Figure 4 provides an example with high and low cross-treatment correlations. In the left panel, the cross-treatment correlation is positive, so the same individuals who respond highly to treatment A also respond highly to B. In this case, despite high within-treatment heterogeneity, personalization will not provide much gain. By contrast, in the right panel, the correlation is negative providing an opportunity to gain from personalization. Proposition 2 in the Appendix formally proves this intuition by showing that the partial derivative of Equation (7) with respect to ρ is negative. To the best of our knowledge, this factor has not been considered previously in the literature on personalization.

Variation in Average Responses (s) When the variation in average outcomes s increases, the chances of a larger difference between the average outcomes of the best uniform and the other treatment are higher.

In this case of a larger difference, the best uniform policy will yield a better outcome, making it harder to achieve a gain with personalization. Figure 5 presents the impact in two ways. First, the left panel shows a case with variation s so large that there is no crossover in potential outcomes

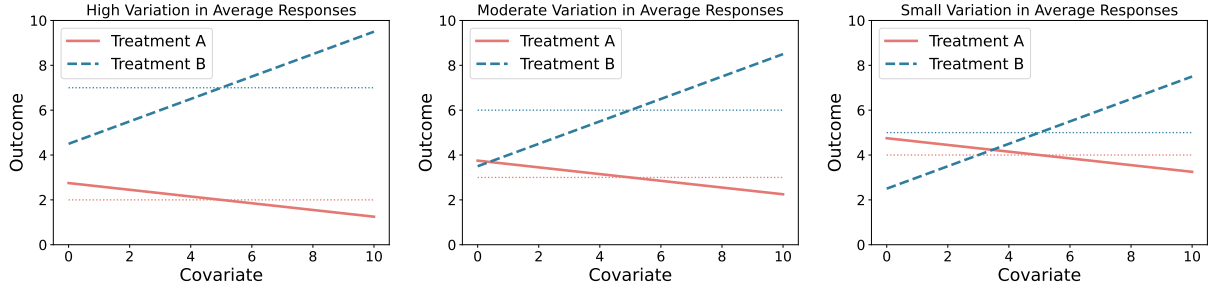
Figure 4: Effect of cross-treatment correlation



The plots display two A/B tests with identical within-treatment heterogeneity and average responses but different cross-treatment correlations: positive (left) and negative (right). Lower cross-treatment correlation yields higher personalization gains.

in the data range. The middle panel shows an example where there is a crossover, but the average outcome of treatment B is still quite high compared to a personalized policy. The right panel shows an example where a small amount of variation generates a crossover and a substantial gain from personalization. Correspondingly, Proposition 3 derives the partial derivative of the expected gain from personalization $\mathbb{E}_{\mu_A, \mu_B}(\text{gain}(\mu_A, \mu_B))$ with respect to s , and shows that it is negative.⁸

Figure 5: Effect of variation in average responses



The plots display three A/B tests that exhibit the same level of within-treatment heterogeneity and cross-treatment correlation but different levels of variation in average responses (higher on left to lower on right). More variation in average responses decreases the probability and magnitude of crossovers, yielding lower gain from personalization.

4 The Effect of Multiple Treatments

When more than two treatments are available, additional forces influence the potential gains from personalization. On one hand, more treatments offer more opportunities to find the optimal choice

⁸The proof assumes a normal prior distribution for μ_A and μ_B , but we show that this applies also in other cases in a later analysis.

for each individual, similar to the effect of σ from the two-treatment model. On the other, selecting the best uniform policy from a larger set of treatments increases the likelihood of a high outcome, akin to the effect of s in the two-treatment model.

We extend the analysis from the two-treatment, normally distributed case, and quantify the effects of the three forces we identified (within-treatment heterogeneity, cross-treatment correlation, and variation in average outcomes). We apply Monte Carlo simulation to an extension of the model in Section 3.1. Algorithm 1 describes the data generating and analysis process.

Algorithm 1: Simulation of the gain from personalization

Input: distribution of average responses F , within-treatment heterogeneity σ , cross-treatment correlation ρ , number of treatments m

- 1 Draw μ from distribution F — an m -vector of average responses for each treatment
 - 2 Construct a $m \times m$ individual responses covariance matrix Σ with σ^2 on the diagonal and $\rho\sigma^2$ off-diagonal
 - 3 Draw potential outcomes Y_i^T from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$
 - 4 The value of the optimal personalization policy corresponds to the maximum of potential outcomes for each individual averaged over individuals: $V_p = \frac{1}{n} \sum_{i=1}^n \max_T Y_i^T$
 - 5 The value of the best uniform policy corresponds to the maximum of average responses $V_u = \max_T \frac{1}{n} \sum_{i=1}^n Y_i^T$
 - 6 The gain from personalization is the difference between the two: $V_p - V_u$
-

We are particularly interested in understanding if there are cases when increasing the number of treatments leads to a *reduction* in the gain from personalization.⁹ Specifically, for peaked distributions of μ_T we expect a tradeoff. On one hand, when the distribution is more peaked, more treatments are close to one another in their average responses. On the other hand, the best-uniform intervention might be more of an outlier and harder to “beat” because it is selected from more treatments. Interestingly, the role of heavy-tailed distributions in generating higher average outcomes in A/B tests was also noted in Azevedo et al. (2020) and Berman and Van den Bulte (2022). However, in our case the mechanism that affects the results is different — a heavier-tailed distribution causes two opposing effects. It affects the benchmark that the personalization policy needs to “beat”, and it also affects the value that a personalization policy generates.

To explore this tradeoff, we generate values for the average potential outcomes per treatment

⁹Even though adding more treatments can increase the *value of a personalization policy*, i.e., the expected outcomes when treatments are personalized, it does not necessarily increase the *gain from personalization*, i.e., the improvement in expected outcomes over the best-uniform policy.

μ_T from the following distribution $F(\cdot)$:

$$\mu_T \sim \begin{cases} M & \text{with prob. } \pi \\ \mathcal{N}(M, s) & \text{with prob. } 1 - \pi \end{cases} \quad (8)$$

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 .¹⁰ 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).

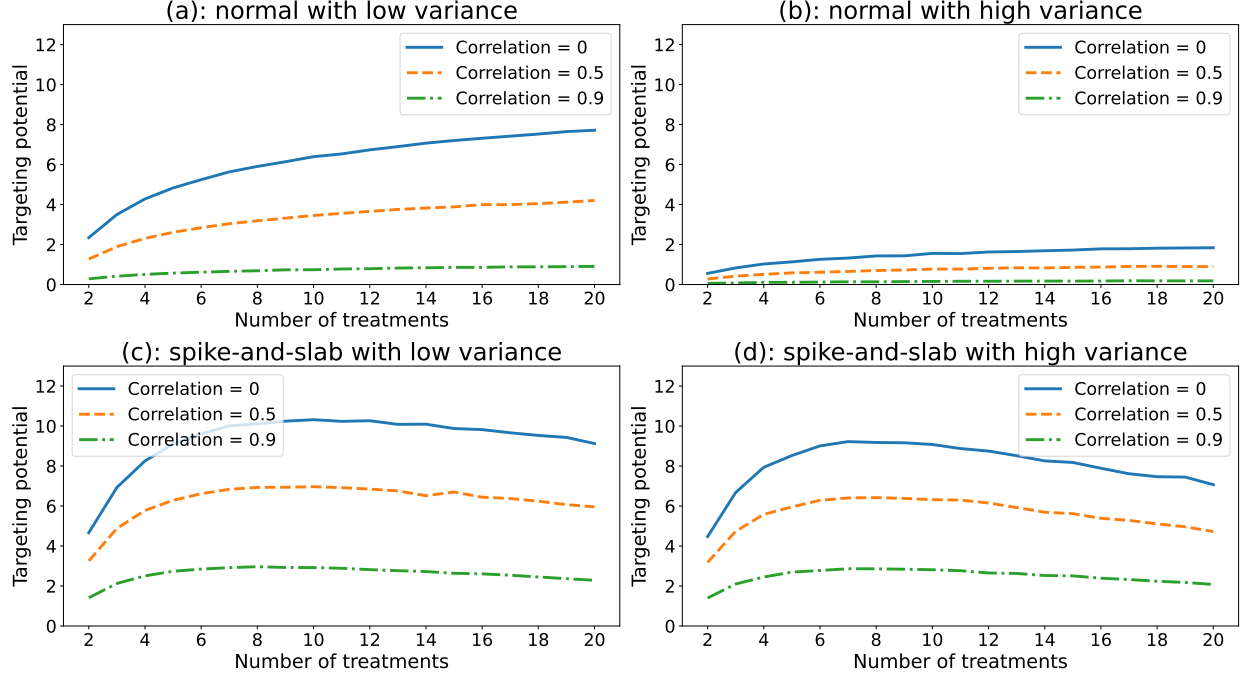
Figure 6 shows the results of the simulation. We normalize the within-treatment heterogeneity (σ) to 10 and vary the cross-treatment correlations $\rho = \{0, 0.5, 0.9\}$. The four panels show the gain from personalization over the best uniform policy as a function of the number of treatments.

Figure 6(a) shows that the results from the two-arm model extend to more arms — higher correlation across potential outcomes of arms is detrimental to the gains from personalization, but more arms generate potential for better outcomes from personalization. In comparison, figure 6(b) illustrates that when the variation of the average potential outcomes s is higher, the personalization potential suffers dramatically. This is because the value of the best uniform benchmark is more likely to be higher, thus harder to beat, lowering the benefit from personalizing treatments.

Continuing to Figure 6(c), we surprisingly see that increasing the number of arms can hurt the gain from personalization, and this effect is even more pronounced in Figure 6(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

¹⁰The value of M does not affect the analysis, as it shifts both the targeting policy and the uniform benchmark by the same value

Figure 6: 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.

one of the interventions will be drawn from the slab component, making its value potentially high and hard to beat, thereby lowering the benefits of personalization. However, when the number of arms is small, there is a high probability that all arms are drawn from the spike component, leading to a low variation in average responses and thus increasing the gain from personalization. As a result, the gain from 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 6(b) and 6(d). For 6(d), while the number of arms is small, all of them are likely to be from the spike component, and the variation in average outcomes is very low, so 6(d) looks very similar to 6(c) for a small number of arms. However, as the number of arms increases, the underlying slab component with a high variance comes into play, leading to an increase in the variation of average outcomes, which reduces the personalization potential. In contrast, in 6(b), the effect of high variation in average outcomes is uniformly applied to any number of arms, making the gain from personalization low for a small number of arms with a slow increase for more arms.

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

5 Empirical Application

The model we developed can be used to estimate the factors that affect the gain from personalization after an experiment was conducted, and can also be used during the design stage of an experiment if it is expected to be useful for personalization later.¹¹ However, our model focuses on population level parameters, but in practice, it should also account for the effects of finite samples, i.e., inaccurate predictions of HTEs. In Algorithm 2 we extend Algorithm 1 to take into account the prediction errors from the method that estimates the HTEs. The revised algorithm accounts for noise by introducing prediction error, effectively degrading the personalization quality. Hence, the quality of the prediction will depend on which statistical method is used to estimate HTEs. By incorporating prediction errors, Algorithm 2 allows us to explore how finite-sample effects impact the personalization potential.

We apply this algorithm to data from two large-scale field experiments that encourage vaccination (Milkman et al., 2021, 2022; Patel et al., 2023), and compare the predicted personalization gain to the ones estimated by five common methods used in practice. Our goal is to explain why potential gains from methods used in practice can be high or low. The next Section describes the data and experimental setting, followed by the empirical application.

5.1 The Megastudies of Milkman et al. (2021) and Milkman et al. (2022)

Our data comes 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

¹¹Our Algorithm 2 complements the RATE algorithm proposed by Yadlowsky et al. (2021). RATE AUTOOC 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.

Algorithm 2: Simulation of the gain from personalization with prediction error

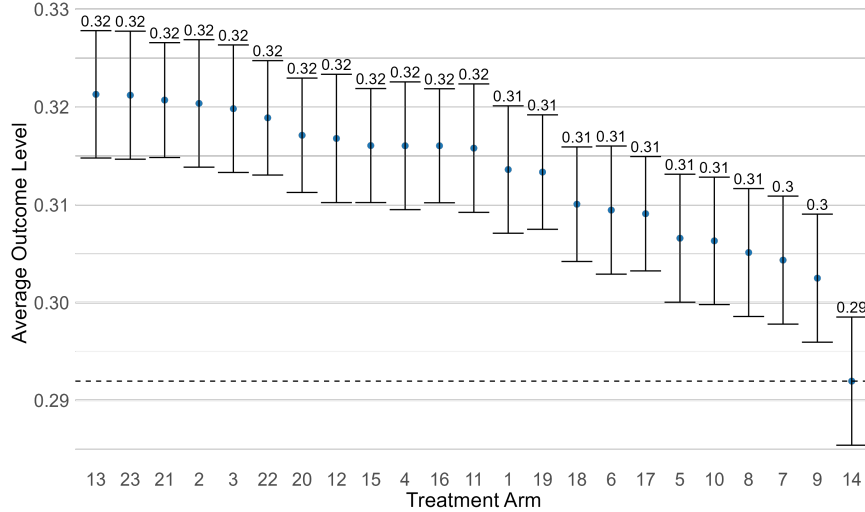
- Input:** distribution of average responses F , within-treatment heterogeneity σ , cross-treatment correlation ρ , number of treatments m , standard deviation of prediction errors σ_ε
- 1 From distribution F draw μ — an m -vector of average responses for each treatment
 - 2 Construct a $m \times m$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

people received at most one intervention.

The first study (Milkman et al., 2022, the “Walmart study”) analyzed the impact of low-cost behavioral nudges in the form of text messages 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, which received no text message. The dataset includes covariates such as gender, age, insurance type, health information, race, and zipcode-level variables such as median income and ethnic composition. Figure 7 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, conducted at Penn Medicine and Geisinger Health, (Milkman et al., 2021; Patel et al., 2023, the “Penn-Geisinger study”) also investigated the impact of text message 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: the

Figure 7: 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).

ones present in the Walmart study, such as gender, age, insurance type, health information, race, zipcode-level median income, as well as additional variables, such as smoking status, height and weight, marital status, flu vaccination history, message sending date, and Penn Medicine/Geisinger Health indicator. Figure 8 summarizes the average vaccination rates for each intervention on a training sample (70% of the population). Interventions are ordered by decreasing response rates.

The best treatment in both experiments has a very similar outcome (32% vaccination rate in Experiment 1 and 33% in Experiment 2), meaning that the best uniform policy in both experiment is likely to achieve similar results. Table 2 also provides summary statistics about other aspects of the experiments.

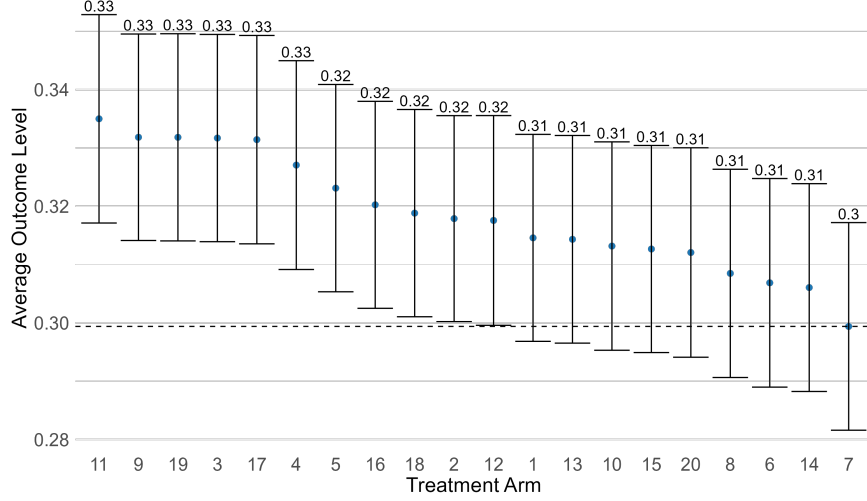
Table 2: Summary Statistics of Megastudies

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

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

While the Penn-Geisinger study includes a broader range of covariates, the Walmart study has a significantly larger sample size. Both of these factors — sample size and available covariates —

Figure 8: 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).

are important, as they impact statistical power and precision in HTE prediction.

5.2 Estimation Procedure

To apply Algorithm 2 to our datasets, we first estimate the input parameters: σ , ρ , s , and σ_ε , as well as assume a distribution F for average treatment outcomes.

Estimating s is done by computing the standard deviation of the average outcome levels \bar{Y}^a , which we find to be 0.007 for both datasets. The algorithm parameters describe the heterogeneity on the level of the potential outcomes and not treatment effects compared to a control. Hence, a machine learning algorithm that allows for flexible inclusion of the covariates (e.g., nonlinearities and interactions) and that operates on the level of the potential outcomes is preferred.

We elected to use an out-of-sample T-Learner implementation of XGBoost (Chen and Guestrin, 2016) which satisfies these criteria (flexibility and modeling of potential outcomes) to estimate σ_ε , σ and ρ :

1. For each intervention a and individual i , estimate the outcome model $Y_i^a = h^a(x_i) + \eta_i$, where $h^a(x_i)$ is a flexible function that maps the observed covariates to a part of the potential outcome that they explain, and η_i is the residual component that cannot be attributed to

the observed covariates. We assume that η_i is independent of $h^a(x_i)$ and that $\mathbb{E}[\eta_i] = 0$. The output of the model is a prediction function $\hat{h}^a(x_i)$.

2. On the holdout sample, we predict the outcome $\hat{y}_i^a = \hat{h}^a(x_i)$ for each individual and treatment.

To estimate σ_ε , we compute $\hat{\sigma}_\varepsilon^2 = \text{Var}(y_i - \hat{y}_i)$ on the holdout sample.

The parameter σ measures $\text{Var}(h^a(x_i))$ and similarly ρ measures the correlation between $h^a(x_i)$ and $h^{a'}(x_i)$ for some treatment a' . Estimating these parameters is not straightforward. One naive way to estimate them would be to compute $\text{Var}(\hat{h}^a(x_i))$ and the corresponding correlations. However, because $\hat{h}^a(x_i)$ is an estimate, it is expected to contain errors which would inflate the variance and underestimate the correlation. Both of these effects would lead to overestimation of the potential gains from personalization as shown in Section 3. Another naive way to estimate σ is to estimate $\text{Var}(Y_i^a)$, but since Y_i^a contain errors from η_i this approach would also overestimate the true σ .

To address these estimation challenges, we introduce a stratified estimation procedure. Our method uses the out-of-sample data to estimate σ and ρ utilizing both the predicted \hat{Y}_i^a and the observed outcomes Y_i . For all individuals that were assigned treatment a , we divide them into 10 quantiles based on their predicted outcome \hat{Y}_i^a . Within each quantile, we compute the average observed outcome \bar{Y}_i^a and then compute the standard deviation of these 10 averages for σ which gives the estimate $\hat{\sigma}$. The estimation of σ hinges on the properties of stratification and the law of total variance. In Proposition 4 in the Appendix we prove that this estimation procedure yields:

$$\hat{\sigma}^2 = \text{Var}(\bar{Y}_i^a) \approx \text{Var}_{\text{quantiles}}(E[Y_i^a | \text{quantile}]) \approx \text{Var}(h^a(x_i))$$

In the Penn-Geisinger and Walmart studies, this procedure yields estimate of σ of 0.267 and 0.078 respectively, favoring the Penn-Geisinger study as having a larger potential for gains from personalization.

Similarly, the cross-treatment correlation ρ is estimated as follows: For each individual, we assign them the average out-of-sample outcome \bar{Y}_i^a based on their predicted quantile of \hat{Y}_i^a . Hence, each individual would be assigned m values. We then compute the pairwise correlations of these

assigned outcomes for each arm pair, which gives an estimate $\hat{\rho}^{aa'}$ for arms a and a' . We take the average of all these estimates, because in both datasets they do not vary much. The average correlation equals 0.61 in the Walmart study and 0.80 in the Penn-Geisinger study. Hence, the Walmart study interventions are more independent, favoring it as having a larger potential gain from personalization.

To validate the results from Algorithm 2, we estimate and evaluate five popular personalization methods: (a) OLS, (b) S-Learner XGBoost, (c) T-Learner XGBoost (Chen and Guestrin, 2016), (d) Causal Forest (Athey et al., 2019), and (e) Policy Tree (Zhou et al., 2023), which are described in detail in Appendix A.2. We chose a sample of methods to demonstrate various common approaches to personalization: (i) a simple linear method (OLS), (ii) methods based on a general multi-purpose machine learning model (XGBoost), (iii) a method based on a specialized machine-learning model for estimating HTEs (Causal Forest), and finally (iv) a machine learning model fully specialized for personalization (Policy Tree). A summary comparing the characteristics of these methods is provided in Table 3.

Table 3: Personalization 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 allocate 70% of each dataset for training each model, while the remaining 30% is reserved for evaluation of the resulting personalization policy. We use Inverse Probability Weighting (IPW, see, e.g., Rafieian and Yoganarasimhan, 2023; Simester et al., 2020a) on the holdout sample to evaluate the policies. Using a holdout sample for policy evaluation with IPW is crucial to avoid the winner’s curse and ensure an unbiased estimate (Andrews et al., 2024). To evaluate policy π , every individual in the holdout sample is assigned the treatment prescribed by the policy, which is denoted by $\pi(x_i)$. Then, we identify individuals whose realized experimental treatment assignment equals the one prescribed the policy, that is, the individuals for whom $T_i = \pi(x_i)$. Finally, we

reweigh the outcomes of these people according to the propensity scores of receiving the treatment $T_i = \pi(x_i)$, denoted by $\hat{e}(a|x_i)$:

$$\widehat{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}\{\pi(x_i) = T_i\} Y_i}{\hat{e}(T_i|x_i)} \quad (9)$$

Since both datasets come from randomized experiments, the propensity scores $\hat{e}(a|x_i) = e(a)$ equal the probability of treatment assignment and are known.

5.3 Empirical Results

Table 4 shows the gains from personalization across different estimation methods, comparing the IPW scores of each personalization policy to the best uniform benchmark.

Table 4: Personalization Results

	Best Uniform	OLS	S-XGB	T-XGB	MACF	PT-CF	Our Model
Walmart	31.2	31.5	31.4	31.7	32.4	32.2	31.6
Bootstrap SE	(0.5)	(0.6)	(0.6)	(0.6)	(0.6)	(0.6)	
Rel. Improvement		1%	0%	1%	4%	3%	
Penn-Geisinger	31.5	32.5	34.9	34.8	37.3	34.1	37.2
Bootstrap SE	(1.4)	(1.7)	(1.8)	(1.7)	(1.8)	(1.6)	
Rel. Improvement		3%	11%	10%	18%	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 a positive gain from personalization (i.e., all personalization methods we used perform better compared to the best uniform policy out of sample). However, in the Penn-Geisinger study, the machine learning methods achieved at least 8% personalization gain with the best one achieving 18%, while in the Walmart study, the best-performing method only achieved a 4% personalization gain.

The last column of Table 4 reports the estimated gain from personalization using Algorithm 2. The method uses the estimated moments and prediction error from T-XGB as detailed in the previous section. The results mimic the pattern we observe for the machine learning methods: our statistical model predicts a personalization gain of 1% in the Walmart study and 18% in the

Penn-Geisinger study. These values are within the 95% confidence intervals of the estimates from the other methods.

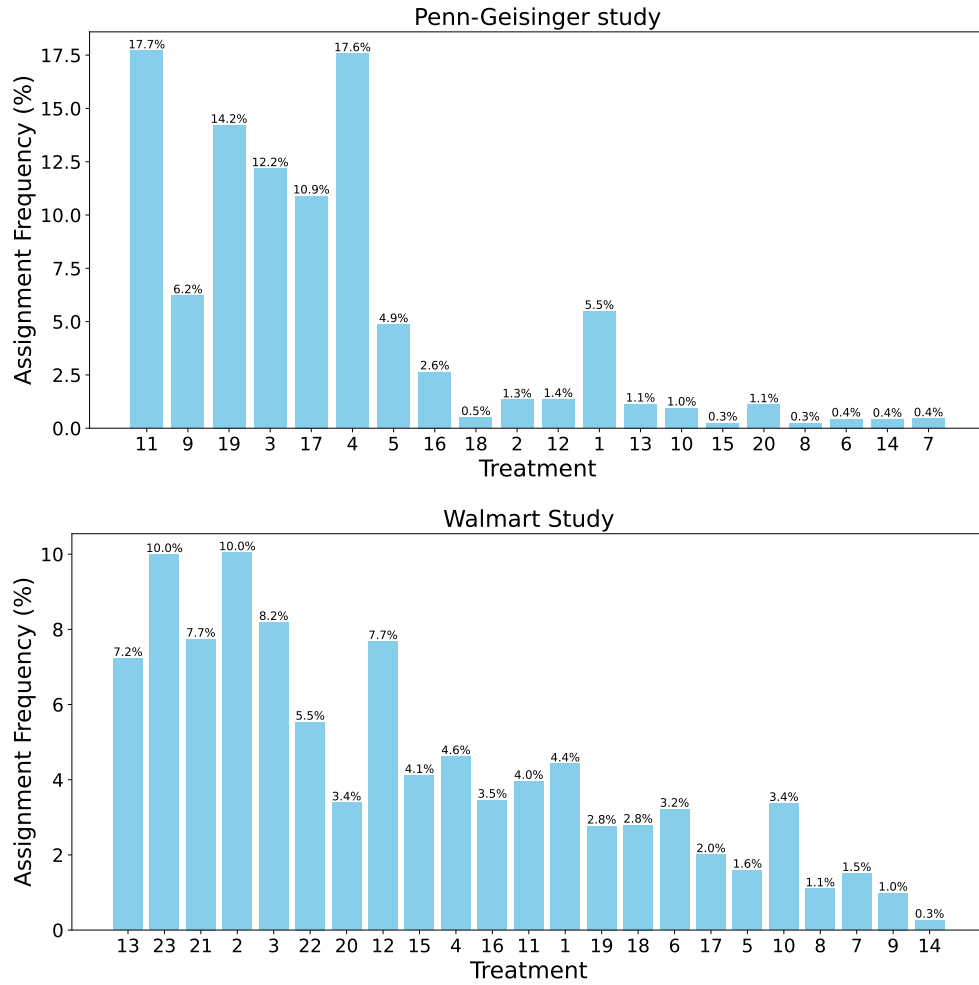
To summarize, after comparing five popular personalization methods, we found that the Penn-Geisinger study, which has higher within-treatment heterogeneity, has a higher personalization gain potential compared to the Walmart study. By contrast, although the Walmart study has lower cross-treatment correlation, its potential for gains from personalization is not as high.

It is also interesting to analyze the leading recommended personalization policies. Figure 9 illustrates the treatment assignment according to the best personalization policy (multi-arm causal forest) in the Penn-Geisinger and Walmart studies. The interventions appear in the order of their average outcomes (vaccination rates), and the height of each bar is the percentage of population the personalization policy allocates them to.

In both studies, each treatment was used with positive probability, but Penn-Geisinger’s personalization policy concentrated on fewer treatments (6 out of 20), covering 80% of the population, while for Walmart, the treatment assignment is more uniform. This is consistent with our finding that the within-treatment heterogeneity in the Walmart dataset is low relative to the prediction error: in this case, a personalization algorithm will not find much difference between assigning an individual to one treatment or another. We also observe that the set of treatments performing the best on average (as indicated by the ordering of bars in the chart) is also the set of treatments mostly used in personalization. However, the ranking is not exactly the same: that is, if a treatment that is high-performing on average has a large overlap in the people it affects with another high-performing treatment, it makes sense for a personalization policy to skip one of those treatments and choose a lower-performing treatment that has less overlap with the previously chosen treatment.

The multi-arm causal forest is not an easily interpretable model, making it challenging to describe the groups of people assigned to each arm. One common way to identify the most salient variables used to estimate treatment heterogeneity in a causal forest is a version of a variable importance metric, which assigns for each variable its frequency of appearance in splits of the trees in the causal forest (see, e.g., Habel et al., 2024; Murakami et al., 2022). It is important to

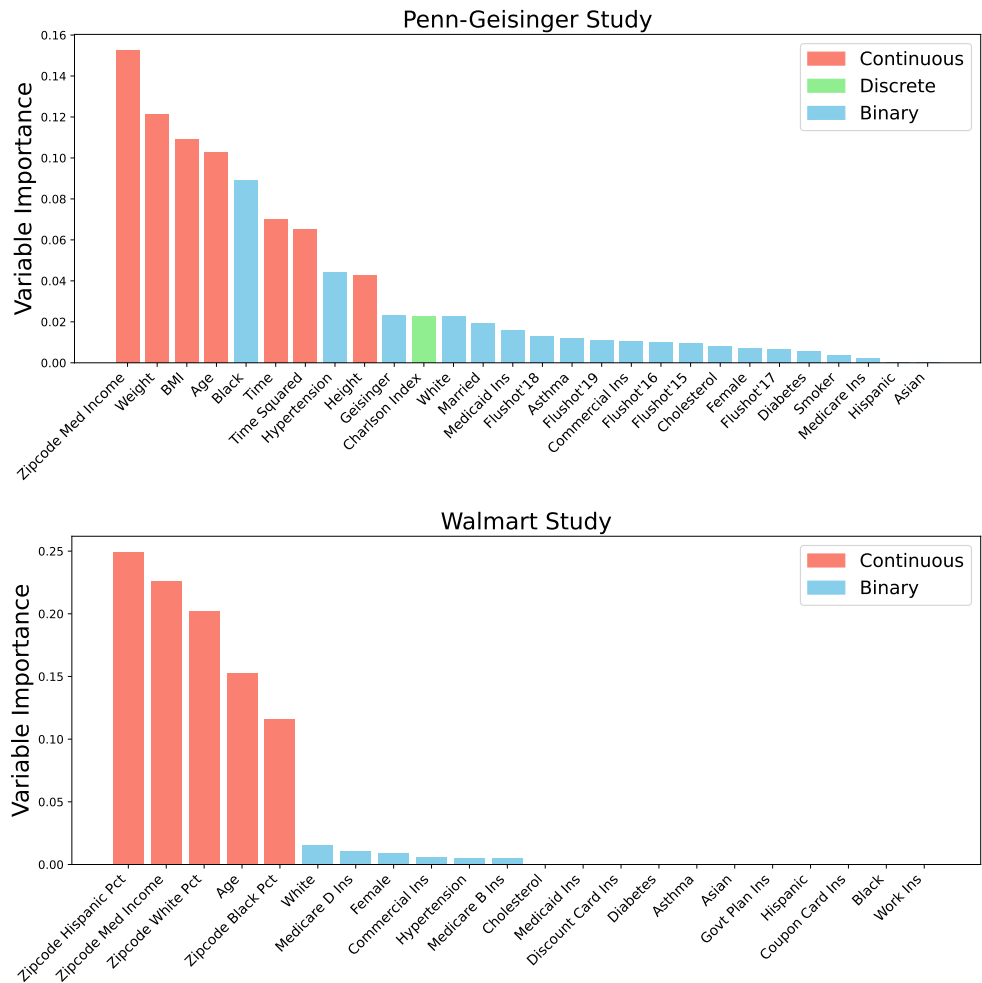
Figure 9: Arm Assignment



Percentages of population assigned to each intervention by the personalization policy estimated using multi-arm causal forest. The order of the interventions from left to right is by the rank of the average outcome levels in the training sample (same as Figures 7 and 8).

note that by design, this metric favors continuous variables. The reason is that if there are no meaningful effects and a forest makes splits at random, the continuous variables will have more chances to be split on compared to discrete variables, inflating their variable importance scores. Hence, the variable scores in the Walmart study (Figure 10) might reflect the true importance of the variables, with the caveat that their pattern is also consistent with the algorithm’s tendency to split on continuous variables more often when there are no meaningful effects. In contrast, in the Penn-Geisinger study a few binary variables (notably *Black* and *Hypertension*) have quite high importance scores, which suggests that the causal forest is learning meaningful heterogeneity.

Figure 10: Variable Importances



The fraction of multi-arm causal forest splits for each variable. A higher value indicates that the variable is used more often as a splitting variable within the nodes in the trees.

Because it is challenging to describe and analyze the personalization policies when there are more than two treatments, we use a visual summary of the normalized means of the covariates of people assigned to each treatment by the personalization policy to draw some insights. The visualization appears in Figure 11 and the means in each cell are normalized by the mean and standard deviation of each column to make the cells comparable. For example, if the weight variable has a positive value, it means that the people assigned to this treatment have a higher than average weight. Red cells indicate overrepresentation or above average value, and blue cells indicate underrepresentation or below average value.

We draw the following insights: in both datasets, the “No message” condition tends to be assigned to older people. In contrast, in the Penn-Geisinger study, “Share a joke about the flu” is recommended to very young people who have not been previously vaccinated. Health behavior quizzes are among other interventions that are recommended to target those who were previously not vaccinated and healthier people in the Penn-Geisinger study. In comparison, in the Walmart study the messages assigned to a healthier group are “Get a flu shot to avoid getting sick” and “Think about risk of catching the flu”. We believe that such visual descriptive analysis of black-box personalization policies in experiments with many treatments and many covariates can be useful to gain data-driven insights for subsequent theory development, identification of relevant constructs and moderators, and additional experimentation.

To summarize, our initial results found significantly higher personalization gains in the Penn-Geisinger study compared to Walmart, consistent with our model’s predictions. In our setting, the multi-arm causal forest was the best-performing personalization method.

6 Interpretation and Managerial Insights

Our analysis has shown a higher potential gain from personalization in the Penn-Geisinger study compared to the Walmart study. This section (1) explores the nature of this difference and (2) examines the sensitivity of the gains from personalization to parameter changes to provide insights for managers on how to increase personalization impact.

First, we perform a sensitivity analysis using our model and visualize the impact of changing

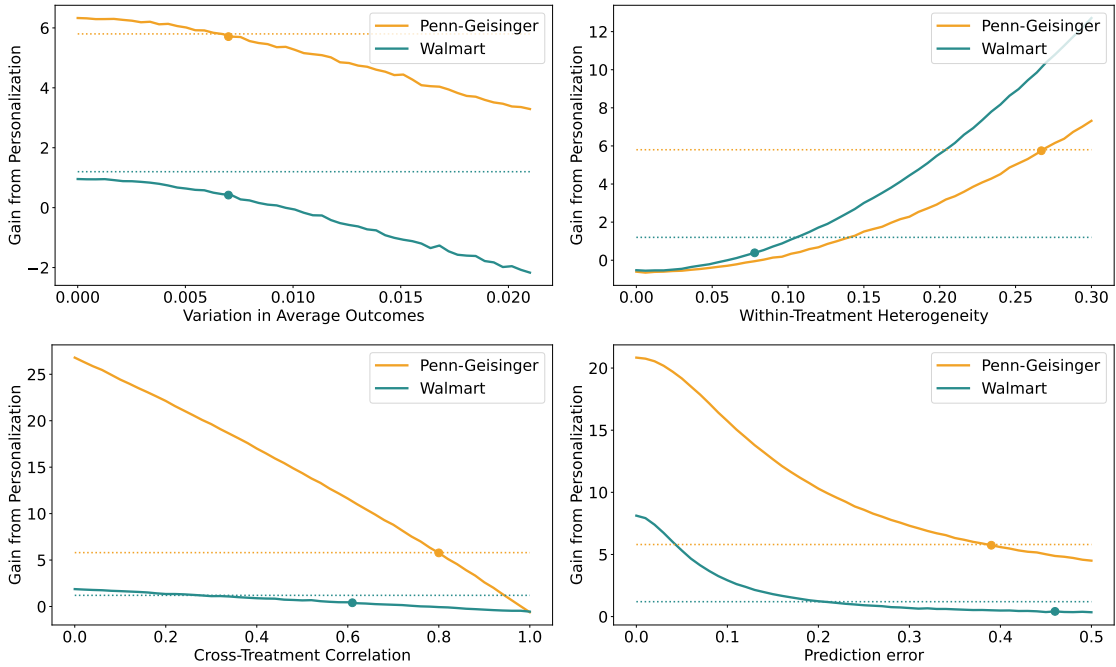
Figure 11: Causal forest: covariates distribution



Colors indicate the average level of each variable for the individuals assigned to each intervention using the multi-arm causal forest. Red colors indicate values higher than the group averages for the variables, and blue indicate average values lower than the variable group average. Darker colors indicate being farther from the mean. The average values are normalized to the group mean average within each column.

each parameter individually on the predicted gain from personalization. Figure 12 illustrates this sensitivity to changes across four dimensions: (i) variation in average outcomes (s), (ii) within-treatment heterogeneity (σ), (iii) cross-treatment correlation (ρ), and (iv) prediction error (σ_ε). In each panel only one parameter is being manipulated, while the other are kept fixed at the original values of each study. The dotted lines indicate the estimated gain from personalization using the multi-arm causal forest, while the bold dot indicates the predicted value by our model. The Penn-Geisinger study sensitivity curve dominates the Walmart one across all parameters except for within-treatment heterogeneity. Notably, the Walmart study could achieve comparable or even higher gains than Penn-Geisinger with only a modest increase in within-treatment heterogeneity, as seen in the top right panel. This suggests that the difference in the levels of within-treatment heterogeneity could be the main driver of the difference in the gain from personalization across the studies.

Figure 12: Sensitivity plots



The figure shows the sensitivity of the predicted gain from personalization for different levels of variation in average outcomes s , within-treatment heterogeneity σ , cross-treatment correlation ρ , and prediction error σ_ε for the Penn-Geisinger and the Walmart studies. The dots depict the gain from personalization predicted using the estimated values of the parameters above. The dotted lines correspond to the IPW estimates of the gain from personalization in Table 4.

To confirm this intuition, we perform a counterfactual analysis where we replace one parameter value of a study with the corresponding value from the other study. The goal of the analysis is to provide an intuition about the importance of each parameter. We do this by looking at how much making the distance in one parameter value disappear (making the studies more similar) affects the predicted gains. Table 5 presents the results, where each column displays the gain from personalization for each study where only one focal parameter was replaced with the value from the other study.

Table 5: Gain from personalization: counterfactuals with factors exchanged

	Original	σ	ρ	σ_ε	m
Penn-Geisinger	5.714	-0.029	11.311	4.890	5.964
Walmart	0.413	10.073	-0.039	0.477	0.407

Counterfactual analysis of personalization gains when individual parameters (σ , ρ , σ_ε , m (number of arms)) are swapped between studies. Each column shows the resulting gain for each study when only the focal parameter is replaced.

For the Walmart study, changes in all parameters except for σ wouldn't affect the gains much, while increasing σ to the level of the Penn-Geisinger study will make the gain from personalization even higher than the original Penn-Geisinger gain. In comparison, for the Penn-Geisinger study, some benefit can be achieved by decreasing ρ to the level of the Walmart study, but because the original gain was already high, this benefit is not as substantial as for the Walmart study. This analysis confirms our previous conclusion that the parameter that drives most of the difference in the gains from personalization is the drastic difference in within-treatment heterogeneity.

We then explore the reasons behind the differences in within-treatment heterogeneity. Possible reasons include (1) availability of relevant covariates, (2) the level of variation in covariates in the study's sample, and (3) the importance weights of each covariate in translating to treatment effects.¹² One way to disentangle (1) from (2) and (3) is to compute the value of within-treatment heterogeneity using only the covariates that overlap between the two studies. We find that this does not change the estimated value of σ for the Walmart study (0.078), while reducing it dramatically for the Penn-Geisinger study (0.267 drops to 0.089). We conclude that the majority of the difference

¹²For example, if the potential outcomes follow the linear DGP $Y_i^A = \alpha_A + \beta_A x_i + \varepsilon_{Ai}$ and $Y_i^B = \alpha_B + \beta_B x_i + \varepsilon_{Bi}$, then the variation in the treatment effect $Y_i^A - Y_i^B$ is higher when the difference $\beta_A - \beta_B$ is higher.

in within-treatment heterogeneity and, consequently, in the gain from personalization, seems to be driven by access to additional relevant covariates in the Penn-Geisinger dataset. The remaining difference might be due to differences in covariate variation or in importance weights. As Figures 13 and 14 in the Appendix show, sample homogeneity might indeed be different: both age and median zipcode-level income are more varied in the Penn-Geisinger study.

Our model is also useful to provide guidance for managers and experimenters who are trying to find the best way to invest resources in designing experiments with higher personalization value. We use the model to compute an “elasticity” measure with respect to each parameter, which provides an estimate of what managers could achieve if they could slightly improve one of the model parameters. Table 6 reports the results of simulations when all parameters are held fixed and one is changed by 1%. For both datasets, we find that reducing the cross-treatment correlation is expected to yield the largest increase in gain from personalization per 1% improvement. This finding emphasizes that although cross-treatment correlation did not receive much attention previously, it has the potential to create an impactful improvement in gains, and managers should carefully consider it.

Table 6: Sensitivity of gains from personalization

	Original	1% Lower s	1% Higher σ	1% Lower ρ	1% Lower σ_ϵ
Penn-Geisinger	5.714	5.723	5.888	6.014	5.827
Walmart	0.413	0.399	0.416	0.420	0.375

Results indicate the expected personalization gain over the best uniform policy per 1% change in the focal parameter while keeping other parameters constant.

How can managers lower the cross-treatment correlation? One solution is to test treatments that are expected to be different in their behavioral mechanisms and constructs they activate. However, in practice, this might be hard to do without also affecting the variation in average outcomes (s): such varied treatments may also produce more variation in average outcomes and thus hurt the gain from personalization (which is not necessarily bad for the *value* of the best policy). In other words, in contrast to the usual recommendation from non-personalized A/B tests to make the treatments more varied and distinct (Berman and Van den Bulte, 2022), the effects of creativity and independence in interventions create a tradeoff: they can decrease ρ but might increase s .

A potential way to improve ρ without affecting s is to collect more covariates and replace those

used in HTE estimation with the ones that minimize correlation. This means collecting additional covariates that hopefully better discriminate among groups of individuals and thus result in a lower correlation of potential outcomes. One way to implement this idea is to extend the multi-arm causal forest algorithm to minimize correlation of predicted outcomes along with maximizing the variance at the tree splits.

Collecting more covariates might not only decrease cross-treatment correlation ρ but can also help increase within-treatment heterogeneity σ : relevant covariates can capture more variation in potential outcomes. Finally, an experimenter can also benefit from decreasing the prediction error σ_ε . To do that, one can build more sophisticated algorithms (and again, collect more covariates) or increase sample sizes. However, complex algorithms often suffer from overfitting, so it is important to measure the improvements in σ_ε out of sample.

7 Conclusion

One potential goal of running experiments with many treatments is to be able to use the results for personalization and achieve better returns than assigning a uniform policy. Traditionally, it is assumed that heterogeneity of treatment effects is necessary to provide such gains. We developed a statistical model to explain what generates gains from personalization, and showed that heterogeneity is necessary but not sufficient for high gains. Along with heterogeneity, other population-level factors such as cross-treatment correlation and variation in average outcomes matter. In our empirical application, we found that cross-treatment correlation, a factor rarely considered in the personalization literature, is expected to yield the highest return in gains per 1% improvement. In addition, we found that increasing the variation in average outcomes can be a double edged sword: while in the classic A/B testing literature (that focuses on uniform policies) higher variation in average outcomes is often recommended, in personalization applications, although it might increase the value of the best policy, it might adversely affect the gain from personalization. Finally, we show that for peaked distributions of average outcomes, adding more treatments to an experiment can sometimes hurt such gains.

To validate our model, we calibrated it to two large-scale field experiments that encouraged flu

vaccination. In one study there were relatively high gains from personalization, while in the other they were low. The predictions of the model are consistent with the empirical estimates of five popular personalization approaches. We found that the multi-arm causal forest achieves the highest personalization gain for both studies. Our model indicates that the difference in gains from personalization between the two studies is driven by a much higher level of within-treatment heterogeneity in the Penn-Geisinger study, which in turn is explained to a large degree (but not completely) by better covariate availability. However, an elasticity analysis shows that the most promising avenue of investment to improve these gains is by lowering the cross-treatment correlation.

This conclusion reiterates the point of Rossi et al. (1996) that heterogeneity is only as good as the available covariates to estimate it. For the same experiment, collecting one set of covariates may generate a lot of actionable heterogeneity, which would enable personalization — while another set of covariates may not find any heterogeneity at all. By applying the model to different experiments and different sets of covariates (e.g., Smith et al., 2023), future work might be able to get insights into which covariates tend to offer heterogeneity that is useful for personalization 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.

There are also some limitations to our model and analysis that can be explored in future work. First, our model assumes a simplified cross-treatment correlation structure where one parameter captures the correlation of all potential outcomes. Naturally it is possible that each pair of treatment will have different levels of correlation. It turns out that in the two studies we analyze the correlation is relatively uniform and this level of flexibility adds little. A second limitation is that our model only captures four factors that affect the gains from personalization, and potentially others exist. We do note that our model’s predictions aligns relatively well with the empirical personalization estimates, but in other contexts this might not be true. A third limitation is that the studies we analyzed were not specifically designed to maximize the heterogeneity and gains from personalization. Researchers were asked to create the behavioral interventions they thought would work best, and not necessarily the ones that would generate the most varied responses. The fact that we do not observe high added

value of personalization might be explained by this design, but does not mean it is necessarily impossible in this context to achieve higher gains.

Our paper has several implications. For practitioners, it provides a tool to gauge the quality of data for personalization either before or after running an experiment. It also allows managers to run counterfactual exercises and decide if they should come up with more treatments, look for a better method, collect additional covariates, or increase sample sizes. For researchers, it sheds light on the determinants of the gain from personalization 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 personalization potential. The two studies have remarkably similar contexts, and yet the gains from personalization in one are four times higher than in the other. Finally, our paper provides an interesting perspective on analysis of heterogeneity that is becoming customary along with main effects. This analysis often aims to provide sharper insights into causal mechanisms at work, identify important moderators and boundary conditions, or suggest a higher effectiveness from personalization. Our present work shows that the latter effect is not as straightforward: not all kinds of heterogeneity are useful for personalization.

Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

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).
- Berman, R. and Van den Bulte, C. (2022). False discovery in a/b testing. *Management Science*, 68(9):6762–6782.
- 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.
- Boutilier, J., Jonasson, J. O., Li, H., and Yoeli, E. (2024). Randomized controlled trials of service interventions: The impact of capacity constraints. *arXiv preprint arXiv:2407.21322*.
- 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). Improv-

- ing 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.
- Goli, A., Lambrecht, A., and Yoganarasimhan, H. (2024a). A bias correction approach for interference in ranking experiments. *Marketing Science*, 43(3):590–614.
- Goli, A., Reiley, D. H., and Zhang, H. (2024b). Personalizing ad load to optimize subscription and ad revenues: Product strategies constructed from experiments on pandora. *Marketing Science*.
- 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.
- Habel, J., Alavi, S., and Heinitz, N. (2024). Effective implementation of predictive sales analytics. *Journal of Marketing Research*, 61(4):718–741.

- Hitsch, G. J., Misra, S., and Zhang, W. W. (2024). Heterogeneous treatment effects and optimal targeting policy evaluation. *Quantitative Marketing and Economics*, 22(2):115–168.
- Hu, A. (2023). Heterogeneous treatment effects analysis for social scientists: A review. *Social Science Research*, 109:102810.
- Huang, T.-W. and Ascarza, E. (2024). Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Marketing Science*.
- 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.
- Mosleh, M., Eckles, D., and Rand, D. G. (2024). Tendencies toward triadic closure: Field-experimental evidence. Technical report, Center for Open Science.
- Murakami, K., Shimada, H., Ushifusa, Y., and Ida, T. (2022). Heterogeneous treatment effects

- of nudge and rebate: Causal machine learning in a field experiment on electricity conservation. *International Economic Review*, 63(4):1779–1803.
- 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.
- Sahni, N. S., Wheeler, S. C., and Chintagunta, P. (2018). Personalization in email marketing: The role of noninformative advertising content. *Marketing Science*, 37(2):236–258.
- 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. (2020a). Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science*, 66(8):3412–3424.
- Simester, D., Timoshenko, A., and Zoumpoulis, S. I. (2020b). Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science*, 66(6):2495–2522.

- Simonov, A., Nosko, C., and Rao, J. M. (2018). Competition and crowd-out for brand keywords in sponsored search. *Marketing Science*, 37(2):200–215.
- Smith, A. N., Seiler, S., and Aggarwal, I. (2023). Optimal price targeting. *Marketing Science*, 42(3):476–499.
- Timoshenko, A., Ibragimov, M., Simester, D., Parker, J., and Schoar, A. (2020). Transferring information between marketing campaigns to improve targeting policies. Technical report, Working Paper.
- 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.
- Yang, J., Eckles, D., Dhillon, P., and Aral, S. (2024). Targeting for long-term outcomes. *Management Science*, 70(6):3841–3855.
- 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, J. Z., Netzer, O., and Ansari, A. (2014). Dynamic targeted pricing in b2b relationships. *Marketing Science*, 33(3):317–337.
- Zhang, W. W. (2023). Optimal comprehensible targeting. Working Paper.
- 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 Appendix

A.1 Proofs

Proposition 1 (Within-treatment heterogeneity) *The gain from personalization in increasing in within-treatment heterogeneity (σ).*

Proof. We compute the derivative of Equation (7) with respect to σ . Let us denote $d = \mu_B - \mu_A$ (we assume $d \geq 0$). The partial derivative is equal to:

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

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

Proposition 2 (Cross-treatment correlation) *The gain from personalization in decreasing in cross-treatment correlation (ρ).*

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

Equation (7) 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 . ■

Proposition 3 (Variation of average responses) *The expected gain from personalization is decreasing in variation of average responses (s).*

Proof. Let us denote the value of Equation (7) by $V(d) = \text{gain}(\mu_B - \mu_A)$, where $d = \mu_B - \mu_A$. Recall that when computing this value we assumed that $\mu_B > \mu_A$, i.e., $d > 0$. Since μ_B, μ_A are

i.i.d. draws, the expectation of the gain from personalization over this distribution can be written using the law of total expectation:

$$\mathbb{E}_{\mu_A, \mu_B}[\text{gain}(\mu_A, \mu_B)] = \mathbb{E}[\text{gain}(\mu_A, \mu_B) | \mu_B > \mu_A] \cdot P(\mu_B > \mu_A) + \mathbb{E}[\text{gain}(\mu_A, \mu_B) | \mu_B \leq \mu_A] \cdot P(\mu_B \leq \mu_A) \quad (10)$$

By symmetry, $P(\mu_B > \mu_A) = \frac{1}{2}$, and $\text{gain}(\mu_A, \mu_B) | (\mu_B > \mu_A) = \text{gain}(\mu_A, \mu_B) | (\mu_B \leq \mu_A)$. Therefore,

$$\mathbb{E}_{\mu_A, \mu_B}[\text{gain}(\mu_A, \mu_B)] = \mathbb{E}_d[V(d) | d > 0], d = \mu_B - \mu_A \quad (11)$$

Since μ_A, μ_B 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: $\mu_A, \mu_B \sim \mathcal{N}(M, s^2)$ and $\mu'_A, \mu'_B \sim \mathcal{N}(M, s'^2)$ with $s' > s$. We first prove two lemmas.

Lemma 1 *The conditional distribution of $d' = \mu'_B - \mu'_A, d' \geq 0$ stochastically dominates the conditional distribution of $d = \mu_B - \mu_A, d \geq 0$.*

Proof. 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:

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

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

Lemma 2 *$V(d)$ is decreasing in d when $d > 0$.*

Proof. For simplicity, we will let $v = \sigma\sqrt{2(1-\rho)}$.

Equation (7) 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, $V(d)$ is decreasing in $\mu_B - \mu_A$. ■

Combining the two lemmas above and the first-order stochastic dominance theorem,

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

In other words, the expected gain from personalization is decreasing in the variation of average responses s . ■

Proposition 4 (Stratified estimation of within-treatment heterogeneity) *The estimation procedure in Section 5.2 yields:*

$$\hat{\sigma}^2 = Var(\bar{Y}_i^a) \approx Var_{quantiles}(E[Y_i^a|quantile]) \approx Var(h^a(x_i))$$

Proof. We can write the variance of the observed outcomes as:

$$Var(Y_i^a) = E_{quantiles}[Var(Y_i^a|quantile)] + Var_{quantiles}(E[Y_i^a|quantile]) \quad (12)$$

Because η_i is independent of $h(x_i)$ then $E[Y_i^a|quantile]$ would not contain any bias from η_i assuming the quantile has enough observations. Further, because the groups comprise of people within the same predicted quantile, we expect their $h(x_i)$ to be close and homogeneous. Hence $E_{quantiles}[Var(Y_i^a|quantile)] \approx Var(\eta_i)$. Consequently, our estimation procedure computes $Var_{quantiles}(E[Y_i^a|quantile]) = Var(\bar{Y}_i^a) \approx Var(h^a(x_i))$. ■

A.2 Description of the Personalization Methods

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

OLS. In this personalization 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:

$$Y_i = \beta X_i + \gamma_a T_i + \delta_a X_i \times T_i + \varepsilon_i \quad (13)$$

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

$$\widehat{Y}_i^a = \hat{\beta} X_i + \hat{\gamma}_a a + \hat{\delta}_a X_i \times a \quad (14)$$

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

$$\pi_{OLS}(X_i) = \arg \max_a \widehat{Y}_i^a \quad (15)$$

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:

$$Y_i = f(X_i, T_i) + \varepsilon_i \quad (16)$$

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

$$\pi_{S-XGB}(X_i) = \arg \max_a \hat{f}(X_i, a) \quad (17)$$

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:

$$Y_i = f_a(X_i) + \varepsilon_i \quad (18)$$

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:

$$\pi_{T-XGB}(X_i) = \arg \max_a \hat{f}_a(X_i) \quad (19)$$

Multi-arm Causal Forest. Since the objective of the previous methods is a prediction of outcome levels, they are not necessarily optimal for uncovering heterogeneity (Athey and Imbens, 2016), which is necessary for personalization. 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 $\widehat{\tau}_{X_i}^a$ — 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 $\widehat{\tau}_{X_i}^{a_0} = 0$):

$$\pi_{MACF}(X_i) = \arg \max_a \widehat{\tau}_{X_i}^a \quad (20)$$

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 personalization policy derived via a policy tree.

A.3 Additional Figures

Figure 13: Age distribution in Walmart and Penn-Geisinger studies

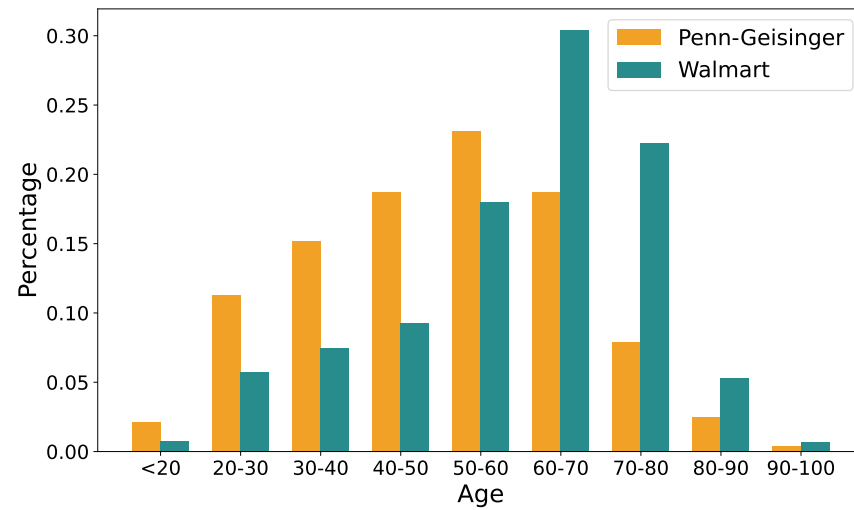


Figure 14: Zipcode median income distribution in Walmart and Penn-Geisinger studies

