Machine Learning Who to Nudge: Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal

Susan Athey Niall Keleher Jann Spiess

October 12, 2023

Abstract

In many settings, interventions may be more effective for some individuals than others, so that targeting interventions may be beneficial. We analyze the value of targeting in the context of a large-scale field experiment with over 53,000 college students, where the goal was to use "nudges" to encourage students to renew their financial-aid applications before a non-binding deadline. We begin with baseline approaches to targeting. First, we target based on a causal forest that estimates heterogeneous treatment effects and then assigns students to treatment according to those estimated to have the highest treatment effects. Next, we evaluate two alternative targeting policies, one targeting students with low predicted probability of renewing financial aid in the absence of the treatment, the other targeting those with high probability. The predicted baseline outcome is not the ideal criterion for targeting, nor is it a priori clear whether to prioritize low, high, or intermediate predicted probability. Nonetheless, targeting on low baseline outcomes is common in practice, for example because the relationship between individual characteristics and treatment effects is often difficult or impossible to estimate with historical data. We propose hybrid approaches that incorporate the strengths of both predictive approaches (accurate estimation) and causal approaches (correct criterion); we show that targeting *intermediate* baseline outcomes is most effective, while targeting based on low baseline outcomes is detrimental. In one year of the experiment, nudging all students improved early filing by an average of 6.4 percentage points over a baseline average of 37% filing, and we estimate that targeting half of the students using our preferred policy attains around 75% of this benefit.

Susan Athey, Stanford University, Stanford, CA, USA, athey@susanathey.com; Niall Keleher, Innovations for Poverty Action, New Haven, CT, USA, niall.keleher@gmail.com; Jann Spiess, Stanford University, Stanford, CA, USA, jspiess@stanford.edu. We thankfully acknowledge support from the Alfred P. Sloan Foundation and Schmidt Futures through the "Computational Applications for Behavioral Science" project with ideas42. Matthew Schaelling, Janelle Nelson, and Keshav Agrawal provided exceptional research assistance. For their support we also thank Octavio Medina, Rebecca Nissan, Rachel Rosenberg, and Josh Wright of ideas42, Vitor Hadad and Henrike Steimer during their time as postdocs at the Golub Capital Social Impact Lab at Stanford GSB, and the Office of Institutional Research and Assessment of the City University of New York. For helpful comments and discussions we are grateful to Michal Kolesár and seminar audiences at Stanford, ETH Zurich, and the Machine Learning in Economics Summer Institute.

1 Introduction

A growing number of randomized experiments set out to measure the effectiveness of behaviorally-informed nudges. Typically, these experiments are designed and analyzed to assess whether a nudge works well on average. In this article, we utilize causal machine learning to move beyond average treatment effects towards optimal targeting of nudges. In a large-scale experiment that randomized behaviorally-informed reminders to increase student financial-aid renewal applications, we estimate not just whether the nudge worked on average but also whether it worked for some students better than others. We then ask how such heterogeneous treatment effect estimates can improve the effectiveness of policy interventions.

Our application considers data from a field experiment among over 53,000 college students. The experiment aimed to measure the causal effect of behavioral nudges on timely applications for financial aid. Across two randomized controlled trials run in 2017 and 2018 by ideas42 and the City University of New York, enrolled students were randomly assigned to receive behaviorally informed text and email reminders about renewing their federal financial aid. The average treatment effect of the behavioral nudges was noteworthy. Students who received nudges were on average 6.4 ± 0.6 (2017) and 12.1 ± 0.7 (2018) percentage points more likely to submit their Free Application for Federal Student Aid (FAFSA) forms by the priority deadline, increasing early filing rates from 37% to 43% and 38% to 50%, respectively (as previously reported in Nissan et al., 2020).

Our goal is to estimate for whom the nudges are most effective, and thus whom to target if there is a limited budget of time or attention that creates opportunity cost associated with the nudges. For example, alternative messages might be used or developed for those where the nudges are less effective. The problem of whom to target for an intervention arises in many settings, ranging from precision medicine to prioritization of salespeople to allocation of advertising spend. A common approach in practice is based on predicting which individuals are most at risk of some undesirable outcome, such as customer churn or a poor medical outcome. Predictive approaches are attractive because they can be applied with observational, historical data without the need to run an experiment; they can be used even for a treatment that has never been tried before, or for which limited or non-representative historical data is available. In the context of nudges to file financial aid forms, we could imagine forming a hypothesis about which type of student would be most influenced by a nudge and building a model to predict that outcome using data about student behavior in the absence of the treatment. If we hypothesized that students who were otherwise at risk of not filing would be most influenced by the nudge, we would target the students with the lowest predicted probability of filing in the absence of the treatment.

However, in general, it is an empirical question as to what type of student is most likely to be influenced and what factors drive heterogeneity in treatment effects. In this article, we use causal machine learning to estimate how treatment effects vary with individual characteristics, that is, we estimate conditional average treatment effects (CATEs). We can predict differences in the response to receiving the nudge based on information available before the nudges are sent out. We identify subgroups across which differences in treatment effects are statistically significant, and the magnitudes of differences can be as large as a factor of 2. We estimate the difference between those with below-median predicted effect and those with above-median predicted effect to be around three to five percentage points. We find that enrollment status, once it becomes available, is highly predictive of treatment effects. Students who are unenrolled at the time of the behavioral nudge campaign had smaller treatment effects. However, we also uncover additional variation in CATEs among the enrolled students.

We then estimate the benefits of a policy that assigns students with the highest estimated CATEs to receive the nudge (counterfactually, under a budget constraint, which may be financial or based on opportunity cost). We find that non-parametric estimates of the CATE are noisy, so we explore several hybrid models that incorporate predicted baseline outcomes and nonparametric estimates of CATEs as covariates in a parametric model. We find that both non-parametric and hybrid models do significantly better than a random policy, with hybrid models substantially outperforming the non-parametric model. A policy that targets those *least* likely to file does very poorly, while targeting those *most* likely performs well; but the best approach is to prioritize those with *intermediate* predicted baseline outcomes first, and those with low predicted baseline outcomes last.

Our results provide an example where naive risk-based targeting from a machine-learning prediction of outcomes (particularly, one that targeted those who might have seemed to need the nudges most) performs substantially worse than targeting based on estimated treatment effect heterogeneity. On the other hand, if we had correctly guessed that it was effective to target those who were already most likely to file, we would have achieved acceptable performance. Our findings thus highlight the value of augmenting machine-learning algorithms, which provide powerful prediction tools, with careful causal inference to tackle policy problems.

Our application also speaks to the value of modeling treatment effects carefully when treatment effect variation is moderate and noisy. A simple causal model based on baseline predictions substantially outperforms the fully non-parametric causal machine learning model, thereby substantially increasing the gain from personalization in the 2017 study year. This semi-parametric model estimates a simple logistic regression on top of a random-forest prediction of baseline filing probabilities, which is straightforward to implement as a prediction within the control group only. Treatment effects are assumed to be constant in log-odds (and thus U-shaped in baseline outcomes), and we do not find evidence that modeling the treatment effects in more complex ways substantially improves the performance of our targeting policy. In a semi-synthetic simulation study, we show that a hybrid model that incorporates both predicted baseline outcomes and non-parametric CATE estimates performs well in a variety of settings, where the value of incorporating the CATE estimates is greater when the signal-to-noise ratio is higher for CATE estimation.

Noisy estimates are common in empirical practice, especially when we aim to learn about treatment effect heterogeneity, and finding ways of improving efficiency can be very valuable in such cases. One way of improving efficiency is to impose simple functional forms as a form of regularization. In our application, we find that assuming that treatment effects on log-odds are constant provides an effective restriction on the relationship of treatment effects to the baseline probability of filing. At the same time, leaving the baseline unrestricted allows our simple logistic model to capture substantial heterogeneity. This approach is related to the insight in Athey et al. (2021) that it is often much easier to produce quality estimates of baseline responses, while estimating treatment effects directly may be noisy and therefore benefit from parametric models. In this sense, we extend the semiparametric strategy of Athey et al. (2021) from average to conditional treatment effects.

While the simple semi-parametric logistic regression model captures treatment effects well in our example, it has the disadvantage that it may perform poorly if the relationship between baseline outcomes and treatment effects is not a good fit with the logistic functional form. To address this, we estimate a non-parametric model of treatment effects as a function of baseline predictions. In general, this approach can capture flexible relationships between baseline predictions and treatment effects. In the FAFSA application, it performs on par with the parametric model. To the extent that it differs, it indicates that the inverted-U relationship between baseline outcomes and treatment effects is steeper and more pronounced than that implied by the logistic functional form; however, this does not matter much for prioritization of individuals for the treatment.

Our results suggest three general conclusions. First, they clarify the importance of integrating causal inference and randomized trials into machine learning to analyze and improve policy, rather than relying on predictive tools based on non-experimental baseline data alone. Second, we demonstrate the value of combining non-parametric predictive tools with simple econometric models for causal estimation and targeting. Third, the effects of treatments that are only small interventions, as is often the case for nudges, may not accrue mainly for those with low baseline outcomes, but rather for those who would already have been more likely to obtain a better outcome in the absence of treatment – and just need to be "nudged" over the finish line.

Our analysis uncovers important challenges in applying machine learning to improve the analysis and targeting of nudges. The environment we study has a fairly low signal-to-noise ratio, and we find that treatment effect estimates are unlikely to be well-calibrated. Thus, describing heterogeneity requires additional diagnostic tools to avoid small-sample biases, such as group-wise analysis discussed in Chernozhukov et al. (2019). We also move beyond describing the performance of machine-learning policies in terms of raw predictive power and instead provide analogs to receiver-operating characteristic (ROC) diagnostics adapted to the problem of treatment assignment, where the performance of the model is quantified in policy-relevant units (following e.g. Rzepakowski and Jaroszewicz, 2012; Zhao et al., 2013; Hitsch et al., 2023; Yadlowsky et al., 2021). In terms of evaluating different targeting policies and estimating their value from data with randomized treatment, our approach is very similar to Hitsch et al. (2023), which also provides a comparison to targeting based on purely predictive "scoring" models.

We build upon a growing literature that combines causal estimation with prediction techniques from machine learning (Mullainathan and Spiess, 2017; Athey and Imbens, 2019) and discusses the application of machine learning to policy problems (Kleinberg et al., 2015). Methods for estimating heterogeneous treatment effects using machine learning have been proposed in, among others, Imai and Ratkovic (2013); Athey and Imbens (2016); Wager and Athey (2018); Athey et al. (2019); Künzel et al. (2019); Nie et al. (2021). Chernozhukov et al. (2019) discusses the analysis of heterogeneous treatment effects with arbitrary machine-learning estimators and suggests diagnostic tools, while Yadlowsky et al. (2021) proposes metrics for assessing the benefits to targeting similar to those applied in this article. Athey and Wager (2021) and Zhou et al. (2023) analyze efficient estimation of targeted policies with constraints on the policy class. Hitsch et al. (2023); Knaus et al. (2022); Yang et al. (2020), among others, carry out empirical studies analyzing targeted policies derived using machine learning methods. Zhang and Misra (2022) proposes a two-step procedure involving treatment-effect estimation and discretization in order to decide which treatments to offer to which individuals.

Until recently, there was little empirical evidence about the relative performance of predictive and causal targeting. An early article to do so is Ascarza (2018), which studies the effect of a customer retention program on reducing churn in two field experiments and shows that targeting based on predicted churn performs considerably worse than targeting based on estimated treatment effects. A few subsequent studies in marketing have similar findings (e.g. Devriendt et al., 2021), where targeting based on treatment effects is referred to as "uplift modeling." Like us, Fernández-Loría and Provost (2022) compares targeting by negative and positive baseline to causal targeting, and shows that the former can be beneficial when causal targeting is noisy. The topic has also recently gained attention in development economics, where Haushofer et al. (2022) considers targeting by "impact" (treatment effect) vs "deprivation" (baseline), and in medicine, where Inoue et al. (2023) analyzes the tradeoff between "high-risk" and "high-benefit" approaches when targeting treatments for high blood pressure.

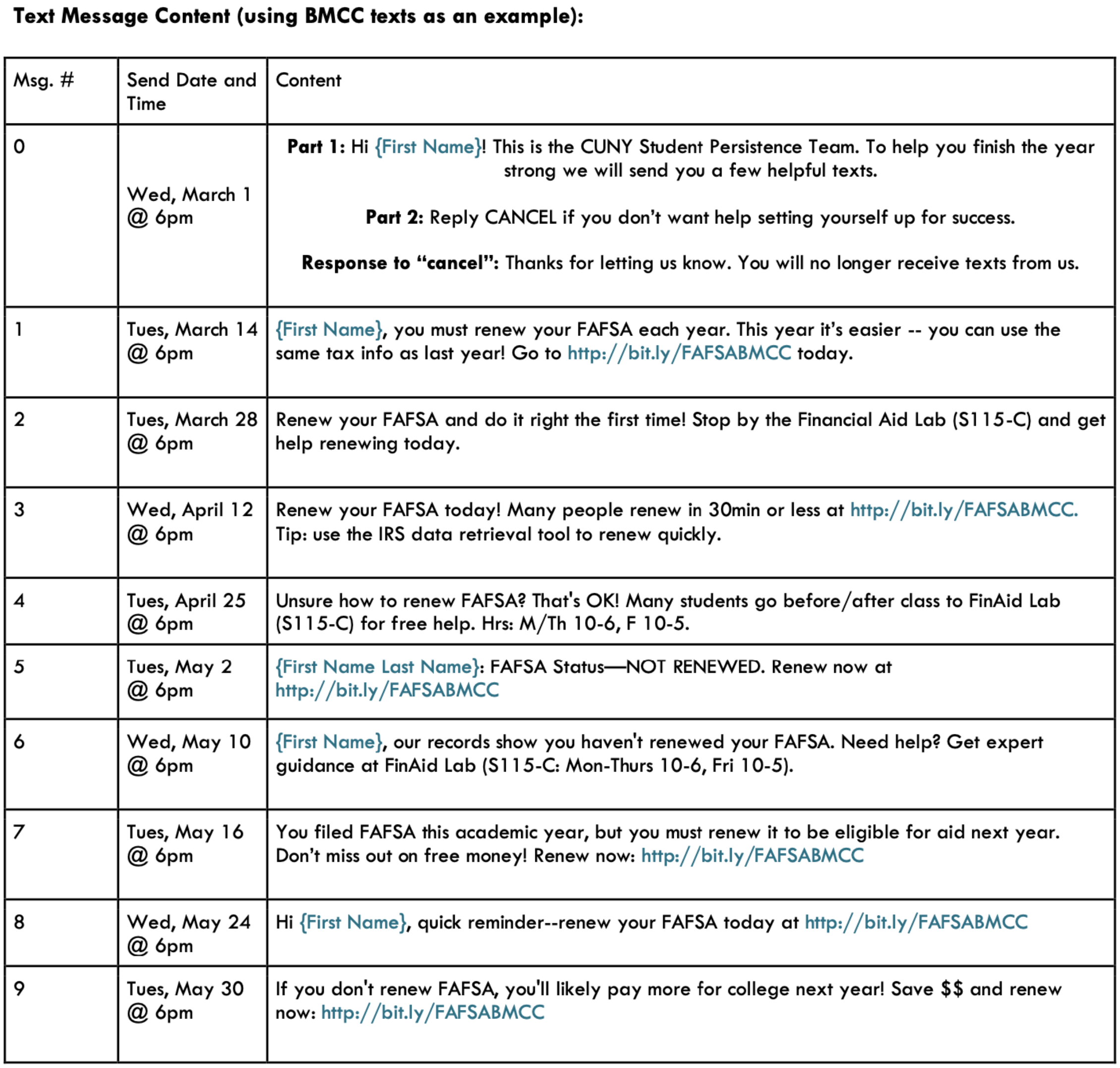
We also connect to a literature on behaviorally-informed nudges (Sunstein and Thaler, 2008) and their empirical validation. In the context of student financial aid, behavioral science has informed multiple cost-effective strategies for increasing FAFSA submissions that the experiment we analyze in this article is based on. Castleman and Page (2016) shows that a simple text-based intervention that encouraged FAFSA submission increased sophomore year retention by 14%. ideas42 research at Arizona State University showed that behaviorally informed student reminder emails, which include devices to trigger loss aversion, plan making, and commitment, increased priority deadline FAFSA renewal by 11 percentage points, from 29% to 40% (ideas42, 2016).

2 Experiment and Data

This article analyzes data from a multi-year experiment conducted in New York City. Students were randomly assigned to receive behaviorally informed text and email reminders to renew their federal financial aid. The field experiment, run in 2017 and 2018 by ideas42 and the City University of New York (CUNY), aimed at increasing applications for Free Application for Federal Student Aid (FAFSA) financial support by the June 30 priority deadline. Students randomly assigned to the control group received only business-as-usual emails from the college. Students assigned to the treatment groups also received supplementary behavioral emails and text messages. These emails and text messages were designed to trigger loss aversion, plan-making, and commitment.[1] Figure 1 shows example text messages sent to students in the treatment group.

The experiment involved matriculated students from CUNY community colleges. Eligible students were

[1] During 2017, the experiment had one treatment arm. The two treatment arms in the 2018 experiment differed in whether they used one-way texts or two-way texts that prompted students to respond. For this article, we pool the two treatment arms in the 2018 study.



Text Message Content (using BMCC texts as an example):

Msg. #	Send Date and Time	Content
0	Wed, March 1 @ 6pm	**Part 1:** Hi {First Name}! This is the CUNY Student Persistence Team. To help you finish the year strong we will send you a few helpful texts. **Part 2:** Reply CANCEL if you don't want help setting yourself up for success. **Response to "cancel":** Thanks for letting us know. You will no longer receive texts from us.
1	Tues, March 14 @ 6pm	{First Name}, you must renew your FAFSA each year. This year it's easier -- you can use the same tax info as last year! Go to http://bit.ly/FAFSABMCC today.
2	Tues, March 28 @ 6pm	Renew your FAFSA and do it right the first time! Stop by the Financial Aid Lab (S115-C) and get help renewing today.
3	Wed, April 12 @ 6pm	Renew your FAFSA today! Many people renew in 30min or less at http://bit.ly/FAFSABMCC. Tip: use the IRS data retrieval tool to renew quickly.
4	Tues, April 25 @ 6pm	Unsure how to renew FAFSA? That's OK! Many students go before/after class to FinAid Lab (S115-C) for free help. Hrs: M/Th 10-6, F 10-5.
5	Tues, May 2 @ 6pm	{First Name Last Name}: FAFSA Status—NOT RENEWED. Renew now at http://bit.ly/FAFSABMCC
6	Wed, May 10 @ 6pm	{First Name}, our records show you haven't renewed your FAFSA. Need help? Get expert guidance at FinAid Lab (S115-C: Mon-Thurs 10-6, Fri 10-5).
7	Tues, May 16 @ 6pm	You filed FAFSA this academic year, but you must renew it to be eligible for aid next year. Don't miss out on free money! Renew now: http://bit.ly/FAFSABMCC
8	Wed, May 24 @ 6pm	Hi {First Name}, quick reminder--renew your FAFSA today at http://bit.ly/FAFSABMCC
9	Tues, May 30 @ 6pm	If you don't renew FAFSA, you'll likely pay more for college next year! Save $$ and renew now: http://bit.ly/FAFSABMCC

Figure 1: Example reminder text messages sent to students in the treatment group.

those who had not yet renewed FAFSA in February of the study year. The 2017 study sample includes 25,167 students from three community colleges, of which 50% were randomly assigned to treatment. The 2018 sample includes 40,638 students from five community colleges, which were included in the intervention in two batches: an early batch of 30,627, of which 45% were assigned to each of the two treatment arms and 10% to control, and a late batch of 10,011 with a larger control group of 25% and roughly equal treatment groups. We pool the late and early cohorts from 2018 for a total combined fraction of 86% treated across the two treatment arms. Throughout our analysis of 2018 data, we adjust estimates for the varying propensity scores between early and late cohorts by inverse probability-weighted estimators.

Our data include baseline demographic, academic, and administrative information about the community-college students in the experiment. On average, students were around 24 years old, with a considerable standard deviation of almost seven years. Our sample includes more women than men, with 57% women in the 2017 experiment, as well as 56% (early schools) and 53% (late schools) women in 2018. A plurality of students was Hispanic (52% in 2017, 45% in 2018), followed by Black non-Hispanic students who made up around a third of the student body in this study. Almost 20% of students were enrolled part-time. Overall, we do not observe large imbalances between treatment and control groups; for nine baseline characteristics we tested across the 2017 and 2018 cohorts, only one variable, GPA for late 2018 schools, is significantly different between treatment and control at the 5% level. Estimated propensity scores are concentrated around their batch-wise mean and balanced between the respective treatment and control groups. Details are available in Tables 2 and 3 in the appendix.

Across the two randomized controlled trials, those who received the treatment interventions were on average 6.4±0.6 (2017) and 12.1±0.7 (2018) percentage points more likely to submit their FAFSA forms by the priority deadline, increasing early filing rates from 37% to 43% and 38% to 50%, respectively. These estimates, which are based on simple averages between treatment and control units within batches, are robust with respect to two alternatives, augmented inverse propensity weighted (AIPW) estimators that leverage covariate information to reduce noise (see Table 4 in the appendix for details). The first of these estimators assumes constant propensity scores within batches, while the second corrects for possible imbalances by estimating the propensity score. Both are based on random forest estimation of the outcome model and propensity score.

3 Estimating Treatment-Effect Heterogeneity

Above, we noted a sizable average effect of behaviorally-informed reminders in this study of increasing FAFSA filing rates by the priority deadline by 6.4 ± 0.6 (2017) and 12.1 ± 0.7 (2018) percentage points. In this section, we use machine-learning tools to estimate treatment effects as a function of available individual covariates. We repeat the analysis for each study year (2017 and 2018) as well as for two sets of explanatory covariates – first, those available at the time of randomization before the spring semester, and second, all information available halfway through the semester (which adds enrollment information and additional academic records) just before reminders were sent out. We report tests and diagnostics based on an analysis of the full sample of 25,167 students in 2017 and 40,638 in 2018.

Our goal in this section is to estimate conditional average treatment effects (CATEs), which are defined as average treatment effects conditional on observed covariates. Before describing how we estimate these treatment effects, we formally define the object of interest. We denote by $Y \in \{1, 0\}$ the random variable that expresses whether a student has filed by the priority deadline ($Y{=}1$) or not ($Y = 0$), and $T \in \{1, 0\}$ for whether the student is in the treatment group ($T{=}1$) or not ($T{=}0$). We use standard potential-outcomes notation and write $Y(1)$ for the filing status a student would have had had they been assigned to treatment, and $Y(0)$ for the filing status had they been assigned to control. The treatment of that student is then $Y(1) - Y(0)$, and we are interested in how this treatment effect varies with some baseline student characteristics X. Specifically, we aim to estimate the conditional average treatment effect

$$\tau(x) = \mathrm{E}[Y(1) - Y(0)|X{=}x]$$

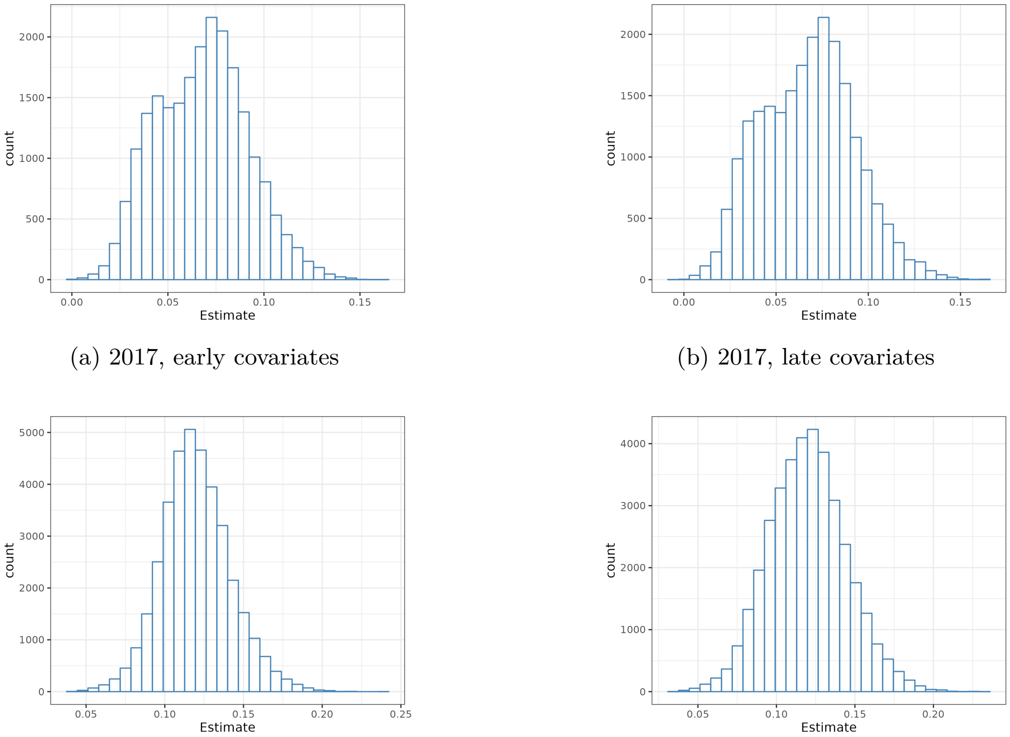
of students with characteristics $X{=}x$. When estimating $\tau(x)$, we face the challenge that for every student we only observe one of the potential outcomes $Y(1)$ and $Y(0)$, namely the realized filing decision $Y = Y(T)$ for their actual treatment status T. However, since treatment has been randomized, the realization of $Y(1)$, $Y(0)$ and X are independent of T, so we can identify treatment effects from $\tau(x) = \mathrm{E}[Y|T{=}1, X{=}x] - \mathrm{E}[Y|T{=}0, X{=}x]$. In words, while we cannot compare outcomes within student, we can compare outcomes across similar students who have been treated or assigned to control, which yields the same conditional average effect when treatment is assigned randomly.

We employ the causal forest algorithm, an instance of generalized random forests (Athey et al., 2019) that is specifically adapted to solve the causal-inference problem of estimating CATEs ($\tau(x)$) in settings like ours. Causal forests recursively compute multiple partitions of the covariate space based on treatment effect heterogeneity, so that estimated average treatment effects vary as much as possible between subsets of the covariate space. To estimate the CATE $\tau(x)$ for a particular vector x of covariates, a weighted average treatment effect

$$\hat{\tau}(x) = \frac{\sum_{T_j=1} \hat{w}_j(x) Y_j}{\sum_{T_j=1} \hat{w}_j(x)} - \frac{\sum_{T_j=0} \hat{w}_j(x) Y_j}{\sum_{T_j=0} \hat{w}_j(x)}$$

of nearby observations (Y_j, T_j, X_j) is used, where weights $\hat{w}_j(x)$ are based on how often observations X_j share the same cell as the target vector x in different partitions. Moreover, because the process of constructing the partition to estimate treatment effects uses a careful rule for sample splitting ("honesty," Wager and Athey, 2018) that ensures that Y_i is not used in the estimation of $\hat{w}_j(X_i)$, the resulting estimates are guaranteed to be consistent and asymptotically normal.

Figure 2 shows the distribution of estimated treatment effects across the two study years we consider (2017, top, and 2018, bottom), differentiating between covariates available before the beginning of the semester ("early covariates", left) and those only available later, but still before the interventions ("late covariates", right). The histogram shows treatment effects ranging from around 0 to 15 percentage points in the 2017 data, and from around 5 to 20 percentage points for 2018. Heterogeneity seems to be similar between early and late covariates for 2017. In 2018, treatment effect estimates using the late covariates are somewhat more spread out than those using information from before the start of the



(a) 2017, early covariates

(b) 2017, late covariates

(c) 2018, early covariates

(d) 2018, late covariates

Figure 2: Histograms of estimated treatment effects across years and set of covariates, using honest estimates from the causal forest method.

semester only. We do not find any evidence for negative treatment effects, so reminders are unlikely to have caused any students not to file by the priority deadline.

Despite the theoretical guarantees for the honest treatment-effect estimates, in practice, the signal-to-noise ratio is often such that a very large sample is required for the theory to be a useful guide. For more realistic sample sizes, estimates $\hat{\tau}(x)$ of CATEs $\tau(x)$ are often miscalibrated. For a particular covariate vector x, the estimates $\hat{\tau}(x)$ may be biased towards the overall average treatment effect $\tau = \mathrm{E}[Y(1) - Y(0)]$, as there will not, in general, be enough observations with similar covariate vectors to a particular target. On the other hand, in a setting with very little true heterogeneity relative to the noise, sampling variation will still induce a distribution of estimated treatment effects over different covariate vectors, potentially overstating heterogeneity. We do not rely on the theoretical guarantees about the estimator in this article; rather, we use our estimates of CATEs as an input to other analyses, such as constructing policies, and evaluate those policies using model-free methods.

In order to assess the importance of treatment effect heterogeneity, we first provide a calibration-based analysis of the estimates of heterogeneous treatment effects based on Chernozhukov et al. (2019). Across both years and both sets of covariates, those available earlier and those later, we report results in Table 1 of a cross-fitted calibration regression of actual outcomes Y_i on treatment-effect estimates $\hat{\tau}(X_i)$ interacted with normalized treatment $T_i - p$, where p is the overall probability of being treated (Table 1).[2]. If treatment effects were perfectly calibrated, we would expect the slope estimate to be close to one, while it would be close to zero if treatment effect estimates are purely spurious. Our estimates in Table 1 are significantly larger than zero (at the 5% significance level), providing evidence that our estimates indeed do capture treatment-effect heterogeneity. For robustness, we present results based on two separate approaches to estimating heterogeneous treatment effects, namely using the fact that we know that the propensity score is constant across groups of schools ("known") or estimating the propensity score from the data to be robust against possible issues with randomization ("AIPW"). Although we find evidence of heterogeneity across these specifications, the model is not perfectly calibrated (the slope coefficients are substantially less than one), and as such we cannot assume that the magnitudes of our CATE estimates $\hat{\tau}(x)$ are unbiased for $\tau(x)$.

Even if CATE estimates are miscalibrated and thus cannot be taken at face value, they may still be reliable for assessing which units have higher treatment effects than others. While it is impossible to assess the accuracy of a treatment effect estimate for a single observation, since we can only observe the outcome Y for an individual with one of the two possible treatment assignments $T \in \{1, 0\}$, we can construct an unbiased estimate of the average treatment effect $\mathrm{E}[Y(1) - Y(0)|G]$ for a sufficiently large group of individuals, where the group $G = g(X)$ is defined by covariates X. Following the "Sorted Group Average Treatment Effects (GATES)" methodology of Chernozhukov et al. (2018), we proceed by creating such groups based on our CATE estimates. We make use of ten-fold cross-fitting to remove bias, as follows. We partition the data into folds, and let $k(i)$ be the fold that leaves out observation i. For each fold $k \in \{1, \ldots, 10\}$, we estimate a mapping $\hat{\tau}_{-k}$ from covariates x to the CATE using

[2] To avoid biases, we estimate treatment effects $\hat{\tau}(X_i)$ for units (Y_i, T_i, X_i) within our sample by ten-fold cross-fitting. Specifically, we randomly divide the sample into ten folds, and for an observation i in a given fold estimate their CATE $\tau(X_i)$ by $\hat{\tau}(X_i)$ using a causal forest $\hat{\tau}$ fitted only on the other folds. We then run the calibration regression by fold and aggregate the resulting coefficient and standard error estimates.

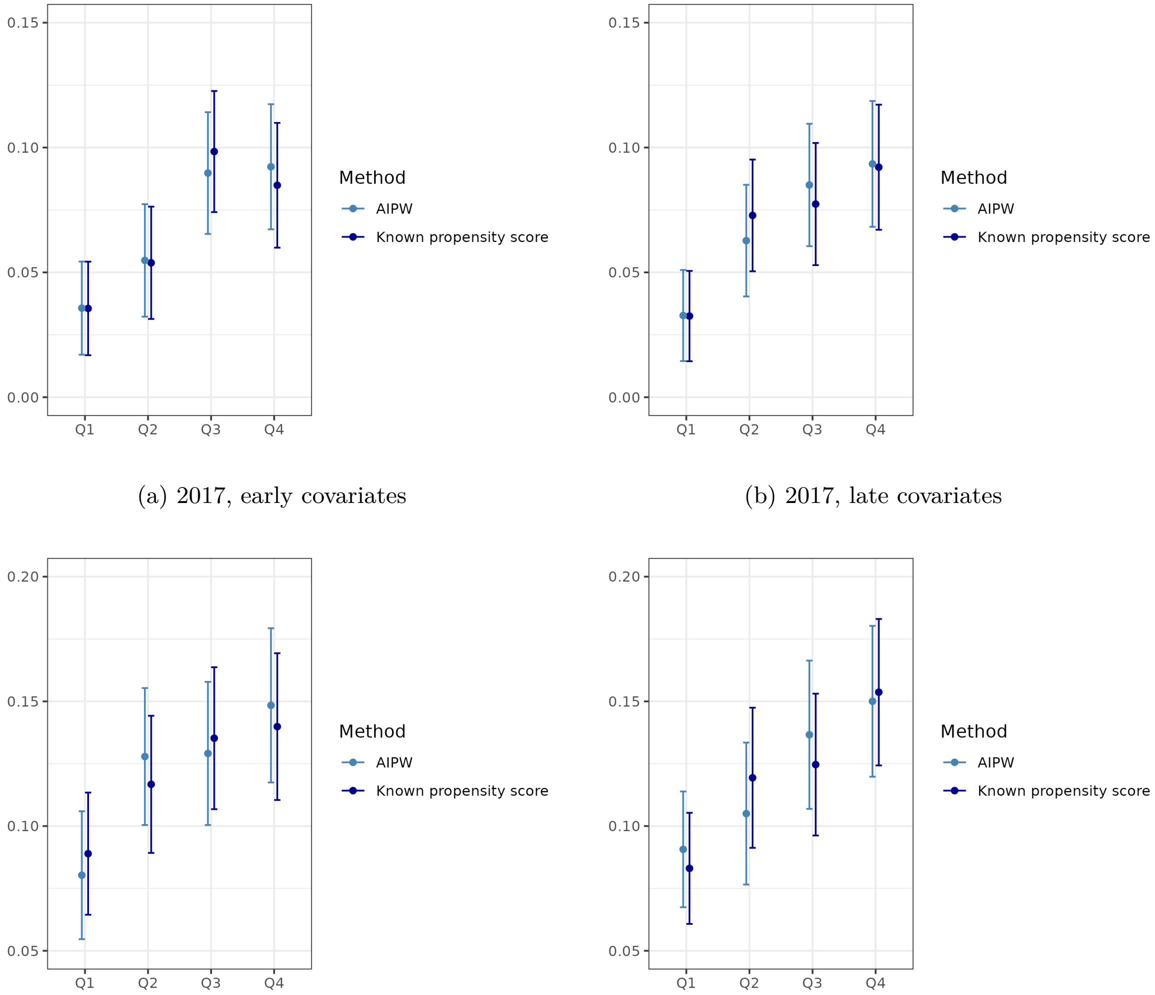
Year	Covariates	Propensity score	Slope estimate	SE	p-value
2017	early	known	0.76171	0.23041	0.00047
		AIPW	0.74504	0.22908	0.00057
	late	known	0.78549	0.21804	0.00016
		AIPW	0.62928	0.21639	0.00182
2018	early	known	0.55267	0.29159	0.02902
		AIPW	0.65692	0.28802	0.01128
	late	known	0.75689	0.25451	0.00147
		AIPW	0.67651	0.25451	0.00393

Table 1: Slope coefficient estimates for the calibration regression of actual outcomes on treatment-effect estimates interacted with normalized treatment following Chernozhukov et al. (2019).

data that leaves out observations in fold k. Within each fold k, we then divide the covariate space into four groups $G \in \{1, 2, 3, 4\}$ based on the quartile of estimated treatment effects $\hat{\tau}_{-k}(x)$. Finally, we estimate average treatment effects $\mathrm{E}[Y_i(1) - Y_i(0)|G = g]$ for each of the four groups g by the average difference between treated and control outcomes within that group, combining groups across all folds. These estimates are unbiased estimates of the average treatment effect for the groups (recalling groups are defined by the covariates), since the outcomes of the units in fold k were not used in any part of the process of assigning units to groups. The resulting estimates are model-free, unbiased estimates of average treatment effects in each group.

Our resulting estimates of treatment effects are noisy, and the heterogeneity across groups is moderate. Figure 3 plots unbiased estimates of the group-wise average treatment effect across quartiles of estimated treatment effects for the 2017 (top) and 2018 (bottom) cohort, differentiating between covariates available before the beginning of the semester (left) and those only available later, but still before the interventions (right). Both graphs document that treatment effects $\tau(x)$ vary across the distribution, although the forest-based estimates $\hat{\tau}(x)$ differ from unbiased estimates and even produce non-monotonic rankings in one instance. Across both years and irrespective of the set of covariates, we find statistically significant heterogeneity in treatment effects. For 2017, the average treatment effect in the two bottom quartiles (early covariates) or in the bottom quartile (late covariates) is significantly below that of the remaining sample, with a difference of around five percentage points separating the respective group averages. For 2018, the average outcome in the bottom quartile is around 7 percentage points lower than the average among the rest. Effects for the set of covariates available earlier are not considerably noisier. The differences across quartile AIPW estimates are reported in Table 5 in the appendix along with standard error estimates.

We next inspect which variables drive treatment effect heterogeneity. To do so, we report a simple variable-importance measure based on the causal forest used for estimating heterogeneous treatment effects. This variable importance measures counts how often each variable is used in a split within the causal forest, and is thus a way of assessing the contribution of a variable to estimating heterogeneous treatment effects. The top variables identified in this way across both years and early and late data include age, GPA, and how many credits a student attempted and actually earned. While suggestive of



(a) 2017, early covariates

(b) 2017, late covariates

(c) 2018, early covariates

(d) 2018, late covariates

Figure 3: Average treatment effects by quartiles of estimated treatment effects. The x-axis divides the sample into quartiles of predicted cross-fitted treatment effects using ten folds. The y-axis plots groups and estimates based on an augmented inverse-propensity weighted estimator using an estimated ("AIPW") or the known propensity score ("Known propensity score"), along with a 95% confidence interval.

drivers of heterogeneity, this measure tells us neither the direction nor the degree to which treatment effects are affected. It may also understate the importance of variables with only a few levels. Some binary variables, for example, may have a large impact on treatment effects, but some partitions may split on them only once.

One potential driver of heterogeneity is whether students continue to be enrolled. There is no enrollment restriction on who can apply for FAFSA, and students could unenroll in a given year yet remain eligible to apply for FAFSA for the subsequent academic year. However, students who drop out midyear may not only be harder for administrators to track but they may also be less likely to re-enroll for the subsequent academic year. For the randomized experiment, unenrolled students were still eligible for behavioral nudges. Using the later set of covariates that become available during the semester, we can identify students that were enrolled at the start of the academic year, yet dropped out by the time of the behavioral nudge treatment.

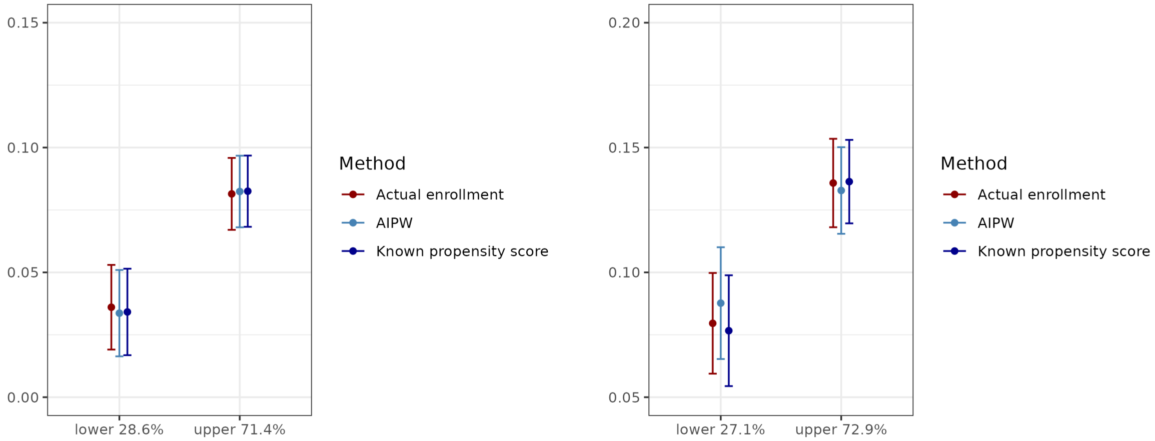
We find that enrollment status, once it becomes available, is indeed highly predictive of treatment effects. Indeed, if we calculate average treatment effects across enrollment status, we find that it partitions treatment effects just as well as if we had formed two groups of students based on our treatment-effect estimates (Figure 4), holding the size of the two groups fixed. However, there seems to be additional heterogeneity in treatment effects even among those that are enrolled. Comparing effects across the top three quartiles on the right side of Figure 3 to the average effect of enrolled students in Figure 4 suggests that the highest quartile of estimated effects has a higher treatment effect than enrolled students overall, although the difference is noisy.

We note that the results on enrollment as an important treatment-effect moderator are intuitive, but not mechanical. Indeed, reminders affect the filing of both students who are enrolled and who are not enrolled at the time enrollment is measured for the spring semester. Since students may drop out and re-enroll, it remains effective to provide reminders for unenrolled students. Rather, we consider it an interesting finding that there seems to be only small additional heterogeneity, making enrollment a powerful proxy for the effectiveness of reminders already by itself.

4 Causal vs Predictive Targeting

In the previous section, we found that there is heterogeneity in students' responses to the reminder. We now ask how we can use this predictable variation in the response to reminders to improve their targeting. We start this section by leveraging the heterogeneous-treatment effect estimates $\hat{\tau}(x)$ we constructed above to target reminders, which is closely related to the approach in Hitsch et al. (2023) towards estimating and evaluating targeting policies. We then compare the performance of this policy to a policy purely based on predictions of the probability of filing at baseline. Throughout, we focus on the 2017 cohort of students, where we have a large control group available.

One finding of our analysis of heterogeneous treatment effects was that a good part of the heterogeneity in conditional average treatment effects $\tau(x)$ can be predicted by enrollment once we know whether a student has dropped out. Hence, one natural way of targeting reminders that we investigate below is to



(a) 2017

(b) 2018

Figure 4: Average treatment effects by enrollment status (red) and by whether predicted cross-fitted treatment effects are below or above the quantile corresponding to the proportion of enrolled students (blue), using all data up to the start of the intervention. The y-axis plots a simple difference-in-averages for actual enrollment as well as two estimators based on augmented inverse-propensity weighted estimators using the estimated ("AIPW") or the known propensity score ("Known propensity score") constant and with estimated propensity scores, along with 95% confidence intervals.

send them only to enrolled students. Notwithstanding this finding, there are two reasons why targeting based on individual covariates is still meaningful: First, we ask whether we could have used predictions of heterogeneous treatment effects to select students to send behaviorally-informed reminders before the beginning of the semester, when enrollment information was not yet available. Second, even among those students who are known to have enrolled, there is still some heterogeneity in the response to reminders.

4.1 Targeting based on non-parametric treatment effect estimates

We start by evaluating the value of the estimates $\hat{\tau}(x)$ of heterogeneous treatment effects from Section 3 for targeting reminders better. That is, rather than estimating the quality of our predictions $\hat{\tau}(x)$ in terms of how well they estimate true treatment effects $\tau(x)$, we quantify which proportion of total gain from the reminders we could have realized from targeting only a selected fraction of students. Specifically, we consider assignment policies

$$\hat{\pi}_t^{\text{causal}}(x) = \mathbb{1}(\hat{\tau}(x) \geq t) \tag{1}$$

that assigning all students to treatment whose estimated treatment effect is above some threshold t.

The results of our assignment exercise are summarized in Figure 5, separately for leveraging early covariates (those available before the beginning of the semester) and late covariates (those available by the time the reminders are sent) for the 2017 cohort of the FAFSA experiment. Similar to the approach taken in Hitsch et al. (2023) for evaluating targeting policies, we plot (under the counterfactual policy

(1)) the estimated fraction of students who file by the priority deadline against the fraction t of students who are assigned to treatment. Our main benchmark is a random assignment policy that chooses a random group of students to receive reminders, leading to a linear relationship between the fraction of students assigned to treatment and the estimated fraction of students who file (black line in Figure 5). The value of using heterogeneous treatment effects estimates $\hat{\tau}(x)$ from the causal forest is represented by the "CATE" line, which shows a persistent gain over the random assignment policy.[3]

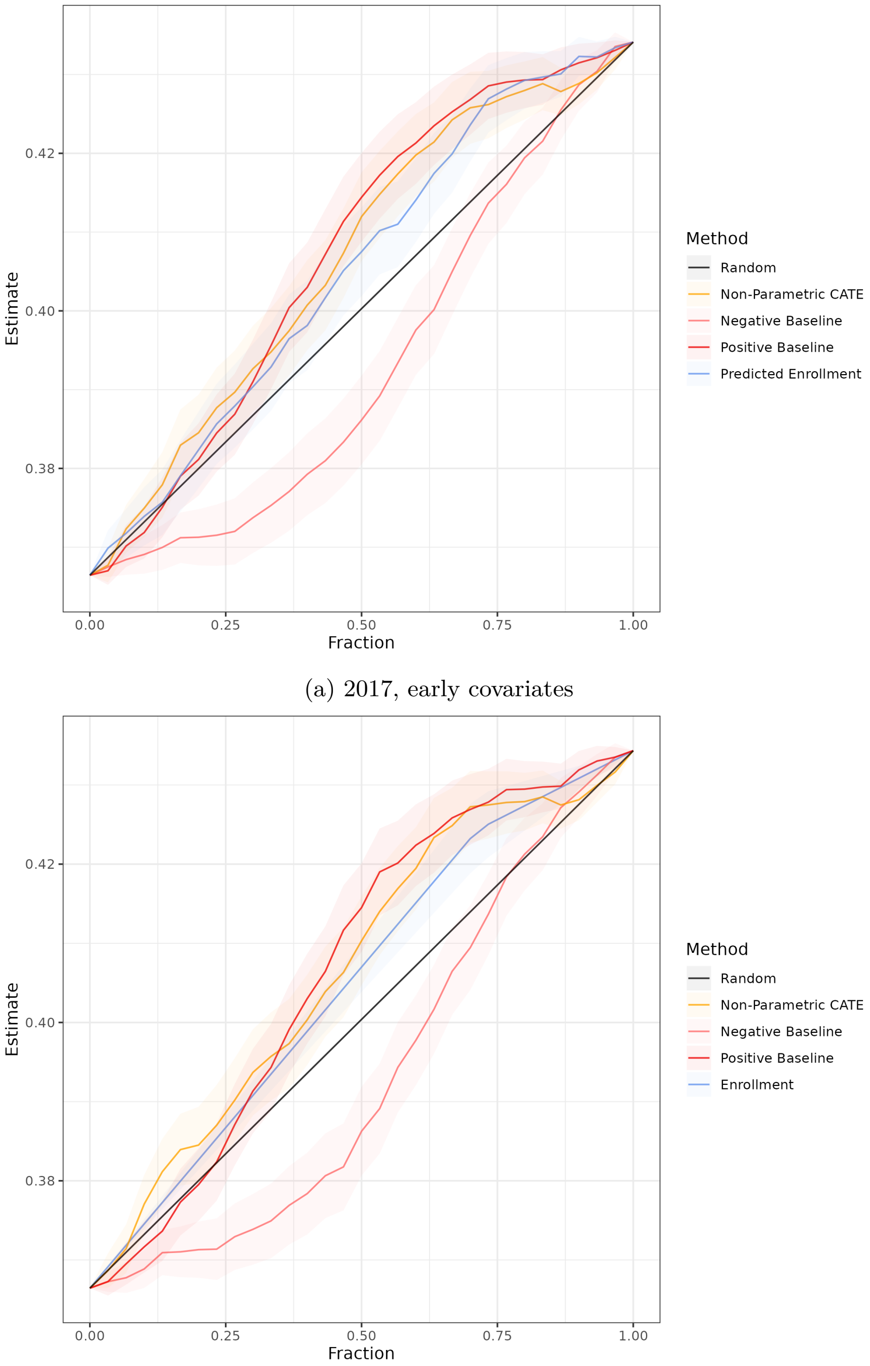
In order to estimate the numbers in Figure 5, we leverage the insight, which is also pointed out by Hitsch et al. (2023), that the outcome under alternative assignment policies can be estimated from existing trial data by exploiting randomized assignment. This permits the estimation of $\mathrm{E}[Y(\pi(X))] = \mathrm{E}[\mathrm{E}[Y|T{=}1, X]\ \pi(X) + \mathrm{E}[Y|T{=}1, X]\ (1 - \pi(X))]$ for an assignment policy that maps characteristics x to an assignment $\pi(x) \in \{1, 0\}$. For each fraction, q, we estimate (using ten-fold cross-fitting) the average outcome $U^{\text{causal}}(q) = \mathrm{E}[Y(\hat{\pi}^{\text{causal}}_{t(q)}(X))]$ we could have achieved using this policy, where the threshold $t(q)$ is chosen such that a fraction $q \in [0, 1]$ of individuals receives the treatment. For every fraction q, Figure 5 plots the resulting average estimate $\hat{U}^{\text{causal}}_k(q)$ of $U^{\text{causal}}_t(q)$ across folds k on the y-axis based on within-fold rankings using cross-fitted treatment effect estimates $\hat{\tau}_{-k}(X_i)$, allowing us to evaluate and compare policies based on the total increase in FAFSA renewal relative to the number of students who are sent reminders.[4] We obtain unbiased and model-free estimates using the augmented inverse propensity weighted estimators using fixed (known) propensity scores, with details provided in Appendix B.

In economic terms, our estimates suggest a moderate gain from targeting using treatment-effect estimates from the causal forest. For example, we could have increased FAFSA renewal at the priority deadline from 36.5% to 41% (realizing 65% of the gain) by targeting those 50% of students with the highest predicted effect based using early covariates, with basically the same performance for late covariates. In statistical terms, the point-wise 95% confidence bands in Figure 5 document a significant gain over random assignment for a substantial range of fractions of assigned students. However, since these confidence bands are estimated point-wise, we complement our analysis with a joint test based on the RATE framework of Yadlowsky et al. (2021). Figure 9 in the appendix shows average treatment effects by targeting fraction, and Table 6 provides corresponding estimates of the area under this curve. An omnibus test of average gain over random assignment based on these estimates (which is also a test of the null hypothesis of no heterogeneity) rejects at the 5% level for early and late covariates in the 2017 sample, and also for late covariates in the 2018 data.

We benchmark the performance of targeting based on causal-forest estimates of heterogeneous treatment effects against targeting based on predicted and realized enrollment. In Section 3 we argued

[3] In the main text, we discuss estimates of the value of targeting based on the robust cross-fitting estimator from Appendix B with constant propensity scores, applied to assignments based on CATE estimates obtained similarly to Section 3 but with tuning parameters chosen to optimize targeting performance. In the appendix, we provide additional robustness checks. Specifically, Figure 11 changes how the same policies are estimated, by separating the estimation of the CATE used for targeting from the estimation of nuisance components used in evaluation. Figure 12 evaluates the targeting performance of the causal forest across different tuning parameter values, specifically varying the minimum node size and the number of trees. Our results appear largely insensitive to these changes, and the chosen tuning parameters seem to perform well.

[4] Such policy ROC curves (also called "uplift curve", "profit curve", or "cost curve") have also been used to represent benefits of targeting at varying costs e.g. in Rzepakowski and Jaroszewicz (2012); Zhao et al. (2013); Sun et al. (2021); Yadlowsky et al. (2021); Hitsch et al. (2023).



(a) 2017, early covariates

(b) 2017, late covariates

Figure 5: Total estimated FAFSA renewal rate (y-axis) by targeting a given fraction (x-axis) of students according to different cross-fitted predictions in the 2017 data, including targeting by estimated treatment effects using the causal forest ("Non-Parametric CATE"), by a random-forest prediction of outcomes absent treatment ("Negative Baseline" for low baseline treated first, and "Positive Baseline" the reverse), and by predicted or actual enrollment ("Predicted Enrollment" and "Enrollment"). Shown are model-free unbiased augmented inverse propensity weighted estimates with 95% confidence intervals that represent the point-wise uncertainty of the difference in renewal rate relative to the random policy ("Random") that assigns the same fraction to treatment, with details provided in Appendix B.

that a good fraction of heterogeneity in treatment effects can be attributed to enrollment. Here, we therefore consider what happens if we prioritize reminders to go to those students who have the highest predicted probability of being enrolled. The "Predicted Enrollment" method in Panel (a) of Figure 5 evaluates that policy, and shows that it does not perform as well as the causal policy, suggesting additional variation in treatment effects beyond predicted enrollment, although differences are noisy. In Panel (b), we instead target by realized enrollment, which is available as part of the late covariates. That policy has a similar performance to predicted enrollment.

4.2 Targeting based on baseline predictions

We compare targeting by causal treatment effects to a purely predictive targeting rule. A natural approach to targeting would have been to prioritize reminders for those students who, absent treatment, would have been least likely to file for FAFSA renewal. Such a policy is intuitive because students with low baseline probability of filing have the largest potential to gain from the nudge. It is also practical because it only requires information from the control group, which could be learned before the introduction of the intervention. A similar comparison to predictive approaches is performed by Ascarza (2018) in the context of churn, and by Hitsch et al. (2023) for targeting based on the predicted potential outcome under treatment in a catalog-mailing application. Fernández-Loría and Provost (2022) provides a theoretical and empirical comparison of targeting based on predicted outcomes vs treatment effects.

In order to implement a targeting policy by baseline predictions, we estimate the probability $f(x) = \mathrm{E}[Y(0)|X{=}x]$ that a student would have filed by the priority deadline absent the behaviorally-informed reminders, and give treatment first to those with the lowest predicted probability. That is, we let

$$\hat{\pi}_b^{\text{predictive}}(x) = \mathbb{1}(\hat{f}(x) \leq b)$$

where $\hat{f}(x)$ is an estimate of $f(x)$ and b is a threshold. We can implement this policy efficiently using any off-the-shelf machine-learning predictor that predicts filing by the deadline from available variables in the absence of an experiment, since $f(x) = \mathrm{E}[Y|X{=}x, T{=}0]$.

We provide an evaluation of this predictive policy using random-forest predictions $\hat{f}(x)$ in Figure 5 as the "Negative Baseline" method. This specific policy performs significantly worse than the policy based on the causal estimation of treatment effects. Indeed, we estimate that the outcome $\mathrm{E}[Y(\hat{\pi}_b^{\text{predictive}}(X))]$ of the prediction-based approach is considerably worse than assigning the same number of people randomly, across choices of the threshold b: If we chose the half of students with a below-median probability of filing at baseline, we would only increase total filing from 36.5% to around 38.5%, for a gain of less than a third of the total.

At the same time, we could have ranked students by who would have been *most* likely to file for renewal by the deadline absent of the treatment ("Positive Baseline" in Figure 5). This policy even outperforms the non-parametric CATE estimate in our example for large fractions of targeted students. Here, a likely driver of this observation is that students who are unlikely to file for FAFSA at baseline are also unlikely to be convinced to do so by the reminder. Rather, the relatively weak behavioral nudge seems to work best for those students who are already close to filing. A baseline policy that targets unlikely filers first therefore achieves exactly the opposite of the desired effect.

While both the positive and negative baseline policies may have been plausible ex-ante, we learned their properties only through the ex-post evaluation from the experiment. The causal approach has the advantage that it directly estimates a policy that we can expect to work well based solely on the empirical relationships of covariates to treatment effects, rendering guessing a policy that may work well (or testing a large number of them explicitly) unnecessary. At the same time, the example suggests gains from leveraging baseline predictions. Below, we therefore analyze how we can let the data decide how to use baseline predictions in order to get the best of both worlds.

5 Improving Targeting by Combining Predictive and Causal Modeling

Our above results demonstrate the importance of modeling the causal effect of an intervention in order to target it better, rather than relying blindly on ad-hoc predictive targeting rules. At the same time, these results show a strong relationship between baseline predictions and causal effects. In this section, we therefore ask whether we can profitably combine predictive information and causal modeling to achieve better targeting.

5.1 Model-based targeting from baseline predictions

We start with a simple model that combines non-parametric baseline predictions with a logistic regression model of treatment effects. Specifically, we model filing status as

$$\mathrm{P}(Y{=}1|X{=}x, T{=}t) = \frac{1}{1+\exp(-\alpha(x) - \beta\, t)},$$

or equivalently,

$$\operatorname{logit} \mathrm{P}(Y{=}1|X{=}x, T{=}t) = \log(\mathrm{P}(Y{=}1|X{=}x, T{=}t)/(1-\mathrm{P}(Y{=}1|X{=}x, T{=}t))) = \alpha(x) + \beta\, t.$$

Here, $\alpha(x)$ can be any non-parametric function and β represents a fixed treatment effect in log-odds. This model could be derived from a binary choice model where $\alpha(x)$ is the perceived expected net utility gain from filing, which is unrestricted and can vary arbitrarily across students. In this model, β can be interpreted as the effect of the reminder in terms of a reduction in hassle cost or an increase in the perceived value of renewing financial aid. Note that even though the mean utility is additive in the baseline and the treatment, the nonlinear form of the logit model implies that treatment effects vary with the baseline in a particular way. Specifically, treatment effects are largest at intermediate values of the mean outcome. That is, when the probability of filing is close to zero or one, the incremental impact of the treatment is low. This relationship is discussed more fully below and illustrated in Figure 8.

In order to estimate our simple model of FAFSA filing, we note that the probability of filing at baseline fulfills logit $\mathrm{P}(Y(0) = 1|X{=}x) = \alpha(x)$. Hence, a simple way we can estimate the model based on a baseline prediction $\hat{f}(x)$ of the probability $f(x) = \mathrm{P}(Y(0){=}1|X{=}x)$ of filing is to let $\tilde{f}(x) = \mathrm{logit}(\hat{f}(x))$ and to estimate the logistic regression

$$\mathrm{logit}\, \widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}t) = \hat{\alpha} + \hat{\alpha}_{\tilde{f}}\, \tilde{f}(x) + \hat{\beta}\, t. \tag{2}$$

We implement this procedure with the random-forest estimates $\hat{f}(x)$ of baseline filing from the previous section.[5] We then estimate the implied difference in predictions $\tau(x) = \mathrm{P}(Y{=}1|X{=}x, T{=}1) - \mathrm{P}(Y{=}1|X{=}x, T{=}0)$ by $\hat{\tau}^*(x) = \widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}1) - \widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}0)$, and prioritize students with high estimated treatment effects, that is,

$$\hat{\pi}_t^{\mathrm{logit}}(x) = \mathbb{1}(\hat{\tau}^*(x) \geq t).$$

The targeting policy derived from our simple logit model is presented as the “Logit from Baseline” method in Figure 6. Across both early and late covariates, this simple policy outperforms the non-parametric policy based on the causal forest. With this semiparametric policy, we can achieve around 75% of the gain from targeting only half of the students. The policy also improves substantially over targeting based on predicted or realized enrollment, showing that there is predictable heterogeneity in treatment effects beyond enrollment. It also outperforms the policy that targets those students with *highest* baseline probability of filing first (“Positive Baseline”).

5.2 Targeting based on a hybrid model that adapts to heterogeneity

While the simple logistic regression model from (2) performs well in our example, it may perform poorly when constant treatment effects in log-odds do not approximate the relationship of treatment effects to the baseline well, or when there is additional variation in treatment effects that is not captured by variation in the baseline. We therefore consider the hybrid logistic regression

$$\mathrm{logit}\, \widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}t) = \hat{\alpha} + \hat{\alpha}_{\tilde{f}}\, \tilde{f}(x) + \hat{\alpha}_{\tilde{g}}\, \tilde{g}(x) + (\hat{\beta} + \hat{\beta}_{\tilde{f}}\, \tilde{f}(x) + \hat{\beta}_{\tilde{g}}\, \tilde{g}(x))\, t \tag{3}$$

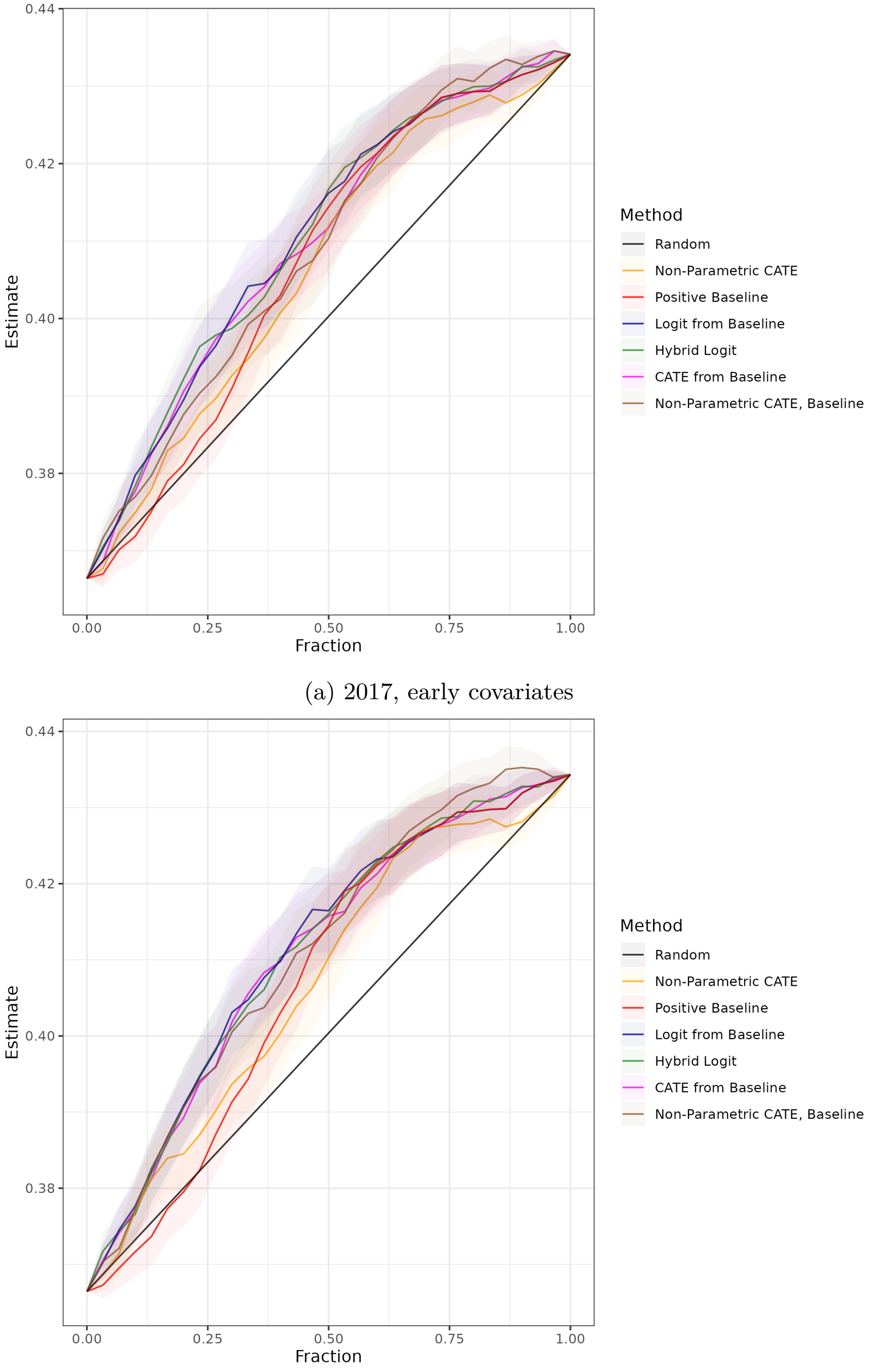
that includes the transformed baseline $\tilde{f}(x) = \mathrm{logit}(\hat{f}(x))$ as well as a logit transformation $\tilde{g}(x)$ of the non-parametric CATE estimates $\hat{\tau}(x)$ from Section 3, where we construct $\tilde{g}(x)$ by

$$\tilde{g}(x) = \mathrm{logit}(\hat{f}(x) + \hat{\tau}(x)) - \mathrm{logit}(\hat{f}(x)).$$

Here, we assume that the estimates $\hat{f}(x) + \hat{\tau}(x)$ of the probability of $Y(1){=}1$ as well as the estimates $\hat{f}(x)$ of the probability of $Y(0){=}1$ are all within the unit interval, which is the case in our data.

This specific way of parametrizing treatment effect estimates and baseline predictions has the advantage that it allows the model to capture both the simple model of constant log-odds treatment effects from

[5] We use honest out-of-bag estimates when estimating the model parameters to avoid over-fitting. When evaluating the policy by ten-fold cross-validation, we estimate all parameters on the training folds, including the logistic regression parameters, and only apply the model on the evaluation fold.



(a) 2017, early covariates

(b) 2017, late covariates

Figure 6: Total estimated FAFSA renewal rate (y-axis) by targeting a given fraction (x-axis) of students according to different cross-fitted predictions in the 2017 data as in Figure 5, with additional targeting rules based on logistic regression with a random-forest prediction of the baseline response as covariate ("Logit from Baseline"), a hybrid logit model that uses transformed baseline and treatment-effect estimates as features ("Hybrid Logit"), a non-parametric estimate of treatment effects from predicted baseline only ("CATE from Baseline"), and a non-parametric estimate of treatment effects from predicted baseline and estimated treatment effects ("Non-parametric CATE, Baseline").

(2) (by setting $\alpha_{\tilde{g}} = \beta_{\tilde{g}} = \beta_{\tilde{f}} = 0$) as well as to recover the nonparametric treatment effect estimate $\hat{\tau}(x)$ itself by setting $\hat{\alpha}_{\tilde{f}} = \hat{\beta}_{\tilde{g}} = 1$ and $\hat{\alpha} = \hat{\alpha}_{\tilde{g}} = \hat{\beta} = \hat{\beta}_{\tilde{f}} = 0$ in which case

$$\widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}1) - \widehat{\mathrm{P}}(Y{=}1|X{=}x, T{=}0) = \frac{1}{1+\exp(-\tilde{f}(x) - \tilde{g}(x))} - \frac{1}{1+\exp(-\tilde{f}(x))} = \hat{\tau}(x).$$

By fitting the model, we can let the data decide which of these models works best.

The performance of this hybrid logit model is documented in Figure 6, where it performs on par with or slightly worse than the simpler logit model that uses baseline information only, while outperforming targeting based on the fully non-parametric estimate throughout.

To further document the ability of this hybrid model to adapt to the level of heterogeneity, we present the results of a simulation study in Figure 7. Here, we take the non-parametric baseline and treatment-effect estimates $\hat{f}(x)$ and $\hat{\tau}(x)$,[6] re-randomize treatment assignment, and re-draw outcomes by the model

$$\operatorname{logit} \mathrm{P}(Y{=}1|X{=}x, T{=}t) = \tilde{f}(x) + \lambda\,(\tilde{g}(x) - \mathrm{E}[\tilde{g}(X)])\,t + \mathrm{E}[\tilde{g}(X)]\,t.$$

λ expresses the degree of treatment effect heterogeneity. For $\lambda = 1$, this model simply simulates based on the non-parametric estimates $\hat{f}(x)$ and $\hat{\tau}(x)$. For $\lambda = 0$ this model would assume that treatment effects are constant in log-odds, while $\lambda > 1$ corresponds to additional variation.

The simulation results highlight the benefit of the hybrid model: for low heterogeneity ($\lambda = 1/2$ and $\lambda = 1$, top row in Figure 7), the simple logit regression from baseline from (2) performs well, while the hybrid model comes close to its performance. For high variation ($\lambda = 3$, bottom right panel), the fully non-parametric estimate of heterogeneous treatment effects performs best, with the hybrid model achieving the same performance. In the intermediate regime ($\lambda = 2$, bottom left panel), the non-parametric model performs best for small treatment fractions, while the simple logit performs well for high fractions. Here, the hybrid model represents a compromise between both, coming close to the non-parametric model, while outperforming the simple logistic model. Overall, the hybrid targeting model thus adapts to the level of signal and provides comparatively good performance throughout.

5.3 Non-parametric targeting with baseline predictions as a feature

As an alternative to the parametric model in Equation 2, we also model treatment effects as a non-parametric function of the estimated baseline probability of filing by the priority deadline. The performance of a policy that estimates treatment effects using a strongly regularized causal forest from predicted baseline probabilities is plotted as the "CATE from Baseline" method in Figure 6, where it is shown to obtain performance slightly worse than and on par with the simple logit, respectively, and substantially outperforms the direct non-parametric estimate of treatment effects. In general, we would expect this method to do well whenever treatment effects mainly vary with the baseline, and a simple relationship of treatment effects to baseline may not be sufficient to express them.

[6] We re-calculate non-parametric honest estimates $\hat{f}(x)$ and $\hat{\tau}(x)$ using the random forest and causal forest, respectively, for which we use more complex trees than for the results reported in Section 3 to obtain enough heterogeneity.

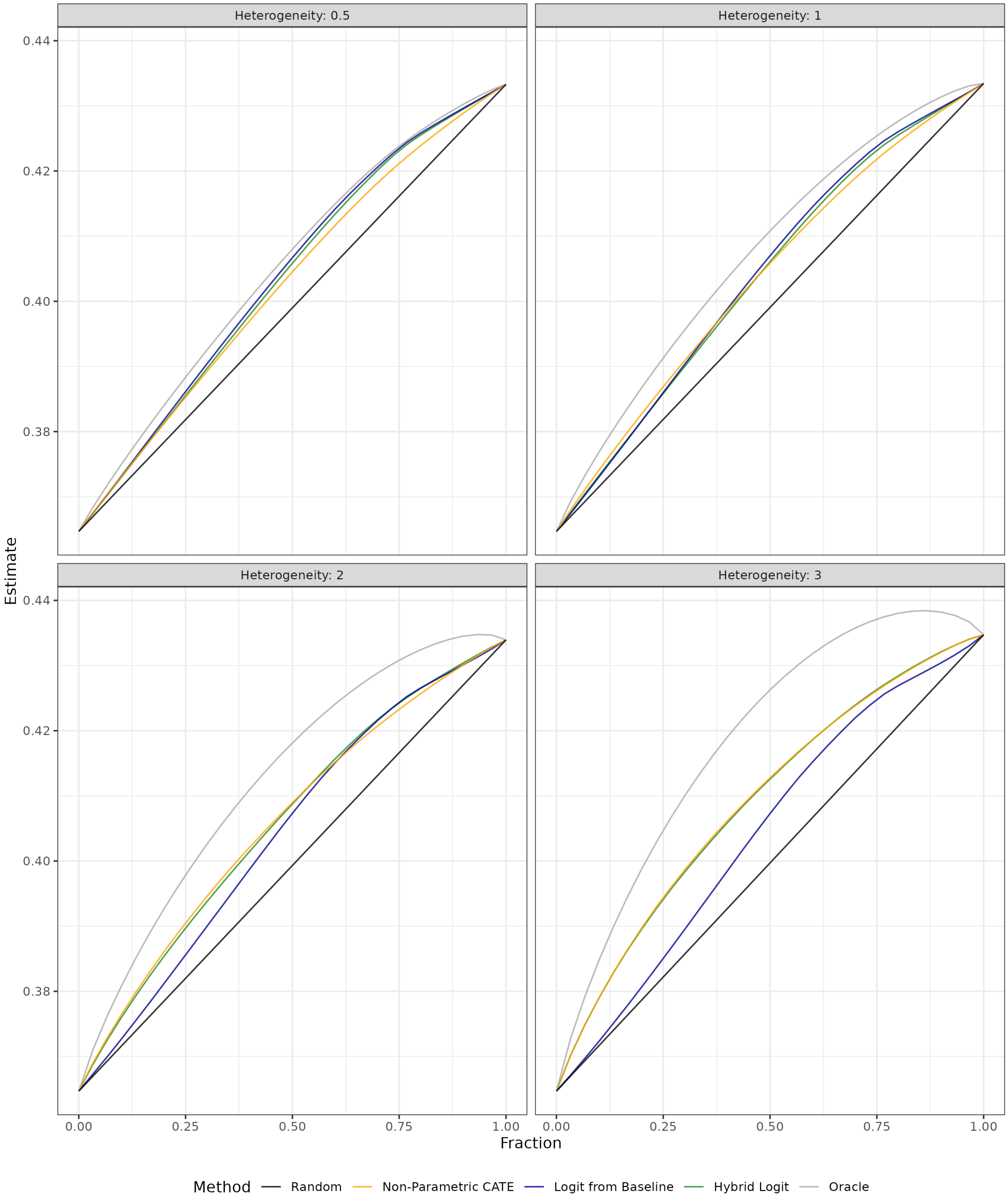


Figure 7: Performance of selected models from Figure 6 in a simulation based on the 2017 data with early covariates and varying strengths of heterogeneity. Shown are averages over 20 random draws of treatment assignment, using the same sample and ten-fold cross-fitting scheme as in Figure 6 and comparing based on known simulated treatment effects. The "Oracle" method ranks students based on these known effects.

In order to visualize this strategy, Figure 8 plots estimated treatment effects against baseline predictions of filing for the 2017 data with early covariates, presenting data from a single fold. The non-parametric treatment-effect estimates themselves only vary moderately, while the models that explicitly leverage baseline predictions are able to capture additional variation. Among them, the hybrid logit model has additional spread in treatment effect estimates for the same baseline level, but also shows a clear overall curved relationship similar to that of the simple logit.

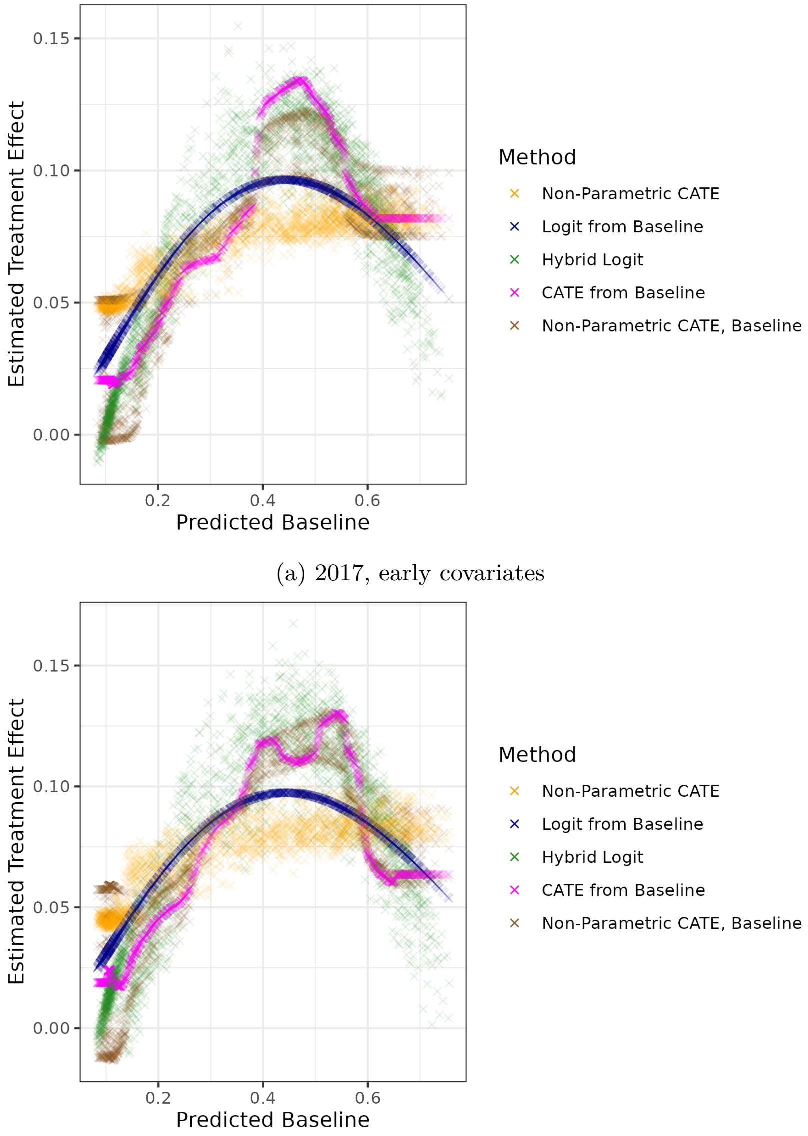
Modeling treatment effects directly as a function of the predicted baseline is only one way in which we can use baseline predictions. In principle, we could obtain additional targeting models by combining baseline predictions with additional features, such as predicted enrollment, quartiles of estimated treatment effects, or some of the original covariates. Figure 10 in the appendix presents such additional specifications. However, the simple logit model introduced above generally performs at least comparably to those more complicated alternatives. We also note that it improves over a simple parametric logit model based on a few selected covariates.

A specific model that is able to adapt to treatment effect heterogeneity could be obtained by estimating treatment effects non-parametrically from baseline predictions and non-parametric CATE estimates. Specifically, we could estimate treatment effects from a heavily causal forest that takes as input out-of-bag baseline estimates $\hat{f}(x)$ as well as non-parametric CATE estimates $\hat{\tau}(x)$ obtained as in Section 3. Like the semi-parametric logistic regression model from Section 5.2, this approach is able to recover the non-parametric CATE estimates when they capture heterogeneity well, while also allowing for modeling heterogeneous treatment effects as a function of the baseline. This policy is labeled "Non-Parametric CATE, Baseline" in Figure 6 and Figure 8. It performs better than targeting by the non-parametric CATE alone, but not quite as well as the logistic-regression model.

We interpret our results in this section in terms of the importance of regularization, choice of features, and functional form. When treatment effect variation is small and their estimation is noisy, then leveraging first-order variation in the baseline and modeling its relationship to treatment effects in simpler ways provides a form of regularization that preserves variation in treatment effect estimates and can improve performance. This finding extends the semiparametric approach of Athey et al. (2021) from the estimation of average to the estimation of heterogeneous effects. The performance of simple semi-parametric models depends crucially on their functional form. In our example, a standard logit model with constant treatment effect coefficient and non-parametric baseline predictions appears to capture the structure of treatment effects remarkably well. At the same time, the baseline itself is straightforward to estimate in our data, and staying fully non-parametric for the baseline appears to provide better performance than imposing a model on it as well.

6 Conclusion

In this article, we compared and combined predictive and causal approaches to targeting interventions. Our analysis provides an example of the value of integrating careful experimentation, causal inference, and predictive modeling. By itself, predictive machine learning without experimental evaluation could have led to a bad policy, while combining predictive modeling and causal targeting based on a coherent



(a) 2017, early covariates

(b) 2017, late covariates

Figure 8: Heterogeneous treatment effect estimates by estimated baseline (“Non-Parametric CATE”), along with treatment effects re-estimated by a simple logistic regression with a constant coefficient on treatment (“Logit from Baseline”), by the hybrid logit model from Section 5.2, from baseline using the causal forest (“CATE from Baseline”), and from baseline and non-parametric CATE using the causal forest (“Non-parametric CATE, Baseline”), shown here for a single fold.

analysis of heterogeneous treatment effects can ultimately lead to sizeable gains.

Our analysis also points to the challenges of evaluating existing experiments with machine learning. While sample and effect sizes seem large for estimating average treatment effects, an intervention designed to work well on average in an experiment powered for estimating averages makes the precise estimation of heterogeneous treatment effects statistically and technically challenging. Future experiments could also rely on an integration of targeting into their design.

Our results come with important caveats that limit statistical power, generalizability, and policy applicability. Since this experiment was not designed for heterogeneous-treatment effect analysis, treatment arms were chosen to work well on average, rather than for specific subgroups. Overall treatment effects are moderate since they come from relatively modest nudges, and the experiment was powered to detect average effects rather than effects on many subgroups. Finally, sending behaviorally informed reminders is cheap and does not appear to have any negative treatment effects in the experiment, so while these results can help target reminders to those for whom they will work best, the main effect remains limited to avoiding inundating students with reminder texts and emails for who the effect would be small.

We believe that overcoming these shortcomings in future studies requires designing experiments *ex ante* to estimate heterogeneous treatment effects in the first place. This includes designing individual treatment arms that are likely to affect different people differently so that differentiated treatments can be matched to appropriate individuals and situations. It also involves updating power analyses to the higher sample size demands for estimating heterogeneous treatment effects, rather than average effects alone. Finally, policies based on heterogeneous treatment effects will be particularly important when treatment delivery is costly, or when we need to make choices between treatments for which none dominates others across individuals.

We close this article by discussing questions that our results raise about fairness and equity in targeting. Our analysis suggests that nudge-type interventions may be most effective for those who are already more likely to engage in the desired behavior at baseline, and that not targeting those with low expected outcomes may therefore be efficient. However, there may be reasons why a planner may attach higher importance to improving outcomes of individuals with a low baseline. While we leave a more careful treatment to future research, we note that our approach can help capture such welfare considerations, in two ways: First, we could directly optimize for an assignment that puts higher weight on the outcomes of individuals with, say, a low expected baseline. Second, we could use our estimates of causal effects to understand better for which group the current intervention is *not* effective and thereby inform the design of better interventions.

References

Ascarza, Eva (2018). Retention futility: Targeting High-Risk customers might be ineffective. *J. Mark. Res.*, 55(1):80–98. (Cited on pages 4 and 16.)

Athey, Susan, Peter J Bickel, Aiyou Chen, Guido Imbens, and Michael Pollmann (2021). Semiparametric estimation of treatment effects in randomized experiments. Technical report, National Bureau of Economic Research. (Cited on pages 2, 3, and 22.)

Athey, Susan and Guido Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.*, 113(27):7353–7360. (Cited on page 3.)

Athey, Susan and Guido W Imbens (2019). Machine learning methods that economists should know about. *Annu. Rev. Econom.*, 11(1):685–725. (Cited on page 3.)

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). Generalized random forests. *Ann. Stat.* (Cited on pages 3 and 7.)

Athey, Susan and Stefan Wager (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161. (Cited on page 4.)

Castleman, Benjamin L and Lindsay C Page (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *J. Hum. Resour.*, 51(2):389–415. (Cited on page 4.)

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research. (Cited on page 9.)

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val (2019). Generic machine learning inference on heterogenous treatment effects in randomized experiments. (Cited on pages 3, 9, and 10.)

Devriendt, Floris, Jeroen Berrevoets, and Wouter Verbeke (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, 548:497–515. (Cited on page 4.)

Fernández-Loría, Carlos and Foster Provost (2022). Causal classification: Treatment effect estimation vs. outcome prediction. *The Journal of Machine Learning Research*, 23(1):2573–2607. (Cited on pages 4 and 16.)

Haushofer, Johannes, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael W Walker (2022). Targeting impact versus deprivation. Technical report, National Bureau of Economic Research. (Cited on page 4.)

Hitsch, Günter, Sanjog Misra, and Walter Zhang (2023). Heterogeneous treatment effects and optimal targeting policy evaluation. *SSRN Electronic Journal*. (Cited on pages 3, 4, 12, 13, 14, 16, and 35.)

ideas42 (2016). Meeting the FAFSA priority deadline. Technical report. (Cited on page 4.)

Imai, Kosuke and Marc Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. (Cited on page 3.)

Inoue, Kosuke, Susan Athey, and Yusuke Tsugawa (2023). Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *International Journal of Epidemiology*, page dyad037. (Cited on page 4.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). Prediction policy problems. *Am. Econ. Rev.*, 105(5):491–495. (Cited on page 3.)

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter (2022). Heterogeneous employment effects of job search programs. *Journal of Human Resources*, 57(2):597–636. (Cited on page 4.)

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165. (Cited on page 3.)

Mullainathan, Sendhil and Jann Spiess (2017). Machine learning: An applied econometric approach. *J. Econ. Perspect.*, 31(2):87–106. (Cited on page 3.)

Nie, Xinkun, Emma Brunskill, and Stefan Wager (2021). Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409. (Cited on page 3.)

Nissan, R., S. Kenney, S. Lensing, J. Anderson, T.-A. Richards, A. Barrows, and D. Palmer (2020). Student success toolkit. (Cited on page 1.)

Rzepakowski, Piotr and Szymon Jaroszewicz (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and information systems*, 32(2):303–327. (Cited on pages 3 and 14.)

Sun, Hao, Shuyang Du, and Stefan Wager (2021). Treatment Allocation under Uncertain Costs. (Cited on page 14.)

Sunstein, Cass R and Richard Thaler (2008). *Nudge.* Yale University Press. (Cited on page 4.)

Wager, Stefan and Susan Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.*, 113(523):1228–1242. (Cited on pages 3 and 7.)

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966.* (Cited on pages 3 and 14.)

Yang, Jeremy, Dean Eckles, Paramveer Dhillon, and Sinan Aral (2020). Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835.* (Cited on page 4.)

Zhang, Walter W and Sanjog Misra (2022). Coarse personalization. *arXiv preprint arXiv:2204.05793.* (Cited on page 4.)

Zhao, Lihui, Lu Tian, Tianxi Cai, Brian Claggett, and L J Wei (2013). Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association*, 108(502):527–539. (Cited on pages 3 and 14.)

Zhou, Zhengyuan, Susan Athey, and Stefan Wager (2023). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183. (Cited on page 4.)

A Additional Tables and Figures

	Control $N = 12,658$	Treatment $N = 12,480$	p-value	N
COLLEGE_IN_INTERVENTION_SPR:			0.69	25138
0	3953 (31.229%)	3928 (31.474%)		
1	8705 (68.771%)	8552 (68.526%)		
AGE	23.680 (6.635)	23.556 (6.518)	0.19	19153
GENDER:			0.51	25138
Men	5396 (42.629%)	5372 (43.045%)		
Women	7262 (57.371%)	7108 (56.955%)		
ETHNICITY:			0.89	25138
American Indian or Native Alaskan	39 (0.308%)	33 (0.264%)		
Asian or Pacific Islander	1072 (8.469%)	1091 (8.742%)		
Black, Non-Hispanic	4107 (32.446%)	4067 (32.588%)		
Hispanic, Other	6558 (51.809%)	6422 (51.458%)		
White, Non-Hispanic	882 (6.968%)	867 (6.947%)		
TRANSFER:			0.32	25138
0	9104 (71.923%)	8874 (71.106%)		
1	591 (4.669%)	584 (4.679%)		
'Missing'	2963 (23.408%)	3022 (24.215%)		
FT_PT_STATUS:			0.39	25138
FULL-TIME	6610 (52.220%)	6415 (51.402%)		
PART-TIME	2441 (19.284%)	2473 (19.816%)		
'Missing'	3607 (28.496%)	3592 (28.782%)		
GPA_CUMU_BF	2.475 (0.952)	2.472 (0.947)	0.84	14633
CRD_CUMU_ATMPT_BF	18.706 (16.565)	18.785 (16.524)	0.75	17939
CRD_CUMU_EARN_BF	17.475 (15.661)	17.387 (15.505)	0.70	17939

Table 2: Balance table for the 2017 FAFSA experiment.

	Control $N = 3226$	Treatment 1 $N = 13698$	Treatment 2 $N = 13610$	p-value	N
AGE	23.412 (6.566)	23.625 (6.764)	23.425 (6.601)	0.07	23164
GENDER:				0.40	30534
Men	1464 (45.381%)	6062 (44.255%)	5998 (44.071%)		
Women	1762 (54.619%)	7636 (55.745%)	7612 (55.929%)		
ETHNICITY:				0.83	30534
American Indian or Native Alaskan	13 (0.403%)	69 (0.504%)	68 (0.500%)		
Asian or Pacific Islander	503 (15.592%)	2065 (15.075%)	2000 (14.695%)		
Black, Non-Hispanic	971 (30.099%)	4112 (30.019%)	4019 (29.530%)		
Hispanic, Other	1419 (43.986%)	6092 (44.474%)	6147 (45.165%)		
White, Non-Hispanic	320 (9.919%)	1360 (9.928%)	1376 (10.110%)		
TRANSFER:				0.59	30534
0	2319 (71.885%)	9809 (71.609%)	9776 (71.830%)		
1	132 (4.092%)	593 (4.329%)	535 (3.931%)		
'Missing'	775 (24.024%)	3296 (24.062%)	3299 (24.240%)		
FT_PT_STATUS:				0.67	30534
FULL-TIME	1706 (52.883%)	7350 (53.657%)	7342 (53.946%)		
PART-TIME	643 (19.932%)	2692 (19.653%)	2601 (19.111%)		
'Missing'	877 (27.185%)	3656 (26.690%)	3667 (26.943%)		
GPA_CUMU_BF	2.538 (0.934)	2.508 (0.942)	2.500 (0.948)	0.28	19020
CRD_CUMU_ATMPT_BF	21.635 (17.409)	21.285 (16.829)	21.275 (16.995)	0.63	22334
CRD_CUMU_EARN_BF	19.330 (15.676)	19.001 (15.169)	18.968 (15.317)	0.58	22334

(i) Early schools

	Control $N = 2497$	Treatment 1 $N = 3802$	Treatment 2 $N = 3699$	p-value	N
AGE	23.742 (6.448)	23.906 (6.724)	24.076 (6.803)	0.23	7866
GENDER:				0.50	9998
Men	1200 (48.058%)	1770 (46.554%)	1745 (47.175%)		
Women	1297 (51.942%)	2032 (53.446%)	1954 (52.825%)		
ETHNICITY:				0.54	9998
American Indian or Native Alaskan	9 (0.360%)	10 (0.263%)	6 (0.162%)		
Asian or Pacific Islander	195 (7.809%)	313 (8.233%)	278 (7.516%)		
Black, Non-Hispanic	814 (32.599%)	1241 (32.641%)	1219 (32.955%)		
Hispanic, Other	1100 (44.053%)	1717 (45.160%)	1646 (44.499%)		
White, Non-Hispanic	379 (15.178%)	521 (13.703%)	550 (14.869%)		
TRANSFER:				0.36	9998
0	1799 (72.046%)	2753 (72.409%)	2692 (72.776%)		
1	151 (6.047%)	258 (6.786%)	213 (5.758%)		
'Missing'	547 (21.906%)	791 (20.805%)	794 (21.465%)		
FT_PT_STATUS:				0.18	9998
FULL-TIME	1309 (52.423%)	2109 (55.471%)	2019 (54.582%)		
PART-TIME	497 (19.904%)	689 (18.122%)	682 (18.437%)		
'Missing'	691 (27.673%)	1004 (26.407%)	998 (26.980%)		
GPA_CUMU_BF	2.408 (0.935)	2.459 (0.935)	2.494 (0.936)	0.02	6223
CRD_CUMU_ATMPT_BF	22.097 (17.048)	22.089 (17.149)	22.415 (17.162)	0.74	7305
CRD_CUMU_EARN_BF	19.888 (15.367)	19.900 (15.465)	20.184 (15.339)	0.74	7305

(ii) Late schools

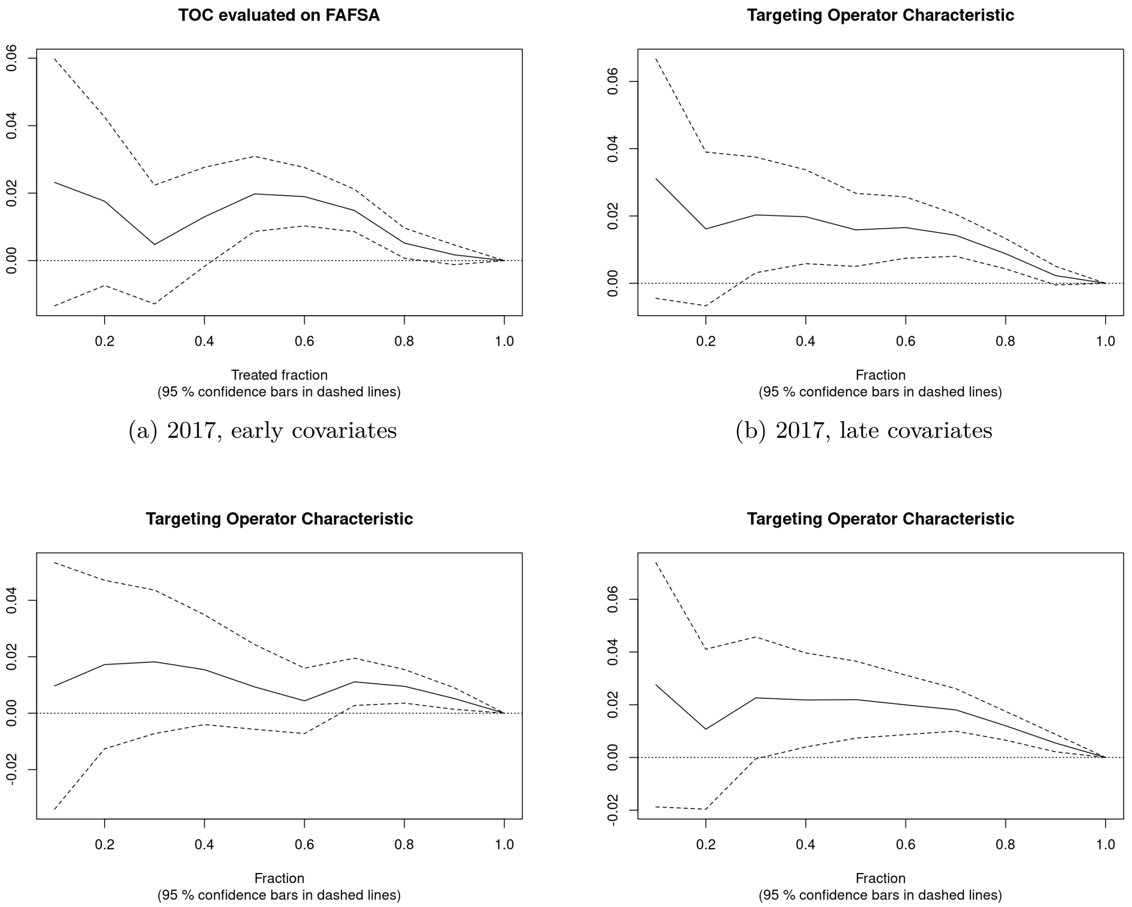
Table 3: Balance tables for the 2018 FAFSA experiment.

Year	School timeline	Method	ATE	SE
2017		Mean difference	0.0641	0.0061
		Constant propensity	0.0687	0.0053
		AIPW	0.0686	0.0053
2018	all	Mean difference	0.1209	0.0074
		Constant propensity	0.1182	0.0065
		AIPW	0.1182	0.0065
	early	Mean difference	0.1213	0.0091
		Constant propensity	0.1198	0.0079
		AIPW	0.1200	0.0080
	late	Mean difference	0.1201	0.0109
		Constant propensity	0.1134	0.0099
		AIPW	0.1132	0.0100

Table 4: Overall average treatment effects, estimated by simple (propensity-adjusted) differences in averages ("Mean difference") as well as by an augmented inverse propensity score estimator based on random forests with constant ("Constant propensity") and flexible propensity score ("AIPW"), respectively.

Year	Covariates		Q1	Q2	Q3
2017	early	Q2	0.018 (0.015)		
		Q3	**0.063 (0.015)**	**0.045 (0.017)**	
		Q4	**0.049 (0.016)**	**0.031 (0.017)**	−0.014 (0.017)
	late	Q2	**0.040 (0.014)**		
		Q3	**0.045 (0.015)**	0.005 (0.017)	
		Q4	**0.060 (0.015)**	0.019 (0.017)	0.015 (0.018)
2018	early	Q2	0.028 (0.018)		
		Q3	**0.046 (0.019)**	0.018 (0.020)	
		Q4	**0.051 (0.019)**	0.023 (0.020)	0.005 (0.020)
	late	Q2	**0.036 (0.018)**		
		Q3	**0.042 (0.018)**	0.005 (0.020)	
		Q4	**0.071 (0.018)**	**0.034 (0.020)**	0.029 (0.020)

Table 5: Pairwise difference of quartile treatment effects from Figure 3 based on AIPW estimates, with standard error estimate in parentheses. **Bold estimates** denote statistically significant increases at the 5% level based on a one-sided test.



(a) 2017, early covariates

(b) 2017, late covariates

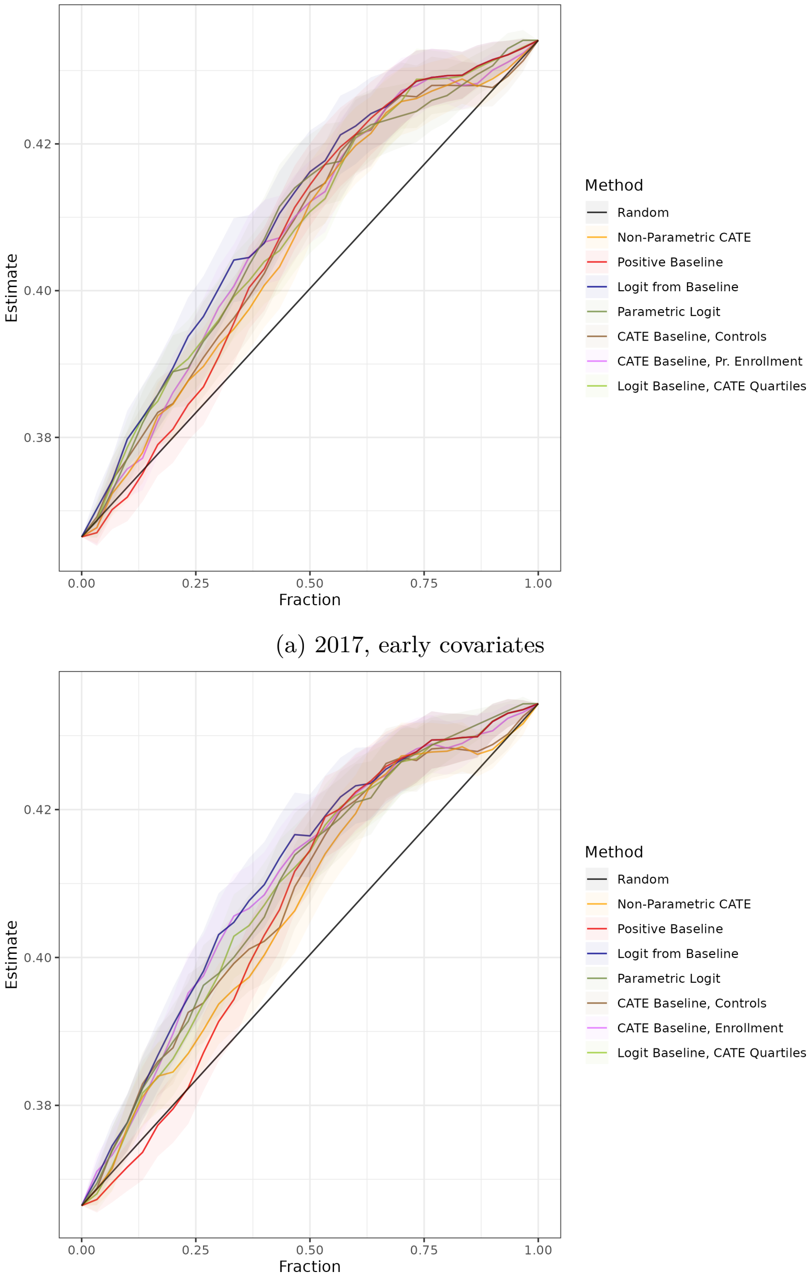
(c) 2018, early covariates

(d) 2018, late covariates

Figure 9: RATE estimates.

Year	Covariates	AUTOC	SE	p-value
2017	early	0.01020	0.00588	0.04143
	late	0.01351	0.00581	0.01001
2018	early	0.01107	0.00792	0.08103
	late	0.01628	0.00768	0.01706

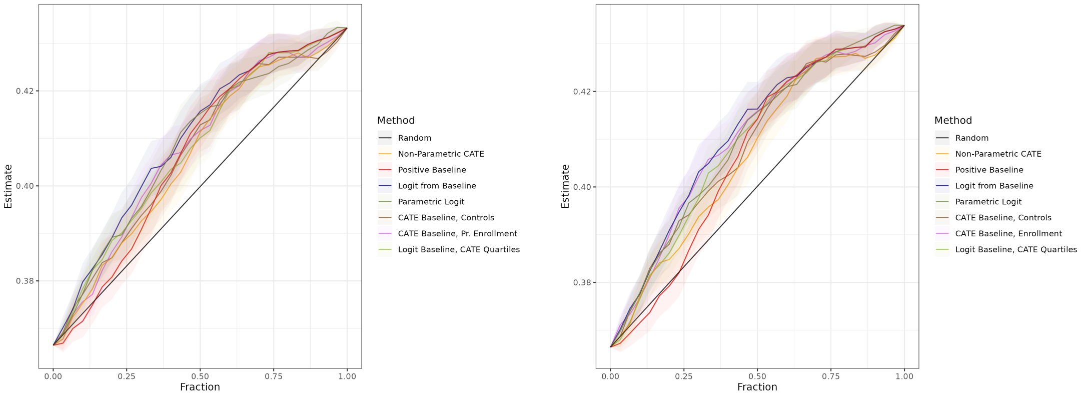
Table 6: RATE-based omnibus test with one-sides p-value against the alternative of no benefit from targeting over random assignment.



(a) 2017, early covariates

(b) 2017, late covariates

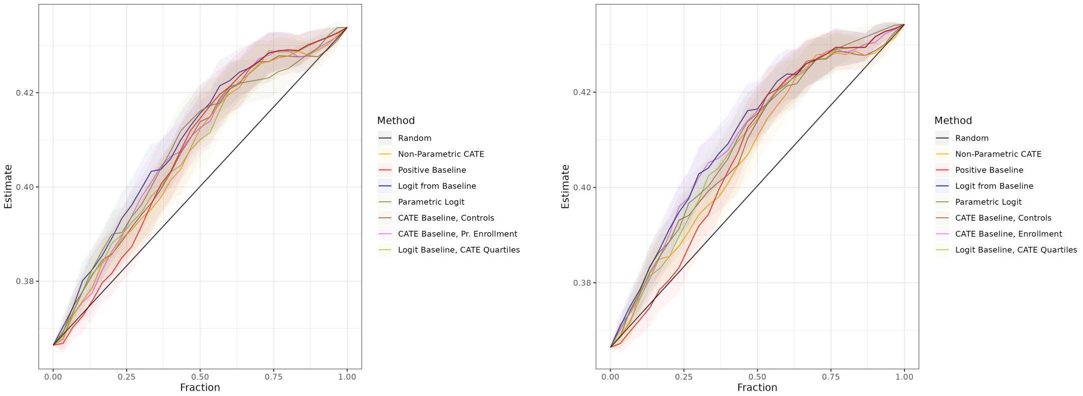
Figure 10: Total estimated FAFSA renewal rate (y-axis) by targeting a given fraction (x-axis) of students according to different cross-fitted predictions in the 2017 data as in Figure 5, with additional targeting rules based on a causal-forest estimate from a prediction of the baseline and covariates ("CATE from Baseline, Controls") from a prediction of the baseline and predicted or actual enrollment ("CATE Baseline, Enrollment"), as well as a logistic regression on baseline and treatment interacted with quartiles of estimated treatment effects ("Logit Baseline, CATE Quartiles") and a treatment-interacted logistic regression on five variables chosen by the logit-LASSO ("Parametric Logit").



(a) 2017, early covariates

(b) 2017, late covariates

(i) Same as Figure 10, but cross-fold estimation of renewal rates are now performed using nuisance parameter estimates (of the CATE and the conditional outcome) that are re-estimated on the left-out fold using honest out-of-bag estimates in order to separate the evaluation from the estimation of policies on the training folds.

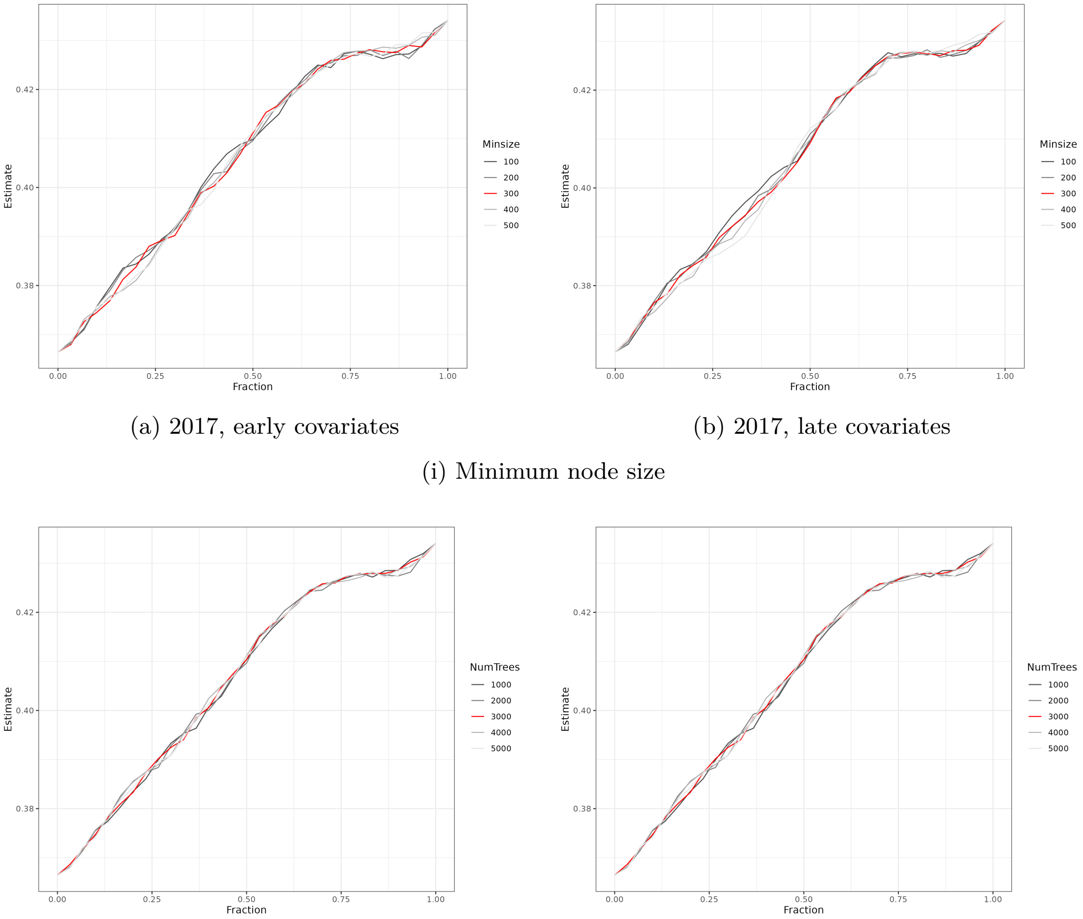


(c) 2017, early covariates

(d) 2017, late covariates

(ii) Same as Panel (i), but the propensity score is now also re-estimated on the left-out fold (rather than assumed constant) using a logistic regression.

Figure 11: The figure evaluates the same policies as Figure 10, but makes changes to the estimation of renewal rates in the held-out folds that serve as additional robustness checks. Panel (i) re-estimates nuisance parameters within fold and uses a fixed propensity score, and Panel (ii) also re-estimates the propensity score.



(a) 2017, early covariates

(b) 2017, late covariates

(i) Minimum node size

(c) 2017, early covariates

(d) 2017, late covariates

(ii) Number of trees

Figure 12: The figures analyze the sensitivity of policies based on the non-parametric CATE estimate from Figure 5 on key tuning parameters of the causal forest, namely the (i) minimum node size and the (ii) number of trees. The red curve on each plot is the choice we use for targeting in Section 4 and Section 5.

B Evaluation of Assignment Policies

In the main article, we compare the precision of different treatment-effect estimates in terms of the average outcomes that can be achieved when we use them for targeting. In this section, we discuss estimation and inference on these average outcomes, which we use to obtain Figure 5, as well as Figure 6 and Figure 10.

Consider treatment assignment policies π that map characteristics $X{=}x$ to probabilities $\pi(x) \in [0,1]$ of being treated. (This may include policies derived from treatment-effect estimates and random assignment, in which case $\pi(x) \equiv q$ with q the probability of assignment.) When treatment is assigned completely randomly (or randomly with known propensity score that only depends on X) in the existing data and X is observed, then the average outcome $\mathrm{E}[\pi(X)Y(1) + (1-\pi(X))Y(0)]$ under this policy, the total lift $\mathrm{E}[\pi(X)(Y(1)-Y(0))]$ relative to baseline, and the average treatment effect $\frac{\mathrm{E}[\pi(X)(Y(1)-Y(0))]}{\mathrm{E}[\pi(X)]}$ of those assigned to treatment are all identified, since $\mathrm{E}[Y(1)|X] = \mathrm{E}[Y|X,T{=}1], \mathrm{E}[Y(0)|X] = \mathrm{E}[Y|X,T{=}0]$ are.

Focusing on the case of average outcomes as in Figure 5, we write $U(\pi) = \mathrm{E}[\pi(X)Y(1){+}(1{-}\pi(X))Y(0)$ for the expected outcome under this policy, which is identified by $U(\pi) = \mathrm{E}[\pi(X)Y|T{=}1] + \mathrm{E}[(1-\pi(X))Y|T{=}0]$ and could be estimated within fold k by its sample analogue

$$\hat{U}_k(\pi) = \frac{\sum_{i;k(i)=k} T_i \pi(X_i) Y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1-T_i)(1-\pi(X_i))Y_i}{\sum_{i=1}^n (1-T_i)} \tag{4}$$

$$= \frac{1}{n} \sum_{i;k(i)=k} \frac{T_i}{\hat{p}} \pi(X_i) Y_i - \frac{1-T_i}{1-\hat{p}}(1-\pi(X_i))Y_i \quad \text{with} \quad \hat{p} = \frac{\sum_{i;k(i)=k} T_i}{\sum_{i;k(i)=k} 1} \tag{5}$$

as in Hitsch et al. (2023), who consider the case of a known propensity score and non-stochastic assignment.

In our implementation for Figure 5, we are specifically interested in making inference on differences $U(\pi) - U(\bar{\pi}_q)$ in outcomes of a policy π that assigns students to treatment based on some rule (such as by ranking by estimated treatment effects) relative to the baseline policy $\bar{\pi}_q(x) \equiv q$ that assigns a *random* fraction q to treatment. We note that for any two policies π and $\bar{\pi}$, we have $U(\pi) - U(\bar{\pi}) = \mathrm{E}[(\pi(X) - \bar{\pi}(X))Y(1) - Y(0)] = \mathrm{E}[(\pi(X) - \bar{\pi}(X))\tau(X)]$, and for this specific baseline policy $\bar{\pi}_q$, we find $U(\bar{\pi}_q) = \mathrm{E}[Y(0)] + q\,\mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[Y|T{=}0] + q\,\mathrm{E}[\tau(X)]$.[7] Since $\mathrm{E}[Y|T{=}0]$ is readily estimated, we therefore now focus on efficient estimation and valid inference on weighted average treatment effects $\tau_w = \mathrm{E}[w(X)\tau(X)]$, where weights can be negative. Once we have established estimation and inference for τ_w, we can estimate all quantities of interest via

$$U(\pi) - U(\bar{\pi}_q) = \tau_{\pi - \bar{\pi}_q}, \qquad U(\bar{\pi}_q) = \mathrm{E}[Y|T{=}0] + \tau_1, \qquad U(\pi) = U(\pi) + (U(\pi) - U(\bar{\pi}_q)).$$

To improve efficiency and robustness of our estimate, as well as to ensure valid inference later on, we consider an augmented inverse propensity weighted (AIPW) estimator of $\tau_w = \mathrm{E}[w(X)\tau(X)]$.

[7] When propensity scores are non-constant, estimating $\mathrm{E}[Y(0)]$ will require additional care, and can be achieved by propensity-score weighting.

Specifically, we assume that we have a consistent estimate $\hat{f}(x)$ of $\mathrm{E}[Y|X{=}x]$, a consistent estimate $\hat{\tau}(x)$ of $\tau(x)$, and a consistent propensity score estimate $\hat{p}(x)$ of $\mathrm{E}[T|X{=}x]$ available. The propensity score may be assumed to be constant when units are randomized unconditionally, in which case we may want to set $\hat{p}(x) \equiv \frac{\sum_{i=1}^n T_i}{n}$ as above. We assume that $\hat{f}, \hat{\tau}, \hat{p}$ are all fitted on separate data or using k-fold cross-fitting. Writing $\hat{f}_1(x) = \hat{f}(x) + (1 - \hat{p}(x))\hat{\tau}(x)$, $\hat{f}_0(x) = \hat{f}(x) - \hat{p}(x)\hat{\tau}(x)$, the AIPW estimator

$$\begin{aligned}
\hat{\tau}_w^{\mathrm{AIPW}} &= \frac{1}{n}\sum_{i=1}^n w(X_i) \overbrace{\left(\hat{\tau}(X_i) + \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}(Y_i - \hat{f}_{T_i}(X_i))\right)}^{=\hat{\tau}^{\mathrm{AIPW}}(Y_i,T_i,X_i)} \\
&= \frac{1}{n}\sum_{i=1}^n w(X_i) \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}(Y_i - \bar{f}(X_i))
\end{aligned}$$

where $\bar{f}(x) = (1-\hat{p}(x))\hat{f}_1(x) + \hat{p}(x)\hat{f}_0(x) = \hat{f}(x) + (1-2\hat{p}(x))\hat{\tau}(x) = \hat{f}_0(x) + (1-\hat{p}(x))\hat{\tau}(x)$. This estimator is $\sqrt{n}$ consistent and asymptotically Normal under standard regularity conditions, and it is exactly unbiased when the propensity score is known and remains consistent even when treatment-effect and outcome estimates are not. We can consistently estimate its asymptotic variance by

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum_{i=1}^m \left(w(X_i)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i,T_i,X_i) - \hat{\tau}_w^{\mathrm{AIPW}}\right)^2$$

to obtain standard error estimates $\frac{\hat{\sigma}_w}{\sqrt{n}}$ and a corresponding 95 % confidence interval $\hat{\tau}_w^{\mathrm{AIPW}} \pm 1.96 \cdot \frac{\hat{\sigma}_w}{\sqrt{n}}$. Applying this estimator to the estimation for Figure 5, we can estimate

$$\begin{aligned}
\hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q) &= \frac{\sum_{i=1}^n (1-T_i)Y_i}{\sum_{i=1}^n (1-T_i)} + q\ \frac{1}{n}\sum_{i=1}^n \hat{\tau}^{\mathrm{AIPW}}(Y_i,T_i,X_i) \\
\hat{U}^{\mathrm{AIPW}}(\pi) &= \hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q) + \underbrace{\frac{1}{n}\sum_{i=1}^n (\pi(X_i) - q)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i,T_i,X_i)}_{=\hat{\Delta}(\pi)}, \\
\widehat{\mathrm{SE}}\left(\hat{U}^{\mathrm{AIPW}}(\pi) - \hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q)\right) &= \sqrt{\frac{1}{n^2}\sum_{i=1}^n \left((\pi(X_i) - q)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i,T_i,X_i) - \hat{\Delta}(\pi)\right)^2}.
\end{aligned}$$

We note that, in particular, the estimation outcome of the random policy does not include any additional randomization, which would only add noise to the estimation. Further, we obtain the estimator in Equation 4 of $U(\pi)$ when we set $\hat{p}(x) \equiv \frac{\sum_{i=1}^n T_i}{n}, \hat{\tau}(x) \equiv 0, \hat{f}(x) \equiv 0$, which is generally inefficient.

So far, we have considered fixed policies π. However, in our application, policies are themselves estimated, such as the policy

$$\hat{\pi}_{\hat{t}}(x) = \mathbb{1}(\hat{\tau}(x) \geq \hat{t})$$

where treatment effects $\hat{\tau}(x)$ and the cutoff $\hat{t}$ chosen to achieve a given proportion q in treatment are all noisy. We use cross-fitting to avoid biases from estimation in this case. Specifically, we estimate the quantities of interest separately on each fold, using rankings estimated based on the other folds only, and then aggregate across all folds. Under regularity conditions is the covariance between estimates from different folds of a lower order than the variance we estimate, allowing us to combine estimates and variance estimates across folds to obtain valid inference.