Reevaluating Causal Estimation Methods with Data from a Product Release*

Justin Young† Muthoni Ngatia‡ Eleanor W. Dillon§

January 9, 2026

Abstract

Recent developments in causal machine learning methods have made it easier to estimate flexible relationships between confounders, treatments and outcomes, making unconfoundedness assumptions in causal analysis more palatable. How successful are these approaches in recovering ground truth baselines? In this paper we analyze a new data sample including an experimental rollout of a new feature at a large technology company and a simultaneous sample of users who endogenously opted into the feature. We find that recovering ground truth causal effects is feasible—but only with careful modeling choices. Our results build on the observational causal literature beginning with LaLonde (1986), offering best practices for more credible treatment effect estimation in modern, high-dimensional datasets.

*For helpful comments, we are grateful to Guido Imbens, Jann Spiess, Vasilis Syrgkanis, Lihua Lei, Julian Nyarko, Stefan Wager, Ramesh Johari, Jason Weitze, Lea Bottmer, David Ritzwoller, Fabio Vera, and conference and seminar participants at NABE TEC 2024, Microsoft Research, and Stanford.

†Microsoft, email: youngjustin@microsoft.com

‡Microsoft, email: mngatia@microsoft.com

§Microsoft Research, email: eleanor.dillon@microsoft.com

1 Introduction

Many econometric techniques aim to estimate causal effects from existing, nonrandomized data. These methods are an important tool to answer applied problems in economics, social science, and business where randomized experiments are infeasible or unethical. Unlike prediction problems, where estimates can eventually be compared with realized outcomes, causal effect estimates typically have no ground truth to compare against, leaving open questions about their practical reliability. Evaluation of these methods against a small set of existing benchmarks, starting with LaLonde's seminal 1986 examination of a job training program, have produced mixed results.

Our study evaluates observational causal inference methods using a new benchmark dataset from Microsoft. In this dataset, a new product feature (masked to allow the data to be released) was rolled out via a randomized experiment for a subset of users, while simultaneously being made available to other users who could choose whether to adopt it. We measure the same set of device performance outcomes in both samples, along with a rich set of user and device characteristics. This creates an ideal testing ground: we observe both an experimental "gold standard" estimate and an analogous observational sample where uptake is nonrandom. Our large and homogenous data samples improve on many of the earlier benchmarks that had limited covariate overlap between treated and untreated units in the observational sample and/or limited similarity between the observational and experimental samples. We use the richness of the data to explore the performance of a range of causal estimation techniques, including modern methods that leverage flexible machine learning, estimating both the average treatment effects (ATE) and conditional average treatment effects (CATE). Our goal is to provide practical advice for researchers: analyzing which causal methods perform well on nonrandomized data, how these methods can be used most effectively, and what checks and tests provide meaningful signals about the quality of estimated effects.

We estimate the causal effect of the new product feature on two outcomes related to device performance. In both cases, the simple difference in mean outcomes across users in the observational sample who did and did not adopt the new features tells the same qualitative story as the experimental benchmark: the new feature causes a small increase in the continuous performance outcome and a larger decrease in the binary performance outcome.[1] However, in both cases these naive mean differences miss the experimentally-estimated causal effects by a meaningful margin, indicating substantial endogenous selection. For the continuous outcome, we are able to remove all the selection bias and recover the experimental benchmark using the available device and user characteristics and carefully deployed causal estimation methods. For the binary outcome,

[1] The details of the performance outcomes are also masked at the request of Microsoft to enable the data to be released. Positive changes do not necessarily align with "good" outcomes.

none of our attempts come close to the experimental benchmark.

Our analysis yields several bits of advice for practitioners. First, observational causal estimation methods can recover true causal effects when used carefully, but the details of estimation matter. When we follow established best practices, we find that doubly robust estimators that use flexible machine learning to capture the effects of confounders on treatment and outcome outperform simpler methods like linear regression. However, this flexibility can become a liability when those best practices are not followed. Most machine learning models have hyperparameters that shape the estimation to prevent overfitting in finite samples.[2] These hyperparameters should be tuned to each data sample using sample splitting to measure out-of-sample performance. When we skip this preliminary step, doubly robust methods can deliver estimated causal effects from the observational sample that are as biased as the unconditional means–indicating that choosing the right estimation model is as important as measuring the right set of confounders.

To our knowledge, this is the first real-world application that quantifies the importance of estimating the effects of confounders on treatments and outcomes precisely, supplementing the simulated illustrations in Bach et al. (2024); Chernozhukov et al. (2018). We also reinforce the importance of trimming the observational sample to remove "always treated" or "never treated" types (Dehejia and Wahba, 1999; Crump et al., 2009; Imbens and Xu, 2025) and averaging across estimates to reduce model uncertainty (Breiman, 1996; Ritzwoller and Romano, 2025). Fortunately, each of these best practices describe concrete, verifiable steps that can be executed with confidence in applied settings, even those without ground truth benchmarks.

The methods we use to estimate causal treatment effects from nonrandomized data all rely on an unconfoundedness assumption (Rosenbaum and Rubin, 1983): that we have measured enough covariates to control for any dependence between treatment assignment and the realization of outcomes. Our second guidance for practitioners is to use multiple metrics to assess the plausibility of this core assumption, which cannot be tested directly. The new product release we study has a very strong effect on the binary performance metric. In the observational sample, we overstate the true treatment effect even using our best methods, estimating declines in the probability of the outcome of 54% rather than the true 43%, indicating that we may be missing some key confounders. We conduct sensitivity analysis following Chernozhukov et al. (2022), commonly used as one check of unconfoundedness. Our estimated treatment effects on the binary outcome strongly "pass" the most common test: they are unlikely to be pushed all the way to zero by an unobserved confounder because the effect is too strong. When only the existence and direction of the effect matters, this check may provide sufficient reassurance. However, when magnitudes matter, for example when treatment

[2]Examples include the free parameters that determine the degree of regularization in LASSO regression or the maximum depth of trees in a random forest model.

is costly, we recommend additional checks including considering the fit of the preliminary prediction models (our available confounders explain none of the variation in the binary outcome) and additional sensitivity checks against benchmarks other than zero.

Finally, we use the scale and richness of our data to conduct a novel real-world test of conditional average treatment effect estimators. We validate the DR-score (Chernozhukov et al., 2024) method to detect heterogeneous treatment effects. In both the experimental and observational samples, this test suggests meaningful variation in treatment effects on the binary outcome but not on the continuous outcome. Our preferred CATE estimators also identify the same user and device features driving the heterogeneity in both the experimental and observational samples, for example devices that are used more frequently experience larger declines in the binary outcome, although every sub-group displays the same persistent selection bias when comparing estimates in the nonrandomized sample to the experimental benchmark.

Our study joins a series of validation papers using other paired observational-experimental datasets, starting with LaLonde (1986) and Friedlander and Robins (1995). Heckman et al. (1997, 1998) identified key conditions for credible observational estimates—such as detailed longitudinal covariate data and geographically proximate comparison groups—to mitigate selection biases. The fundamental importance of comparability across samples was further illustrated in a series of papers exploring the potential and limitations of matching methods (Dehejia and Wahba, 1999, 2002; Hotz et al., 2005; Smith and Todd, 2005). Imbens and Xu (2025) reviews these earlier studies and re-analyses LaLonde's sample, connecting the earlier matching studies to the general principle of overlap (Rubin, 1974). Our new data sample, which is larger than earlier benchmarks with greater overlap, allows us to reinforce some of these earlier findings with more precision, as well as adding novel insights on the importance of design-stage choices and exploring sub-group conditional average treatment effect estimation.

In section 2, we introduce our paired datasets. In section 3, we outline our formal set-up and the various estimators we consider. In section 4, we discuss our main results using the continuous outcome in our new data sample as the benchmark and illustrate the importance of estimation design choices. Section 5 re-tests our core methodologies using the binary outcome in our new data sample as well as LaLonde's job training sample. Section 6 evaluates tests of unconfoundedness and heterogeneous treatment effects, while section 7 concludes.

2 Introducing a New Dataset

We evaluate the capacities of causal analysis methods using a new data set describing the performance and characteristics of a sample of computers running Windows. The treatment of interest is a binary indicator of whether each device had a new feature enabled. The details of the studied feature are masked so the data can be released publicly. Examples of common feature updates in Windows include changes to the layout of the desktop and taskbar, new included applications or capacities, and changes in security procedures. The data cover the first two weeks of December, 2022. At that point, the new feature had been available to Windows users for some months, but users had to manually switch on the feature through settings and very few had done so.

To evaluate the feature before broader release, Microsoft ran an experiment on a sample of existing, privately owned Windows devices that had not previously turned on the feature. Microsoft switched on this feature for these two weeks for a random subset of this experimental sample. The remaining experimental devices did not have the option to turn on the feature for the two weeks of the experiment. We pair this experimental data sample with an observational sample covering the same two weeks in December. The devices in the observed sample were selected on all the same characteristics as the experimental sample except whether they had turned on the new feature. Some devices in this observed sample had already manually opted in to enabling the new feature, and kept the new feature on over the two weeks, while others had not. We therefore have two broadly similar samples: an experimental sample with exogenously assigned treatment assignment, and an observational sample where treatment is endogenously selected by device users.

Table 1 describes the full set of features in the data sample. In addition to the binary new feature indicator, we measure two outcomes related to device performance, one binary and one continuous, and a variety of features describing the technical specifications and patterns of use of the devices included in the sample. While we do not observe lagged outcomes in this sample, several of the included features summarize patterns of past user behavior that could drive endogenous selection into turning on new features. To prepare these data for public release we have masked some features (for example scrambling geographic segment and anonymizing device manufacturer), combined some rare entries in categorical features into an "other" category, and truncated continuous variables at the 97th percentile.

The observational and experimental samples are well balanced across all features, as shown in Appendix Table A.1. Treatment is rare in both samples, with only 4.5% of devices opting into the new feature in the observational sample and only 2.6% of the experimental sample selected for treatment. Appendix Table A.2

illustrates that the observational data sample captures some important elements of endogenous selection into treatment. Devices with users who opted into the new feature were used for more seconds per week before the experiment, both overall and in every internet browser category. These devices were also substantially more likely to be classified in the highly engaged segment and not in the low engagement segment and showed systematic differences in manufacturer and other device specifications. In contrast, the experimental sample passes the standard balance test for randomization, with only one pre-treatment feature barely crossing the threshold of more than a 0.1 standard deviation difference in means between the treated and control devices.[3]

3 Methodology and Estimation

In this section we introduce precise definitions of the treatment effects of interest. We evaluate multiple estimators of these causal treatment effects, all of which rely on two key assumptions for identification: unconfoundedness and overlap.

3.1 Estimands of Interest

Using the potential outcomes framework (Rubin, 1974), we can define two outcomes for each unit $i = 1, \ldots, N$. $Y_i(1)$ represents the outcome of the individual had she activated the new feature, while $Y_i(0)$ represents her outcome had she not. The causal effect for unit i is the difference in these potential outcomes. Due to the "fundamental problem of causal inference" (Holland, 1986), we do not observe both potential outcomes and hence this individual causal effect cannot be directly calculated. Let $D_i \in \{0, 1\}$ denote the treatment status, i.e. whether individual i participates in the new feature. We observe the outcome given by $Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$.

One feasible object of interest is the average treatment effect (ATE), the average effect of treatment across the population.

$$ATE = E[Y_i(1) - Y_i(0)]. \tag{3.1}$$

Another object of interest might be the average treatment effect on the treated,

$$ATT = E[Y_i(1) - Y_i(0) \mid D_i = 1]. \tag{3.2}$$

If the set of observations that are treated in the data used for estimation is representative of all units that

[3]Note that the observational and experimental samples are different on one dimension; the experimental sample was drawn only from devices whose users had not yet opted to the new feature, i.e., the untreated segment of the observed sample. However, opting into the new feature prior to the experiment was sufficiently rare and, as we show in the following section, imperfectly predictable so that there are no significant differences in the samples overall.

are ever likely to be treated, then the average treatment effect on the treated is the most policy-relevant estimand. If decision makers would like to consider the impact of a universal treatment policy, the ATE is more appropriate.

If all units respond in the same way to treatment (homogeneous treatment effects), then ATE=ATT. In contrast, if different units respond differently to treatment, then the group of units treated in the estimation data is only one sub-group of interest. We can also consider the more general case of a conditional average treatment effect (CATE),

$$\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x], \tag{3.3}$$

which captures heterogeneity across any set of covariates X. Estimating the CATE allows researchers to examine subgroup effects and personalize policy recommendations.

3.2 Identifying Assumptions

To identify these estimands, we rely on two core assumptions.

Unconfoundedness: Also known as selection on observables, this key identifying assumption requires that treatment assignment is independent of the potential outcomes conditional on observed covariates X (Rosenbaum and Rubin, 1983):

$$Y(0), Y(1) \perp D \mid X. \tag{3.4}$$

This assumption ensures that conditional on X, treated and untreated units are comparable. A weaker version suffices for ATT estimation:

$$Y(0) \perp D \mid X. \tag{3.5}$$

Since unconfoundedness is fundamentally untestable, researchers often conduct tests such as sensitivity analyses to examine its plausibility (Abadie and Imbens, 2006), as we discuss further in section 6.1.

Overlap: This assumption ensures that all units have a nonzero probability of receiving treatment:

$$0 < P(D = 1 \mid X) < 1 \quad \forall X. \tag{3.6}$$

Overlap ensures that it is possible to use untreated units to construct an estimate of the counterfactual outcome for each treated unit, and vice versa, without extrapolation. Extrapolation can lead to meaningful bias (Imbens and Xu, 2025). For the ATE, this *common support* condition must hold across the entire covariate distribution. For the ATT, it is only necessary that treated units could plausibly have received the

control. Formally, this requires $P(D = 1 \mid X) < 1$ for all X, so that every treated unit has a corresponding control counterpart. Areas of the covariate space populated only by control units are allowed, since they are not needed to identify the ATT.

Fortunately, this condition is easy to assess using estimates of the propensity to be treated, $e(X) \equiv P(D = 1 \mid X)$. If overlap is not met for the full data sample, researchers can estimate treatment effects on a trimmed subsample of data for which overlap is satisfied. For all of the estimators we consider, we begin with this preliminary step, following the optimal trimming rule designed by Crump et al. (2009) where the trimmed sample $\mathbb{A}$ is selected from the full sample $\mathbb{X}$

$$\mathbb{A}^* = \{x \in \mathbb{X} \mid \alpha \leq e(X) \leq 1 - \alpha\}$$

and α is chosen to minimize the asymptotic variance of treatment effect estimates.

After trimming, we can achieve an unbiased estimate of a slightly different estimand: the average treatment effect within the region with strong covariate overlap. This shift reduces sensitivity to model misspecification and improves internal validity at the potential cost of some external validity. This overlap ATE may vary meaningfully from the population ATE, particularly in settings such as ours where treatment is rare. Overall, an unbiased estimate of this subgroup effect is likely preferable to an estimate of a population ATE that is biased in an unknown direction, but researchers should keep the distinction in mind when interpreting their results.

3.3 Methods

To estimate treatment effects, we employ a range of estimators, each employing different strategies to control for confounding. Modern econometric estimators leverage advancements from machine learning to flexibly estimate *nuisance* functions and maximize the predictive power of the set of observed confounders. Specifically, most of the methods we evaluate require estimates of the conditional mean outcome function $\mu_d(X) \equiv \mathbb{E}[Y \mid D = d, X]$, the propensity score $e(X)$, or both.

3.3.1 Outcome Modeling

One strategy for estimating causal effects under unconfoundedness is regression adjustment, or outcome modeling, which aims to estimate the conditional mean of potential outcomes given covariates. Formally, for a binary treatment, this involves modeling two conditional response functions, $\mu_0(X)$ and $\mu_1(X)$. The simplest and most widely used implementation is linear regression, which assumes each conditional response

surface is linear, $\mu_0(x) = X'\beta$), and additive, $\mu_1(x) = \mu_0(x) + \tau$ and $\mu_0(x) = X'\beta$. When these conditions hold, the coefficient on the treatment indicator in a regression of the form $Y_i = \alpha + \tau D_i + X_i'\beta + \varepsilon_i$ consistently estimates the average treatment effect (Angrist and Pischke, 2009a). The linear regression model is simple to implement, easily interpretable, and computationally efficient. However, it relies on strong assumptions and may be biased when these conditions do not hold.

To allow for treatment effect heterogeneity or better finite-sample performance, researchers may instead model $\mu_0(X)$ and $\mu_1(X)$ separately. This two-model approach, often referred to as the Oaxaca-Blinder estimator (Kline, 2011), estimates treatment effects as the difference in these two predicted outcomes for the ATE,

$$\hat{\tau}^{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right], \tag{3.7}$$

or as the difference between realized and predicted counterfactual outcomes for the ATT,

$$\hat{\tau}^{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} \left[Y_i - \hat{\mu}_0(X_i) \right], \tag{3.8}$$

where N_1 denotes the number of treated units.

This more flexible outcome modeling can be efficient when the outcome models are correctly specified (Hahn, 1998). However, because outcome modeling relies on the correct specification of the response surface(s), errors in the functional form can lead to significant bias (Imbens, 2004). This motivates using semiparametric or nonparametric methods to estimate these functions, including machine learning approaches which offer greater flexibility in high-dimensional or nonlinear settings (Athey et al., 2019).

3.3.2 Propensity Score Modeling

Another class of estimators relies on propensity score modeling. Researchers can use estimated propensity scores, $e(X_i)$, to construct counterfactuals for treated units by pairing them with control units that had a similar predicted probability of treatment (Rosenbaum and Rubin, 1983; Abadie and Imbens, 2016). This propensity score matching reduces potentially high-dimensional covariate sets into a single, salient statistic, balancing treatment and control groups while maintaining computational feasibility. Under unconfoundedness and overlap, the propensity score is a sufficient statistic for deconfounding (Rosenbaum and Rubin, 1983).

Let $\mathcal{M}_K(i)$ denote the indices of the K nearest neighbors (in propensity score) of unit i from the opposite

treatment group. The matching estimators are:

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(Y_i - \frac{1}{K}\sum_{j\in\mathcal{M}_K(i)} Y_j\right) \tag{3.9}$$

$$\hat{\tau}_{ATT} = \frac{1}{N_1}\sum_{i:D_i=1}\left(Y_i - \frac{1}{K}\sum_{j\in\mathcal{M}_K(i)} Y_j\right). \tag{3.10}$$

While intuitive, PSM relies on correctly specifying the propensity score model, and its performance depends on overlap between treated and untreated units. Inadequate overlap can lead to poor matches, Balance diagnostics and trimming can help mitigate these issues (Dehejia and Wahba, 1999). Matching *with* replacement and caliper restrictions further improve robustness when common support is limited (Dehejia and Wahba, 2002).

Another approach using propensity scores is inverse propensity weighting (IPW). This method creates a weighted pseudo-population of units using the estimated propensity score $\hat{e}(X_i)$ (Hirano et al., 2003):

$$\hat{\tau}^{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)}\right]. \tag{3.11}$$

$$\hat{\tau}^{ATT} = \frac{1}{N_T}\sum_{i:D_i=1}\left[Y_i - \sum_{j:D_j=0}\frac{\hat{e}(X_j)}{1-\hat{e}(X_j)}Y_j\right]. \tag{3.12}$$

While IPW is unbiased under correct specification, large or unstable weights due to extreme $\hat{e}(X_i)$ values can increase variance and degrade finite-sample performance, making trimming or stabilized weights necessary in practice (Crump et al., 2009). Nonparametric estimation of the propensity score (e.g., via random forests or gradient boosting) is now commonly used to relax modeling assumptions (Li et al., 2018).

3.3.3 Augmented Inverse Probability Weighting

Augmented inverse probability weighting (IAPW) combines IPW and outcome modeling, producing a doubly robust estimator (Imbens, 2004; Glynn and Quinn, 2010):

$$\hat{\tau}^{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{D_i}{\hat{e}(X_i)} - \frac{1-D_i}{1-\hat{e}(X_i)}\right)(Y_i - \hat{\mu}_{D_i}(X_i)) + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right]. \tag{3.13}$$

$$\hat{\tau}^{ATT} = \frac{1}{N_T}\left[\sum_{i:D_i=1}(Y_i - \hat{\mu}_0(X_i)) + \sum_{i:D_i=0}\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}_0(X_i))\right]. \tag{3.14}$$

AIPW earns its namesake as a "doubly-robust" estimator as it remains consistent if *either* $\hat{e}(X_i)$ or $\hat{\mu}(X_i)$ is correctly specified. This method often outperforms single model approaches in finite samples (Robins et al., 1995; Imbens, 2004).

3.4 Conditional Average Treatment Effect Estimation

For personalized policy insights and subgroup analysis, researchers are often interested in heterogeneous treatment effects $\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$. Our evaluation focuses on a family of estimators known as meta-learners, which reframe CATE estimation as a sequence of supervised learning problems and adapt flexible machine learning models to this causal setting (Künzel et al., 2019). Meta-learners use propensity scores and/or outcome models to construct pseudo-labels in structuring the learning task. Our primary CATE estimator is an extension of the doubly robust AIPW estimator.

The CATE DR-learner (Kennedy, 2023) constructs doubly-robust pseudo-outcomes by combining estimates of the outcome regression $\mu_D(X) = \mathbb{E}[Y \mid D, X]$ and the propensity score $e(X) = \mathbb{P}(D = 1 \mid X)$. The pseudo-label for unit i is given by:

$$Y_i^{DR} = \left(\frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \right) (Y_i - \hat{\mu}_{D_i}(X_i)) + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i). \tag{3.15}$$

We fit a regression model to predict Y_i^{DR} using covariates X_i, yielding an estimate of $\tau(X)$. Under standard regularity conditions, this CATE DR-learner is consistent and has the familiar, favorable doubly-robust property where the estimate is first-order insensitive to nuisance function estimation error. We also estimate the CATE with other meta-learners like the R- (Nie and Wager, 2020) and X- (Künzel et al., 2019) learners for comparison.

4 Recovering Experimental Causal Effects

We analyze the product release data to assess how successfully each of the causal inference techniques described in the previous section can recover experimental ground truth causal effects. We begin by considering only the treatment effect of the product release on the continuous outcome, which we consider to be the cleanest testing ground. We explore the second binary outcome, as well as additional data samples, in section 5.

4.1 Establishing Ground Truth

Figure 1 displays the distribution of the estimated probabilities of receiving the new product, estimated using our preferred approach of an ensemble of tuned, boosted trees with cross-fitting (Friedman, 2001). Panel (a) of Figure 1 shows limited predictive power in the experimental sample, with mass concentrated around the mean treatment rate and minimal separation between treated and untreated devices. This pattern is consistent with Appendix Table A.2 and a correctly randomized experiment. In contrast, the observational sample, in Panel (b), shows clear separation in the estimated propensities, consistent with endogenous selection into treatment.

Panel (c) of Figure 1 plots the propensity model $\hat{e}(X_i)$ estimated from the observed sample, which captures the process of endogenous selection into treatment, applied to the experimental sample. In other words, $\hat{e}(X_i)$ is learned from the observed sample and then used to predict the devices for both the observed and experimental samples. This distribution is overlaid with the full distribution of propensity scores in the observational sample (combining treated and untreated units). The range of ex ante propensities to endogenously select treatment are virtually identical in the two samples, highlighting their fundamental similarity[4] This full overlap of the experimental and observational samples stands in contrast to some earlier benchmarks such as the LaLonde (1986) data.

To ensure an apples-to-apples comparison, we estimate the ground truth treatment effects in the experimental sample using our preferred doubly robust ATE estimator. Adjusting for confounder imbalance is theoretically unnecessary in randomized samples, but can improve stability in finite samples, as in Wager et al. (2016). We also trim the experimental sample to the region of propensity scores with common support in the *observational* sample, using the $\hat{e}(\cdot)$ plotted in Figure 1, panel (c). This step gives us an experimental benchmark of the common support ATE, which is all we can estimate in the observational sample and which may differ from the full population ATE if treatment effects are heterogeneous.

4.2 Main Results: Following Best Practices

In the trimmed samples, we can recover the experimental baseline from the observed sample using most causal estimation methods introduced in the previous section. We estimate the average treatment effect (ATE) using the five estimators discussed in section 3: linear regression (Reg), Oaxaca-Blinder outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust AIPW estimation (DR). We plot these estimates against the experimental baseline in Figure 2. All five estimators include the ground truth within their 95% confidence bands, though some hit closer with their

[4]We test similarity feature by feature in Appendix Table A.1.

point estimates than others.

In addition to sample trimming, the estimates in Figure 2 reflect several other best practices for practical causal estimation: careful selection of the nuisance models and model ensembling. For the "nuisance models" we use to construct our estimators, $\mu_d(X) \equiv \mathbb{E}[Y \mid D = d, X]$ and $e(X) \equiv \mathbb{E}[D \mid X]$, we compare multiple models, including flexible machine learning models with hyperparameters tuned using AutoML procedures to maximize predictive performance, and choose the ones with the best fit. Finally, instead of relying on a single model for each nuisance, we ensemble across multiple models to mitigate model uncertainty. Following Breiman (1996), averaging across multiple learners reduces variance and stabilizes treatment effect estimation.

The remainder of this section illustrates the costs of deviating from each of these best practices. While we find that many causal estimators can recover experimental benchmarks following all of our recommended practices, the following sections show the DR estimator is generally the most robust to deviations from these best practices. We therefore treat this estimator as our baseline when exploring extensions to our main results.

4.3 Sample Trimming

Ensuring common support between treated and control units is critical for estimation. In settings with poor overlap, pruning the data to a region of overlap is often more important for credibility than the specific choice of estimation technique (Dehejia and Wahba, 1999, 2002). The resulting estimand is the average treatment effect for the trimmed subpopulation rather than the full-sample ATE or ATT. We follow Crump et al. (2009) and arrive at an optimal trimming threshold of $[0.026, 0.974]$.[5] As shown in Table 2, these cutoffs remove 32% of the observational sample, consisting of 8% of treated units and 33% of control units. Implementing the same cutoffs in the experimental sample drops a similar share of total observations, as we would expect from Figure 1 panel (c), but in this case balanced between treated and untreated units.

Figure 3 reports ATE estimates and 95% confidence intervals for the full, untrimmed sample, following our other best practices. The naive difference-in-means estimate is substantially farther from the experimental benchmark in the full sample. While all estimators close most of this gap, they are still biased somewhat upwards from the true treatment effect. Only the doubly robust estimator still contains the point estimate of the ground truth within its confidence band.

[5] In practice, given the substantial treatment imbalance in our setting, trimming is effectively one-sided.

4.4 Nuisance Model Tuning

Modern causal estimation methods like doubly robust learners outperform linear regression in our baseline comparisons because they use flexible machine learning models to better capture the relationships between confounders, treatment, and outcome. However, their performance is dependent on finding models that actually capture these relationships accurately. Flexibility can become a liability if the hyperparameters that control these models are chosen poorly and the models overfit to the training sample (Bach et al., 2024).

Table 3 illustrates the out-of-sample performance of various models when estimating each of the "nuisance" functions required for doubly robust estimation: the predicted outcome for the treated and untreated samples, $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$, and the probability of treatment, $\hat{e}(X)$. The models are listed in descending order of their average performance across the three functions. Our preferred models are a well-tuned light gradient-boosted machine (LGBM) classifier to predict treatment, and light gradient-boosted machine regressors to predict the continuous outcome.[6] Using tuned LGBM, we can predict close to 13% of the variation in the continuous outcome for treated and untreated units in the trimmed sample. These models also do the best job predicting the probability that each unit is treated, although this top performance still yields a modest AUC of just under 0.6. The performance of the propensity score model is consistently higher in the full sample. This pattern is unsurprising since the trimmed sample excludes groups of devices with very high or very low treatment shares, which are also the easiest to predict.

The remaining rows of Table 3 illustrate a variety of alternative methods for estimating the nuisance functions.[7] In the product release data, an untuned LGBM performs nearly as well as the tuned version, though we expect this is a coincidence of the default choices rather than a general feature of these models. Tuned random forest models also perform nearly as well as our baseline LGBM, though even a tuned neural net does not. Logistic/linear regression and penalized Lasso models perform notably less well, indicating that the linearity assumptions these models impose is not the best fit for our data. The performance of the tuned Lasso is nearly identical to the unpenalized logistic/linear regression models; for this relatively long and narrow data sample, the optimal Lasso penalty is close to zero. However, the bottom of Table 3 is dominated by flexible models estimated at their default hyperparameter values. Without careful handling,

[6] For LGBM and Lasso, we employ Microsoft's FLAML, a Bayesian optimization framework for hyperparameter tuning. We tune each nuisance function estimator for up to 100 seconds. For random forests and neural networks, we use Optuna since the current version of FLAML (as of this writing), has a limited hyperparameter search space for these models. Neural networks are never truly "untuned" in practice, and their weak performance on tabular data is well-documented. Our "weakly tuned" neural network models perform only a cursory hyperparameter search to illustrate this point; see Appendix Table B.1 for full details.

[7] The full propensity score distributions for these models are plotted in Appendix Figures B.1 and B.2. Without tuning, propensity score estimates vary widely across models, often failing to meaningfully separate treated and untreated groups.

these more complex models predict much less of the variation in both treatment and outcomes than a simple linear regression.

To quantify the final impact of modeling choices, we re-estimate the doubly robust estimator across this range of first-stage nuisance models. Figure 4 plots the resulting ATE estimates, with the models again ordered by average performance. Across models, we observe a near-monotonic relationship between first-stage predictive quality and downstream causal estimate quality: models with higher R^2 and AUC scores produce estimates closer to the experimental benchmark. The downstream effects of nuisance model performance are quantitatively important; the worst-fitting models do no better than the unconditional difference-in-means estimator.

This exercise illustrates that choosing the right functional form for causal analysis is as quantitatively important as choosing the right set of conditioning variables. Happily, the best functional form is easy to identify. Modern methods that use machine learning to estimate underlying relationships make this step easier. Rather than ad hoc choices by the econometrician on what potential confounders to include and how to featurize them, ML methods identify complex, non-linear relationships automatically and allow practitioners to err on the side of inclusion[8] while avoiding overfitting through cross-estimation and penalization. For a given data sample, choosing functional forms that maximize predictive performance in the nuisance models is likely to minimize bias in the treatment effect estimates. We recommend trying multiple models, including tuned, flexible ML models, or using an AutoML algorithm to select across many models.

4.5 Nuisance Model Stability

Our preferred estimation approach involves two rounds of arbitrary sample splitting. We first use cross-validation to tune the hyperparameters for each nuisance function. Then, we cross-fit those nuisance functions within the doubly robust estimation to enable valid confidence interval construction in the final stage (see, i.e., Chernozhukov et al. (2018)). The previous section illustrates the importance of this approach for satisfying the unconfoundedness assumption. However, these cross-fitting steps also introduce a new, and often overlooked, source of uncertainty. For each iteration of cross-fitting or cross-validation, the data sample is randomly split into equal sized folds. Even when using well-tuned learners, different splits of the data can yield substantively different nuisance models (Ritzwoller and Romano, 2025). Even if performance metrics for nuisance models are similar across runs, this variation can carry into downstream causal estimates. This *model uncertainty* is distinct from *sample uncertainty*, which arises because we estimate treatment effects on

[8] As long as the conditioning set only includes potential confounders and not "bad controls" such as colliders or downstream outcomes. See Pearl (1995) or Angrist and Pischke (2009b) for more discussion.

a finite sample of data drawn from a larger population of interest and is captured via standard errors.

In our recommended approach, we ensemble 50 tuned LGBM models for each of four preliminary models: the initial propensity model used to trim the sample, which we denote $e^*(X)$, the propensity model re-estimated on the trimmed sample within the doubly robust algorithm, $e(X)$, and the two outcome models, $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$. As discussed in Breiman (1996), averaging across models improves predictive accuracy and mitigates downstream instability. In Figure 5, the point estimate for this fully ensembled baseline is marked by the red dotted line. The remainder of the figure illustrates the potential instability created by foregoing this ensembling. The first shape plots the range of treatment effects estimated using each of the 50 single runs of the outcome models while continuing to average across runs for the two propensity models. The remaining shapes conduct similar exercises, plotting the distribution when:

- Ensembling $\hat{e}^*(X)$, $\hat{\mu}_0(X)$, and $\hat{\mu}_1(X)$, while varying $\hat{e}(X)$ over 50 trials.
- Ensembling $\hat{e}^*(X)$ and varying $\hat{e}(X)$, $\hat{\mu}_0(X)$, and $\hat{\mu}_1(X)$ over 50 trials.
- Varying all components over 50 trials.

Figure 5 illustrates the real variation that can occur between runs of these models with different seeds, particularly the last shape where none of the models are ensembled, a common starting point in empirical work. Comparing this figure with our baseline estimates in Figure 2 illustrates that the instability introduced by random sample splitting is considerably smaller than our standard errors, which capture the uncertainty created by drawing the initial data sample from the population. This is also a smaller source of instability than the model selection discussed in the previous section. Ensembling across multiple runs of the nuisance functions can add considerably to run times, particularly when each run includes re-tuning hyperparameters, as we believe it should. For high-stakes estimates where precision matters, this extra time can be a worthwhile investment.

5 Additional Samples

In the previous section, we illustrate the importance of three key elements of our recommended best practices using the estimated effect of a new product release on one continuous outcome in our data sample. We now briefly explore how robust those recommendations are across different outcome types and datasets

5.1 Product Release Impact on Binary Outcome

We first consider the impact of the same new product release on a second, binary outcome in the same data sample. We use the same trimmed sample as in the continuous outcome analysis, since trimming depends only on estimated treatment propensity and is independent of the outcome. However, as illustrated in Figure 6, for this outcome even the best practices miss the experimental benchmark by a wide margin; in fact, they are barely distinguishable from the naive difference-in-means.

This systemic pattern of bias across many estimates of the treatment effect suggests the presence of one or more unobserved confounders. Consistent with this hypothesis, we do a poor job of predicting the binary outcome with our set of available measured confounders. The AUC for our preferred $\hat{y}_1$ model is 0.54 in the trimmed observational sample, barely better than random guesses.[9] This finding illustrates an important reminder: well-fit, flexible ML models make it easier to achieve unconfoundedness by extracting all the available signal from a set of conditioning variables. However, even the best methods cannot overcome important omissions from this set.

5.2 Re-analyzing LaLonde

Imbens and Xu (2025) deploy most of our suggested best practices to re-analyze the worker training samples introduced in LaLonde (1986). For comparison, we briefly repeat the exercise here. We follow the convention in Dehejia and Wahba (1999), referring to the reanalyzed LaLonde samples as LDW-CPS and LDW-PSID, which pair experimental treated units with controls from the CPS and PSID, respectively. We apply four estimators (doubly robust, propensity score matching, inverse propensity weighting, and outcome modeling) using Crump et al. (2009) trimming, AutoML tuned nuisances, and model ensembling. To ease comparison with much of the long line of earlier studies using these data samples, we estimate the average treatment effect on the treated (ATT) for these samples, rather than the population ATE we consider for the product release sample.

Figure 7 illustrates the results of this exercise, which confirm the findings of Xu and Imbens. In the trimmed LDW-CPS sample, all four estimators yield ATT estimates between approximately \$1,100 and \$2,400, and all fall within the experimental benchmark confidence band (centered at \$1,761). The LDW-PSID results display greater heterogeneity. The PSM estimator is highly negative (near -\$1,300), consistent with the instability reported in earlier studies. IPW, OM, and DR estimators center around \$1,000, well below the experimental benchmark of \$1,775, but have wide confidence intervals that contain the benchmark.

[9] The AUC to predict the binary outcome for untreated units is higher, 0.82, but it is difficult to measure performance for this group as nearly 99% of units have a positive outcome (see Appendix Table A.2).

As discussed in more detail in Dehejia and Wahba (1999) and many others, the lack of overlap in these LaLonde samples is severe. Our main departure from Imbens and Xu (2025) is in how we trim the samples to create a cleaner comparison. Xu and Imbens trim the treated workers based on rule-of-thumb thresholds, $[0.1, 0.9]$ for the LDW-CPS sample and $[0.2, 0.8]$ for LDW-PSID, and then select a 1:1 matched set of control observations. The Crump et al. (2009) procedure we follow selects propensity score thresholds of 0.085 and 0.915 for the LDW-CPS sample and 0.109 and 0.891 for the LDW-PSID sample. These thresholds allow us to retain somewhat more of the sample, particularly for the LDW-PSID sample, while matching or slightly improving the performance of most estimators.

6 Extensions and Caveats

We close by extending our main analysis along several dimensions. First, assess the ability of sensitivity analysis and other tests to detect when the unconfoundedness assumption required by all our methods is likely to be violated. Next, we explore heterogeneous treatment effect estimation, assessing whether modern methods can reliably recover patterns of treatment effect variation across covariates.

6.1 Sensitivity Analysis

In the previous sections we explored the effect of a new product release on two outcomes. For the continuous outcome, we are able to recover experimental benchmark effects using careful causal estimation methods. For the binary outcome, no method recovers the experimental benchmarks. We hypothesize that this failure is consistent with an unobserved feature that confounds only the relationship between the product release and the binary outcome. There are no direct tests for the presence of such a confounder. Sensitivity analyses (Cinelli and Hazlett, 2019; Chernozhukov et al., 2022) can illustrate how robust our estimates are to a hypothetical unobserved confounder.

Our sensitivity analysis follows Chernozhukov et al. (2022) and quantifies how strongly an omitted confounder U would need to be related to the outcome Y and treatment D in order to overturn our causal conclusions. Three key quantities characterize this relationship:

- C_Y^2. The nonparametric partial R^2 of an omitted confounder U with the outcome Y, conditional on observed covariates X and treatment D:

$$C_Y^2 \;=\; \frac{\mathrm{Var}(\mathbb{E}[Y|D,X,U]) - \mathrm{Var}(\mathbb{E}[Y|D,X])}{\mathrm{Var}(Y) - \mathrm{Var}(\mathbb{E}[Y|D,X])}$$

This measures the additional share of unexplained variation in Y that could be explained by U once X and D are controlled for.

- C_D^2. The relative gain in precision in predicting treatment assignment when U is added to the observed covariates:

$$C_D^2 \;=\; \frac{\mathbb{E}\left[\left(P(D=1|X,U)(1-P(D=1|X,U))\right)^{-1}\right] - \mathbb{E}\left[\left(P(D=1|X)(1-P(D=1|X))\right)^{-1}\right]}{\mathbb{E}\left[\left(P(D=1|X,U)(1-P(D=1|X,U))\right)^{-1}\right]}$$

This captures how much sharper the propensity score becomes once U is observed.

- ρ. The correlation between the omitted confounder's effect on Y and its effect on D, conditional on X: Values of $|\rho|$ close to 1 correspond to a confounder that moves both outcome and treatment in the same direction, generating maximum bias.

These quantities bound the potential bias between the feasible estimate, τ, and the unconfounded estimate that conditions on U, τ_S:

$$|\tau_S - \tau| \;\leq\; |\rho|\, C_Y\, C_D\, S,$$

where S is a sensitivity scaling factor. Intuitively, S rescales the bound by the variability of the outcome residuals and the informativeness of the treatment assignment, so that C_Y and C_D are measured on a unit-free scale. Full details on its empirical construction are provided in Appendix G.2.

Given (C_Y, C_D, ρ), the maximum bias in the ATE estimate can be bounded. Two summary metrics are of particular interest. First, we calculate the robustness value ($RV(\rho)$), the minimum explanatory strength of an unobserved confounder, that is a minimum value of $C_Y = C_D$, required to shift the true point estimate of the treatment effect to zero. Second, we report a weaker adjusted robustness value ($\text{RV}_\alpha(\rho)$), the minimum explanatory strength required to shift the confidence interval of the estimate to include zero. For our continuous ATE estimate, assuming a scenario where $\rho = 0.5$, we obtain: RV = 1.31% and $\text{RV}_\alpha(\rho) = 0.60\%$.[10] An omitted confounder would need to explain 1.31% of the residual variation in both Y and D to drive the point estimate to zero, or 0.6% of the residual variation in both Y and D to move the confidence interval to include zero.

These robustness values are small in absolute terms, but they are difficult to interpret without benchmarks of plausible sets of $\{C_Y, C_D, \rho\}$. We can calculate these values for each observed covariate in our sample by excluding them one-by-one and re-estimating the model. For the continuous outcome, the strongest over-

[10] While $\rho = 0.5$ is not the worst-case confounding scenario, it provides a conservative value relative to the ρ we estimate for any of our observed confounders, reported in Appendix Table G.1. Among the observed confounders, all variables with meaningful influence on the treatment or outcome have a ρ well below 0.5.

all confounder is the total device usage. For this feature, $C_Y = 0.023$, $C_D = 0.013$, and $\rho = 0.104$.[11] These influence values are well above the robustness value, suggesting that the presence of a similarly influencial unobserved confounder is plausible, although we emphasize that the ρ we estimate for this feature is well below 0.5. An unobserved confounder with $\rho = 0.1$ would generate far less bias.

The robustness values we calculate for the binary outcome are considerably higher. To move the estimate to zero, an omitted confounder would need to explain approximately 74.8% of the residual variation in both the treatment and outcome. Identifying an appropriate benchmark for these values in the binary outcome is more difficult. As we have already established, none of our observed confounders do a good job predicting the binary outcome; the highest C_Y we estimate is 0.0026 for the indicator Manufacturer=5. Instead of comparing a hypothetical unobserved confounder to this set of weak predictors of the binary outcome, we think a more reasonable benchmark is to use the same influence values as the strongest observed confounder for the continuous outcome.

Figure 8 plots range of impacts on our treatment effect estimates for each outcome of various unobserved confounders. The x-axis shows values for C_D, the y-axis values of C_Y and the isoquants plot the treatment effect we would have estimated if we were able to include an additional confounder of various strengths in our analysis. The “unadjusted" dot at the origin marks our baseline treatment effect estimate. The blue dots plot the treatment effect we would estimate if we added confounders with 1x and 2x the benchmark (C_Y, C_D) of $\mathbb{I}$\{Total_Device_Usage\} for the continuous outcome. These plots reinforce the message of the robustness values. For the continuous outcome, plotted in the left panel, a plausible unobserved confounder could easily push our estimated treatment effect past zero (indicated with a dashed red line). In contrast, for the binary outcome in the right panel, a benchmark unobserved confounder would have only modest effects, moving the estimated treatment effect from just below -0.53 to just below -0.52.

These contrasting results seem puzzling at first. Sensitivity analyses suggest that our estimated treatment effects on the continuous outcome, which we have demonstrated achieve the experimental benchmark, are highly sensitive to a hypothetical unobserved confounder. In contrast, our estimated treatment effects for the binary outcome, which exhibit considerable bias relative to the experimental benchmark, are identified as largely robust to a hypothetical additional confounder. The solution lies in details of these sensitivity tests and the magnitude of the treatment effects.

Comparing the ground truth treatment effects to the summary statistics in Appendix Table A.2 indicates that our treatment increases the continuous outcome by about 0.03 standard deviations. The baseline ATE

[11] Values for all observed confounders are reported in Appendix Table G.1. The indicator that Engagement Segment = Highly Engaged has a stronger role in predicting treatment, $C_D = 0.015$, but far less influence on the outcome, $C_Y = 0.00038$, with $\rho = 0.207$.

is 0.021, while the standard deviation of the continuous outcome is 6.7[12]. While statistically distinguishable from zero, this is a very small effect. The modest treatment effect also means that even the treatment and the other confounders combined explain relatively little of the outcome.[13] An unobserved confounder would not need to explain much of the residual variation in order to push the very small treatment effect to zero.

In contrast, our ground truth estimates from the experimental sample indicate that our treatment decreases the binary outcome by 3.6 standard deviations. Our best estimate from the observational sample suggests a decrease of 4.5 standard deviations, overstating the true effect but telling the same qualitative story. These are quantitatively important treatment effects, and the treatment drives much of the variation in the binary outcome. In the experimental sample, predicting the binary outcome using only the treatment yields an AUC of 0.86. Once we condition on both the treatment and the observed confounders, there is much less residual variation in the binary outcome. An unobserved confounder would need to explain quite a lot of it to meaningfully change our estimates.

This example highlights both the value and the limitations of this type of sensitivity analysis. In this special case, we know that the estimated treatment effect on the continuous outcome, while small, represents a robust relationship that aligns with the experimental benchmark. In the general case, however, it is worth realizing that this very small treatment effect, estimated in a setting where much of the variation in the outcome remains unexplained, could easily be sensitive to even small missing confounders. In this special case, we know that our estimated treatment effects on the binary outcome are somewhat biased and miss the experimental benchmark. However, both the experiment and the observational analysis capture the right qualitative message: the treatment has a strong and important negative effect on the outcome. The sensitivity analysis correctly identifies that adding additional confounders is unlikely to change this result. In some cases, as in the LaLonde samples analyzed in Imbens and Xu (2025), the estimates identified as most vulnerable using a sensitivity check are also the estimates that are most biased away from the experimental ground truth. This exercise is a useful reminder that this alignment is not guaranteed.

6.2 Heterogeneous Treatment Effects

Thus far we have evaluated the ability of causal inference methods to recover the average effect of treatment across the entire population of interest. However, many modern applications of causal analysis also require estimating treatment effects within various subgroups, or at least testing for the presence of treatment effect heterogeneity. The product release sample has the potential to be a good testing ground for methods that

[12] We use the standard deviation for untreated observations in the experimental sample for this exercise.
[13] From Table 3, the confounders explain 13% of the variation in the outcome.

estimate these Conditional Average Treatment Effects (CATE). The large sample size should provide enough statistical power to estimate nuanced heterogeneous effects. In the experimental sample, randomized assignment ensures a thick distribution of treated and untreated observations across the full range of covariates, providing tight benchmark CATE estimates for all combinations of features. Modern tools offer a variety of flexible meta-learners for estimating CATE, including the S-, T-, X-, DR-, and R-learners. We first estimate each of these models separately to see if any fit the data significantly better than a constant treatment effects model, then combine them to compare the results of this best-practice CATE estimator against the experimental benchmark.

We evaluate the presence of treatment effect heterogeneity for each outcome in the product release data. The DR-score (Chernozhukov et al., 2024) measures the out-of-sample reduction in residual variance of the doubly robust pseudo-outcome Y^{DR} relative to a constant CATE model. We begin by constructing the doubly robust variables,

$$Y_i^{DR}(\hat{\mu}, \hat{e}) := \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (Y_i - \hat{\mu}_{D_i}(X_i)) \frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \tag{6.1}$$

and then we train different CATE models $\hat{\tau}$ by regressing $Y_i^{DR}(\hat{\mu}, \hat{e}) \sim X_i$. For any CATE model, we define the (normalized) DR-score:

$$\text{DRscore}(\hat{\tau}) := \frac{E_n\left[\left(Y^{DR}(\mu, e) - \tau_{\text{const}}(X)\right)^2\right] - E_n\left[\left(Y^{DR}(\mu, e) - \hat{\tau}(X)\right)^2\right]}{E_n\left[\left(Y^{DR}(\mu, e) - \tau_{\text{const}}(X)\right)^2\right]} \tag{6.2}$$

A score of 0 corresponds to a constant CATE model that predicts the ATE, and a score of 1 would reflect perfect prediction of Y^{DR}. Any CATE model that captures true heterogeneity should achieve a positive score between 0 and 1.

Figure 9 displays the DR-scores for each learner for our two outcomes, separately for the experimental and observational samples. Beginning with the experimental (left) column, we see that the DR scores are negative or zero across all models for the continuous outcome, indicating that none of the CATE models can outperform a constant effects model, and in fact estimating heterogeneous effects appears to add noise. For the binary outcome, all CATE models modestly outperform the constant effects baseline, suggesting some heterogeneous effects. Reassuringly, conducting the same test in the observational sample (right column) yields the same results: the DR scores reject treatment heterogeneity for the continuous outcome and shows evidence of modest heterogeneity for the binary outcome.

While Figure 9 provides encouraging evidence on the ability of DR-scores to detect treatment heterogeneity in practice, it tells a discouraging story for the value of the product release data as a testing ground for CATE estimators. We only find evidence of variation in treatment effects on the binary outcome. However, as shown in Section 5, we cannot predict this outcome reliably from the available observed covariates and the ATE estimates in the observational sample show substantial bias relative to the experimental baseline, suggesting the presence of unobserved confounders. We proceed with an investigation of how well CATE methods used in the observational sample can replicate the *ordering* of treatment effects on the binary outcome across subgroups, with the caveat that we expect that all subgroups may show the same bias as the ATE estimates.

Our best-practice CATE estimates combine the models plotted in Figure 9. Each of these methods exhibits its own bias-variance tradeoff and susceptibility to overfitting. Q-aggregation addresses this challenge by forming a convex combination of learners that penalizes individual learners that perform poorly (Lan and Syrgkanis, 2024). Q-aggregation selects a vector of weights, w, that solves the penalized optimization problem:

$$\min_{w \in \Delta(M)} (1-\nu)\mathbb{E}_n \left[\left(Y^{DR} - \bar{\tau} - \sum_{m=1}^{M} w_m \tilde{\tau}_m(Z) \right)^2 \right] + \nu \sum_{m=1}^{M} w_m \mathbb{E}_n \left[\left(Y^{DR} - \bar{\tau} - \tilde{\tau}_m(Z) \right)^2 \right], \tag{6.3}$$

where $\tilde{\tau}_m(Z)$ denotes the centered CATE predictions from each model m, and $\Delta(M)$ is the simplex over M models. The parameter $\nu \in [0,1]$ interpolates between convex regression, or OLS regression subject to simplex constraints ($\nu = 0$) and simply selecting the best model ($\nu = 1$).

The key strength of Q-aggregation lies in its ability to penalize learners that overfit to training noise or exhibit poor out-of-sample generalization. Unlike naive averaging, which may dilute strong learners, or hard selection, which risks high variance, Q-aggregation strikes a balance, selecting a convex combination that minimizes doubly robust loss. This procedure is supported by oracle inequality guarantees (Lecué and Rigollet, 2014), and in practice, it has proven effective for stabilizing CATE performance. Q-aggregation serves as both a model selection tool and a heterogeneity regularizer by discarding models that introduce spurious heterogeneity.

To create a benchmark for CATE estimation, we first use a shallow tree to distill what the Q-aggregated CATE ensemble models have learned in the experimental sample about the treatment effect on the binary outcome. As shown in the left panel of Figure 10, the tree splits first on Device Specification A, then on Total Device Usage and an indicator of whether the device was made by Manufacturer 5. These cuts split the experimental sample into groups with substantially different average treatment effects. The group with

the most negative treatment effect has an average CATE of -0.46 (sd = 0.048). The group with the smallest effect in absolute terms has an average CATE of -0.226 (sd = 0.034).

We then split the observational sample into groups using these same feature values and consider the average CATE in each group as estimated from the observational sample. As shown in the right panel of Figure 10, the Q-aggregated CATE ensemble model is able to recreate the ordering of effect sizes across groups: the cluster with the most negative treatment effects in the experimental sample (Device Specification A is false, with high Total Device Usage) also has the most negative estimated effects in the observational sample and so on. However, as anticipated, all groups show the same systemic bias as the ATE estimates. In the observational sample, the group with the most negative treatment effects has an average CATE of -0.561 (sd = 0.027), while the group with the smallest absolute effects has an average CATE of -0.414 (sd = 0.025). Appendix Figure B.3 shows SHAP values showing the feature importance in predicting the estimated CATE in each sample. While not identical, the same device characteristics influence the CATE in similar ways across the experimental and observational samples. In other words, the observational sample is not only able to recreate a similar ordering of average CATE across the groups identified in the experimental sample, but also to identify these key splits in the absence of the experimental benchmark.[14]

Overall, we view this as encouraging evidence that CATE estimators can accurately rank units by treatment effect sizes. This ordering is important for targeting rationed or expensive treatments. Unfortunately, our data sample is not a useful test case for the ability of CATE models to recover the magnitude of heterogeneous treatment effects.

7 Conclusion

This paper provides a new empirical validation study of observational causal inference methods using a unique dataset. Unlike earlier comparisons such as LaLonde (1986), our clean paired design contains both treated and untreated units in both experimental and observational samples. This richer setup allows for a more direct and comprehensive assessment of observational methods relative to an experimental benchmark.

We find that, when modern best practices are employed—namely careful overlap trimming, AutoML-driven hyperparameter tuning, and model ensembling—observational estimators can closely recover experimental treatment effects. Across a variety of estimators, including outcome modeling, inverse probability weighting, matching, and doubly robust methods, our results show strong agreement with the experimental ground truth for a continuous outcome. These findings extend and reinforce earlier conclusions by Dehe-

[14] Appendix Figure E.1 shows the shallow tree re-estimated using only the observational sample. As expected from the SHAP values, the splits include most of the same features as the experimental tree.

jia and Wahba (1999), Crump et al. (2009), Imbens and Xu (2025), Bach et al. (2024), Breiman (1996) among others, regarding the critical role of design-stage choices in observational studies. Moreover, when we reanalyze the LaLonde dataset using our proposed workflow, we find similar improvements: applying modern trimming, tuning, and ensembling strategies helps improve alignment between observational and experimental estimates, highlighting the broad applicability of these lessons.

At the same time, our results underscore the fundamental limits of observational causal inference. Even when we successfully recover the experimental benchmark for the continuous outcome, and even when sensitivity analysis provides some credibility to these estimates, this alignment does not in any way *validate* the unconfoundedness assumption, which remains inherently untestable. The analysis of the binary outcome highlights this challenge starkly: despite applying the same best practices and observing high robustness values, flexible estimators still fail to recover the experimental benchmark. Overall, our findings contribute to the growing literature demonstrating that careful design and validation practices can greatly enhance the credibility of observational causal inference. Yet they also reaffirm that, at its core, causal identification depends as much on substantive assumptions and domain knowledge as on statistical technique.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Angrist, J. D. and Pischke, J.-S. (2009a). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2009b). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., and Spindler, M. (2024). Hyperparameter tuning for causal inference with double machine learning: A simulation study. In Locatello, F. and Didelez, V., editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1065–1117. PMLR.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2022). Long story short: Omitted variable bias in causal machine learning. *National Bureau of Economic Research Working Paper*, (30302).

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Cinelli, C. and Hazlett, C. (2019). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Friedlander, D. and Robins, P. K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *The American Economic Review*, 85(4):923–937.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Heckman, J. J., Ichimura, H., and Smith, J. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hotz, V. J., Imbens, G. W., and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1):241–270.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Xu, Y. (2025). Lalonde (1986) after nearly four decades: Lessons learned. *Journal of Economic Perspectives*. Forthcoming.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17:3008–3049.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3):532–37.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620.

Lan, H. and Syrgkanis, V. (2024). Causal q-aggregation for cate model selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4366–4374. PMLR.

Lecué, G. and Rigollet, P. (2014). Optimal learning with q-aggregation. *The Annals of Statistics*, 42(1):211–224.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Ritzwoller, D. and Romano, J. P. (2025). Reproducible aggregation of sample-split estimators. *Conditionally Accepted, American Economic Review*. Forthcoming.

Robins, J. M., Rotnitzky, A., and and, L. P. Z. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

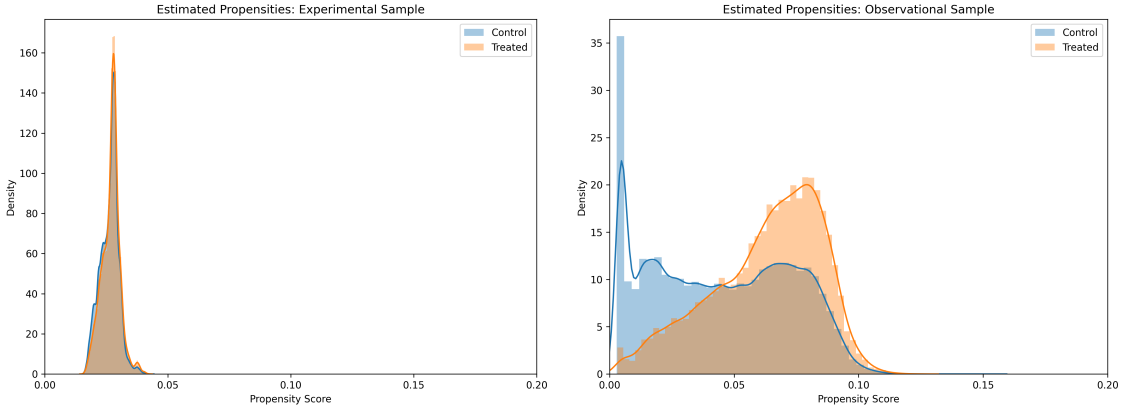
Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

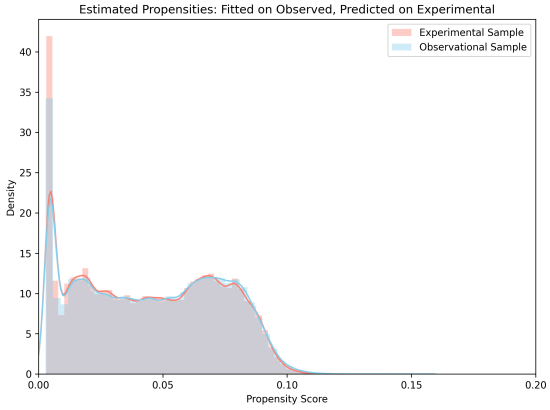
Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

Tables and Figures



(a) Experiment Sample (b) Observed Sample



(c) Observed Fit on Experiment

Figure 1: **Estimated Propensity Score Distributions Across Samples.** Panel (a) shows the distribution of cross-fitted propensity scores estimated on the experimental sample, reflecting the randomized assignment—most scores cluster around the overall treatment rate of 2.64%. Panel (b) shows the distribution of cross-fitted propensity score estimated on the observational sample, where greater separation between treated and untreated groups is evident, consistent with endogenous selection into treatment. Panel (c) plots the observational propensity model $\hat{e}(\cdot)$ evaluated on both the entire observational and entire experimental samples. The complete distributional overlap in Panel (c), coupled with the same range of scores as seen in Panel (b), indicates that the experimental and observational samples cover comparable covariate regions, allowing trimming procedures to be applied consistently across both datasets.

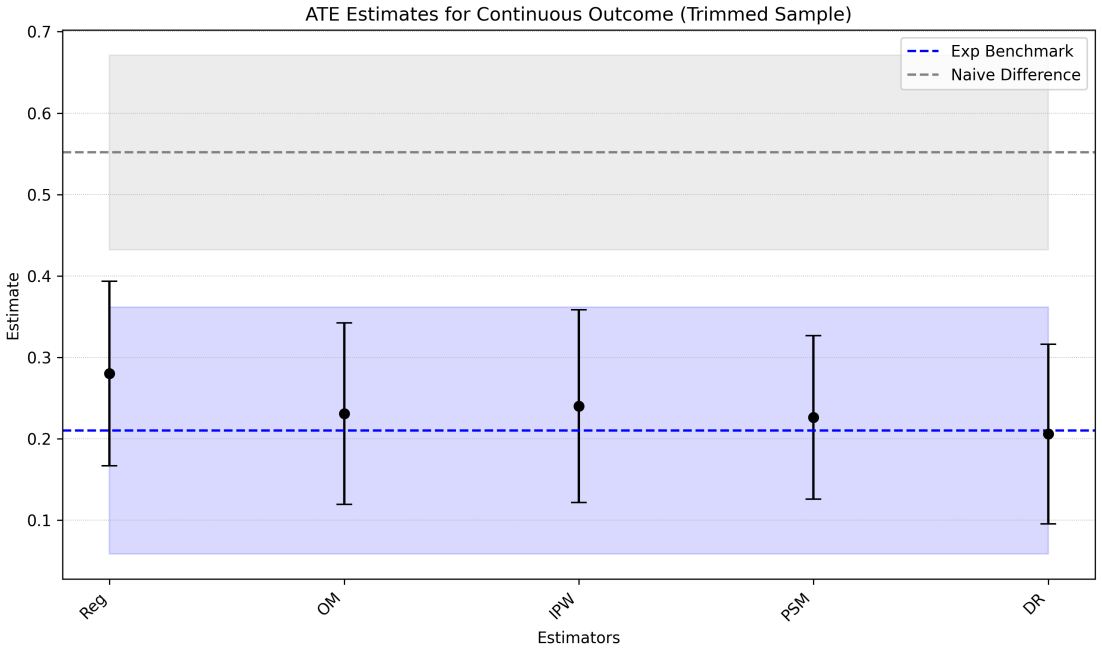


Figure 2: **ATE Estimates for Continuous Outcome (Trimmed Sample):** The figure displays ATE estimates and their 95% confidence intervals for five estimators: regression adjustment (Reg), outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust estimation (DR). After trimming, careful nuisance tuning, and model ensembling, all estimators align closely with the experimental benchmark.

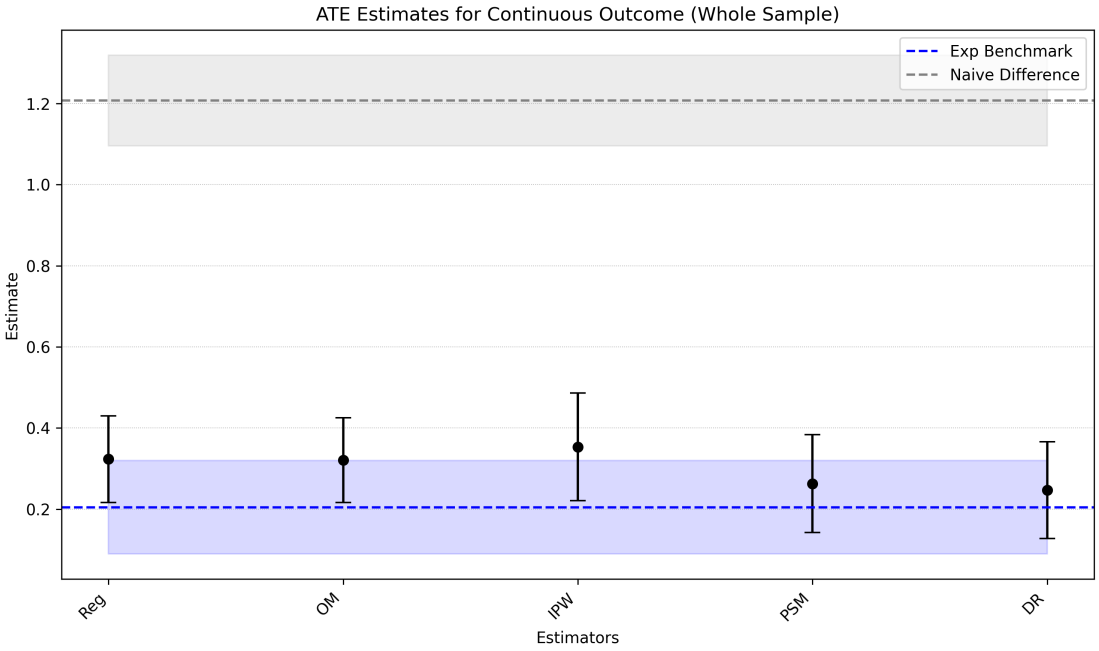


Figure 3: **ATE Estimates for Continuous Outcome (Untrimmed Sample):** The figure displays ATE estimates and their 95% confidence intervals for five different estimators: regression adjustment (Reg), outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust estimation (DR). The black dashed line and associated shaded band represent the naive difference in means and its 95% confidence interval, while the blue dashed line and associated band show the experimental benchmark and its 95% CI. Without trimming, estimates are more dispersed, with only DR encompassing the point estimate of the benchmark.

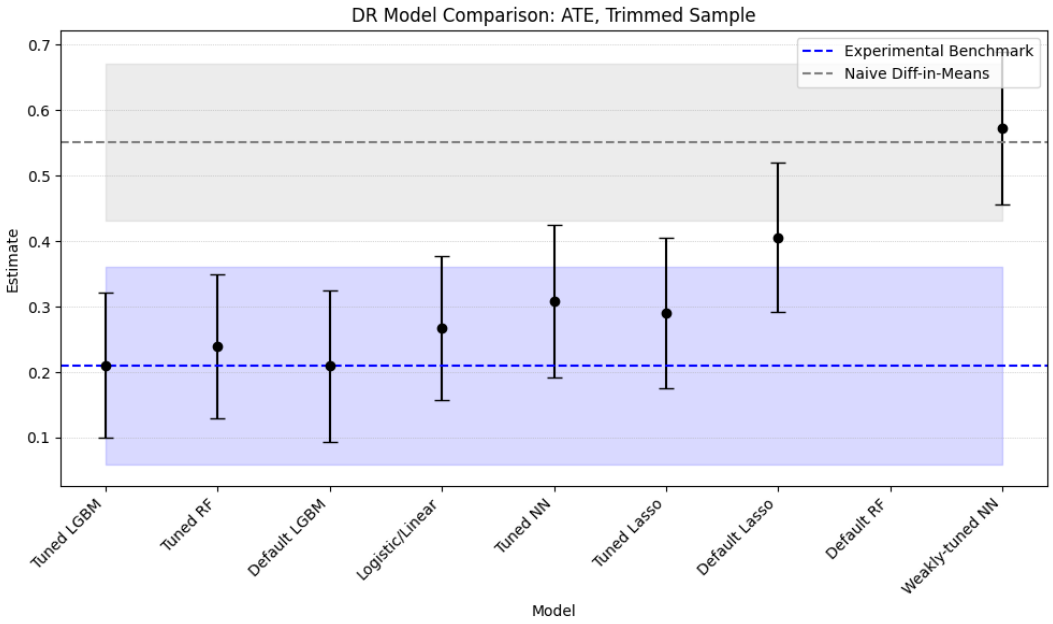


Figure 4: **DR Estimates Across Different First-Stage Models (Trimmed Sample):** This figure compares ATE estimates from doubly robust estimators using a range of flexible and rigid first-stage models for outcome and treatment modeling. Both tuned and sklearn default versions are considered. We find tuning is critical for flexible models, while its effect is less important with rigid models. Models like tuned LightGBM and tuned random forests align closely with the benchmark, while weakly tuned neural networks continue to exhibit bias. The default random forest implementation yields an estimate of -55.409 (83.877), driven by very unstable propensity weights, and is thus not pictured in the figure. In contrast to the untrimmed comparison in Figure C.1, we see trimming substantively reduces estimator variability and brings many models closer to the experimental benchmark, partially mitigating the gains from tuning. These results altogether reinforce that trimming and tuning are complementary tools for improving the reliability of observational causal estimation.

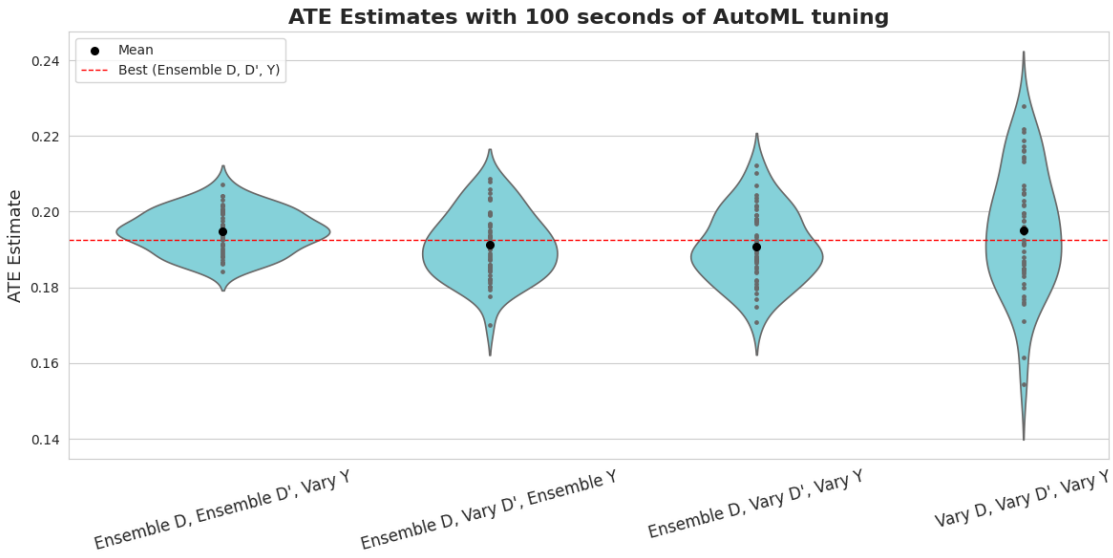


Figure 5: Model Uncertainty: ATE Estimates

Notes: This figure illustrates model uncertainty in ATE estimates arising from variability in nuisance function estimation. The fully ensembled learner is shown by the red dotted line. Starting from the whole sample, AutoML is run a total of 150 times: 50 trials for the initial propensity model $\hat{e}(X)$, and 50 trials each for the re-estimated propensity model $\hat{e}'(X)$ and the outcome models $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$ after trimming, with different random seeds affecting cross-validation splits during hyperparameter search. Fixed models are then combined according to four stabilization strategies: (i) ensemble $\hat{e}(X)$ and $\hat{e}'(X)$, while varying $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$; (ii) ensemble $\hat{e}(X)$, while varying both $\hat{e}'(X)$, $\hat{\mu}_0(X)$, and $\hat{\mu}_1(X)$; (iii) ensemble $\hat{e}(X)$, $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$, while varying $\hat{e}'(X)$; and (iv) vary all three nuisances without ensembling any. Results show that ensembling even a subset of nuisance models reduces dispersion in ATE estimates, highlighting the benefits of such a procedure to stabilize flexible nuisance learners.

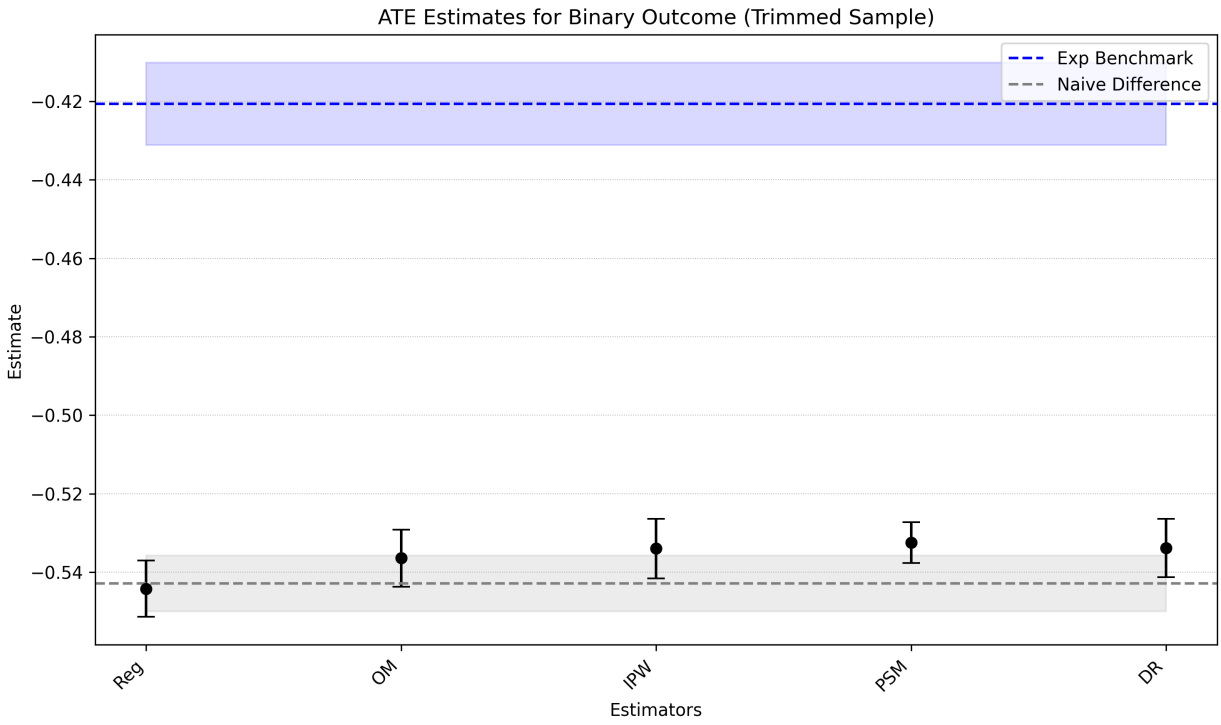
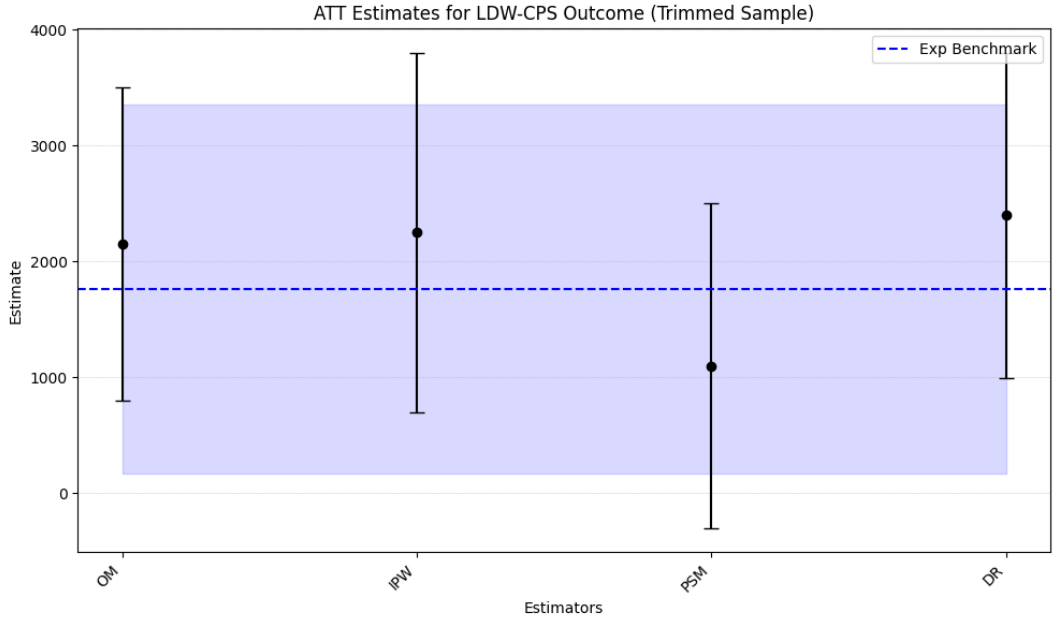
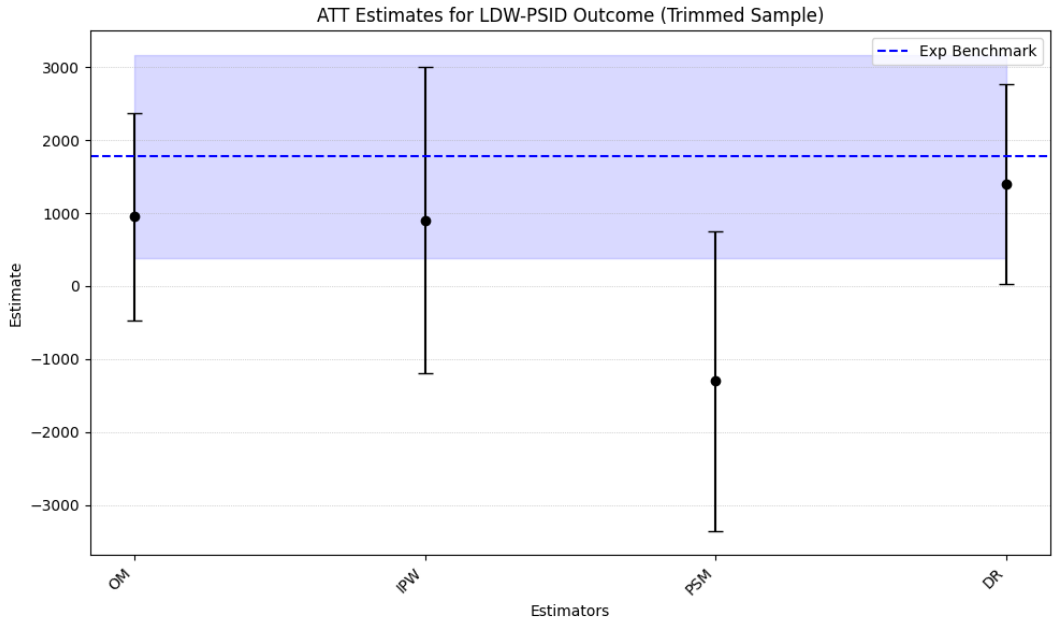


Figure 6: **ATE Estimates for Binary Outcome:** ATE estimates and their 95% confidence intervals for five estimators: regression adjustment (Reg), outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust estimation (DR) in the trimmed product release sample. The blue dashed line and shaded band denote the experimental benchmark and its 95% CI, respectively.

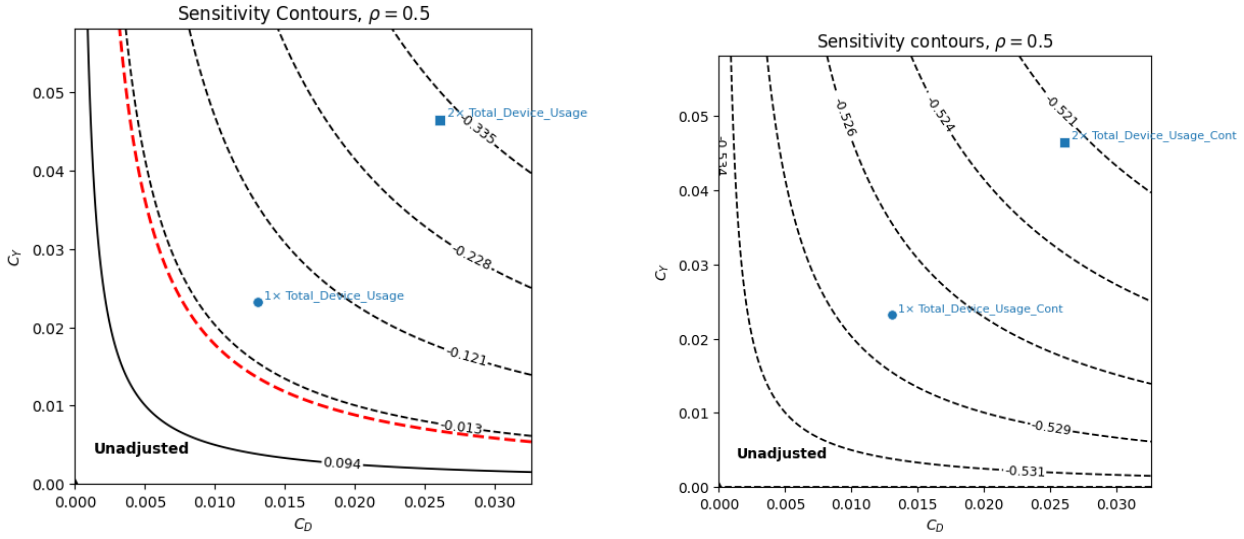


(a) LDW-CPS Sample



(b) LDW-PSID Sample

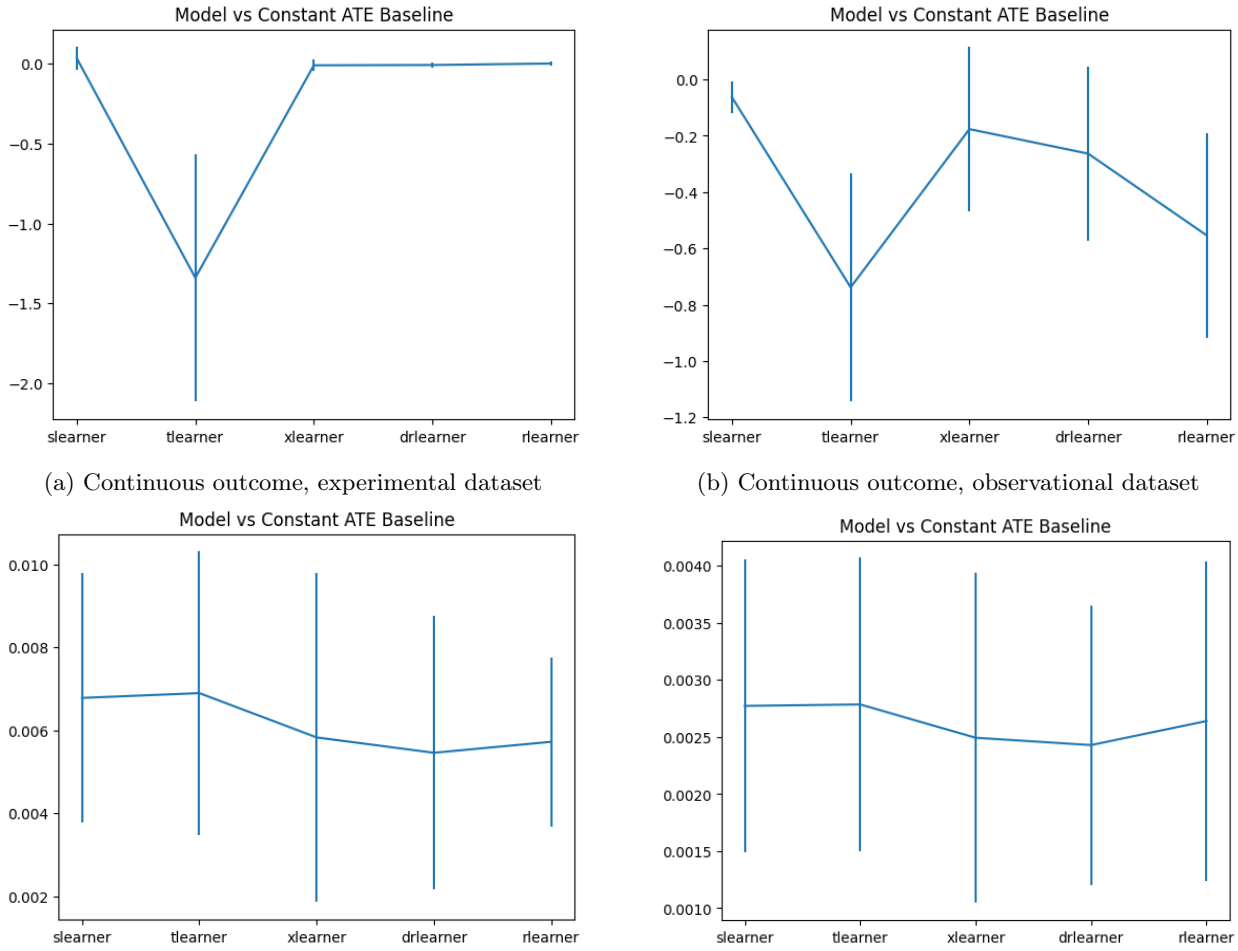
Figure 7: ATT Estimates: Each point represents an estimator's ATT estimate with 95% confidence intervals. The blue dashed line and shaded band denote the experimental benchmark and its 95% CI, respectively.



(a) Continuous outcome

(b) Binary outcome

Figure 8: **Sensitivity Analysis Results:** Panel (a) displays bias contours for the continuous outcome, assuming correlation $\rho = 0.5$ between unobserved confounding effects on treatment and outcome. Moving north-east, each dashed line indicates greater estimated bias relative to the original point estimate. The "unadjusted" is our estimate without unobserved confounding. The dashed red line represents a zero effect. The dots plot the treatment effect we would estimate if we added confounders with 1x and 2x the (C_Y, C_D) as $\mathbb{I}\{\text{Total_Device_Usage}\}$ at $\rho = 0.5$, which is a conservative estimate given what we observe in our data, shown in our benchmarking exercise in Appendix Table G.1. Panel (b) shows the same hypothetical confounder but in the binary outcome setting.



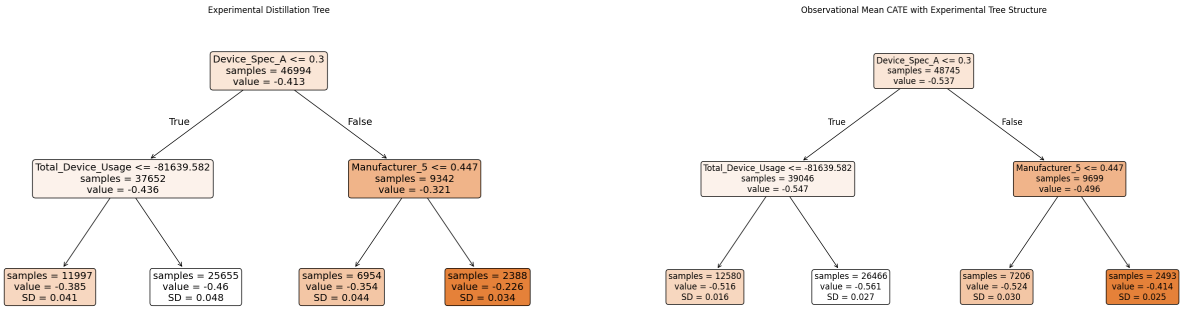
(a) Continuous outcome, experimental dataset

(b) Continuous outcome, observational dataset

(c) Binary outcome, experimental dataset

(d) Binary outcome, observational dataset

Figure 9: DR scoring for presence of heterogeneous effects in the product release data sample



(a) Experimental Sample

(b) Observational Sample

Figure 10: Distilled CATE trees from Q-aggregated models for the binary outcome. In the experimental sample (left), the same variable splits are detected to minimize overlap across leaves. In the observational sample (right), we estimate CATE within the splits detected in the experimental sample. Each final leaf shows the standard deviation of estimated CATE within the group (not the standard error of the estimate).

Table 1: Data Sample Features

Variable	Type	Description
Treatment	binary	1 := new feature enabled
Outcome 1	binary	metric of device performance
Outcome 2	continuous	metric of device performance
Total_Device_Usage	continuous	Total seconds device was in use over the 2-week period
Browser_x_Usage	continuous	Seconds spent in an internet browser over the 2-week period, by anonymized browser (i.e. Chrome, Firefox)
EngagementSegment	ordinal	Indicator of device's past usage patterns (from heavy use to inactive)
A14Region	categorical	Scrambled indicator of geographic region[15]
AppCategoryCohort	categorical	Profile of the device's primary user from past usage patterns across apps (i.e. gamer, worker)
Manufacturer	categorical	Anonymized indicator of device manufacturer
Device_Spec_A	binary	
Device_Spec_B	categorical	Other device characteristics, such as laptop vs desktop
Device_Spec_C	categorical	

Table 2: Optimal Sample Trimming

	Untrimmed	Trimmed	% dropped
Number of Observations	445,286	304,651	31.6%
Number of Treated Units	20,189	18,585	7.9%
Number of Control Units	425,097	286,066	32.7%

Notes: This table summarizes the results of implementing the Crump et al. (2009) trimming procedure on the observational product release sample.

Table 3: Summary of First-Stage Model Performance on Whole and Trimmed Samples

	Whole Sample			Trimmed Sample		
	$\hat{\mu}_0$ R^2	$\hat{\mu}_1$ R^2	$\hat{e}$ AUC	$\hat{\mu}_0$ R^2	$\hat{\mu}_1$ R^2	$\hat{e}$ AUC
Tuned LGBM	0.150	0.127	0.688	0.130	0.126	0.597
Tuned RF	0.138	0.124	0.687	0.123	0.121	0.593
Default LGBM	0.150	0.114	0.686	0.130	0.115	0.592
Logistic/Linear	0.112	0.109	0.675	0.095	0.107	0.578
Tuned NN	0.105	0.117	0.660	0.078	0.076	0.567
Tuned Lasso	0.112	0.109	0.657	0.095	0.107	0.563
Default Lasso	0.067	0.064	0.657	0.050	0.061	0.563
Default RF	0.086	0.069	0.626	0.064	0.061	0.537
Weakly-tuned NN	0.022	0.029	0.510	0.031	0.039	0.498

Notes: This table reports performance metrics for first-stage nuisance models used in the causal analysis. The first two columns report the R^2 scores for the outcome models $\hat{\mu}_0(X_i)$ and $\hat{\mu}_1(X_i)$, evaluated on $\{Y_i : D_i = 0\}$ and $\{Y_i : D_i = 1\}$ respectively. That is, we evaluate our two outcome models' performance based on the subset of observable outcomes on which they were cross-fit and trained. The third column reports the Area Under the Curve (AUC) scores for the propensity score model $\hat{e}(X_i)$. Results are shown separately for the whole sample and the sample after propensity score trimming. Higher scores are indicative of better first-stage nuisance estimation.

Appendix

A Sample Details

We show normalized differences between observational and experimental samples in Figure A.1. We also shown within-sample normalized differences between treated and untreated groups in Figure A.2.

Table A.1: Standardized Mean Differences Between Observed and Experimental Samples

Variable	Observed Mean	Experiment Mean	SMD
Treatment	0.045 (0.208)	0.026 (0.160)	**0.102**
Binary Outcome	0.962 (0.192)	0.976 (0.154)	-0.080
Continuous Outcome	4.116 (6.878)	4.017 (6.666)	0.015
Total_Device_Usage	138,974 (139,359)	136,247 (138,453)	0.020
Browser_1_Usage	22,238 (57,848)	21,805 (57,119)	0.008
Browser_2_Usage	41,787 (79,143)	40,944 (78,301)	0.011
Browser_3_Usage	4,661 (28,999)	4,639 (29,169)	0.001
Browser_4_Usage	38.388 (1,461)	31.971 (1,159)	0.005
Other_Browser_Usage	3,693 (26,972)	3,549 (26,444)	0.005
EngagementSegment_General	0.112 (0.316)	0.112 (0.316)	-0.001
EngagementSegment_Highly Engaged	0.460 (0.498)	0.450 (0.497)	0.021
EngagementSegment_Inactive	0.003 (0.057)	0.003 (0.058)	-0.001
EngagementSegment_Low Engaged	0.145 (0.352)	0.151 (0.358)	-0.017
A14Region_Korea	0.041 (0.199)	0.041 (0.199)	-0.000
A14Region_United States	0.026 (0.158)	0.026 (0.160)	-0.003
A14Region_Japan	0.017 (0.129)	0.017 (0.129)	-0.000
A14Region_India	0.030 (0.171)	0.030 (0.169)	0.003
A14Region_Western Europe	0.029 (0.169)	0.030 (0.171)	-0.004
A14Region_UK	0.124 (0.329)	0.123 (0.329)	0.001
A14Region_Greater China	0.048 (0.214)	0.047 (0.212)	0.004
A14Region_Canada	0.175 (0.380)	0.175 (0.380)	-0.001
A14Region_Latam	0.015 (0.123)	0.015 (0.123)	-0.000
A14Region_France	0.073 (0.260)	0.072 (0.259)	0.003
A14Region_Germany	0.033 (0.178)	0.032 (0.176)	0.005
A14Region_APAC	0.024 (0.152)	0.024 (0.153)	-0.002
A14Region_CEE	0.274 (0.446)	0.275 (0.446)	-0.002
A14Region_MEA	0.068 (0.251)	0.069 (0.253)	-0.003
AppCategoryCohort_Developer	0.036 (0.187)	0.035 (0.184)	0.007
AppCategoryCohort_Gamer	0.105 (0.307)	0.104 (0.305)	0.004
AppCategoryCohort_General	0.112 (0.316)	0.112 (0.316)	-0.001
AppCategoryCohort_Media	0.026 (0.159)	0.026 (0.160)	-0.002
AppCategoryCohort_Productivity	0.170 (0.376)	0.172 (0.377)	-0.003
Manufacturer_1	0.116 (0.320)	0.115 (0.319)	0.003
Manufacturer_2	0.137 (0.343)	0.136 (0.342)	0.003
Manufacturer_3	0.212 (0.409)	0.210 (0.408)	0.004
Manufacturer_4	0.166 (0.372)	0.166 (0.372)	-0.001
Manufacturer_5	0.074 (0.262)	0.076 (0.264)	-0.005
Manufacturer_6	0.219 (0.413)	0.221 (0.415)	-0.005
Device_Spec_A	0.244 (0.429)	0.246 (0.431)	-0.006
Device_Spec_B_1	0.627 (0.484)	0.625 (0.484)	0.004
Device_Spec_B_2	0.186 (0.389)	0.187 (0.390)	-0.005
Device_Spec_C_1	0.040 (0.197)	0.042 (0.200)	-0.007
Device_Spec_C_2	0.518 (0.500)	0.511 (0.500)	0.015
Device_Spec_C_3	0.007 (0.084)	0.007 (0.085)	-0.001

Table A.2: Within-Sample Standardized Mean Differences Between Treated and Untreated Groups

	Observed Sample ($n = 445286$)			Experiment Sample ($n = 435170$)		
Variable	Treated Mean	Untreated Mean	SMD	Treated Mean	Untreated Mean	SMD
Binary Outcome	0.454 (0.498)	0.986 (0.119)	**-1.470**	0.601 (0.490)	0.986 (0.118)	**-1.080**
Continuous Outcome	5.269 (7.983)	4.061 (6.816)	**0.163**	4.438 (6.912)	4.005 (6.659)	0.064
Total_Device_Usage	194,293 (142,868)	136,347 (138,642)	**0.412**	146,718 (139,427)	135,964 (138,416)	0.077
Browser_1_Usage	31,786 (70,701)	21,784 (57,126)	**0.156**	23,077 (59,126)	21,770 (57,064)	0.022
Browser_2_Usage	59,955 (94,314)	40,924 (78,245)	**0.220**	44,301 (80,972)	40,853 (78,225)	0.043
Browser_3_Usage	6,426 (34,542)	4,577 (28,707)	0.058	5,528 (33,460)	4,615 (29,043)	0.029
Browser_4_Usage	60.601 (1,640)	37.333 (1,452)	0.015	39.280 (1,144)	31.773 (1,159)	0.007
Other_Browser_Usage	5,996 (34,885)	3,583 (26,532)	0.078	3,272 (22,746)	3,557 (26,537)	-0.012
EngagementSegment_General	0.106 (0.308)	0.112 (0.316)	-0.021	0.110 (0.313)	0.112 (0.316)	-0.008
EngagementSegment_Highly Engaged	0.694 (0.461)	0.449 (0.497)	**0.511**	0.495 (0.500)	0.449 (0.497)	0.093
EngagementSegment_Inactive	0.001 (0.025)	0.003 (0.059)	-0.062	0.002 (0.042)	0.003 (0.059)	-0.033
EngagementSegment_Low Engaged	0.019 (0.136)	0.151 (0.358)	**-0.487**	0.115 (0.319)	0.152 (0.359)	**-0.110**
A14Region_Korea	0.044 (0.204)	0.041 (0.199)	0.012	0.046 (0.209)	0.041 (0.199)	0.022
A14Region_United States	0.028 (0.164)	0.026 (0.158)	0.013	0.026 (0.158)	0.026 (0.160)	-0.004
A14Region_Japan	0.017 (0.130)	0.017 (0.129)	0.003	0.016 (0.125)	0.017 (0.129)	-0.009
A14Region_India	0.032 (0.177)	0.030 (0.170)	0.015	0.031 (0.175)	0.029 (0.169)	0.012
A14Region_Western Europe	0.026 (0.160)	0.030 (0.169)	-0.020	0.029 (0.169)	0.030 (0.171)	-0.004
A14Region_UK	0.143 (0.351)	0.123 (0.328)	0.061	0.143 (0.350)	0.123 (0.328)	0.058
A14Region_Greater China	0.051 (0.221)	0.048 (0.213)	0.017	0.054 (0.225)	0.047 (0.211)	0.031
A14Region_Canada	0.166 (0.372)	0.175 (0.380)	-0.024	0.166 (0.372)	0.175 (0.380)	-0.023
A14Region_Latam	0.012 (0.111)	0.015 (0.123)	-0.026	0.014 (0.118)	0.015 (0.123)	-0.011
A14Region_France	0.083 (0.276)	0.072 (0.259)	0.039	0.073 (0.260)	0.072 (0.259)	0.004
A14Region_Germany	0.036 (0.185)	0.033 (0.178)	0.015	0.031 (0.173)	0.032 (0.176)	-0.006
A14Region_APAC	0.023 (0.149)	0.024 (0.152)	-0.007	0.024 (0.155)	0.024 (0.153)	0.003
A14Region_CEE	0.247 (0.431)	0.275 (0.447)	-0.063	0.257 (0.437)	0.275 (0.447)	-0.042
A14Region_MEA	0.068 (0.252)	0.068 (0.251)	0.002	0.069 (0.253)	0.069 (0.253)	0.000
AppCategoryCohort_Developer	0.049 (0.216)	0.036 (0.186)	0.066	0.041 (0.199)	0.035 (0.183)	0.035
AppCategoryCohort_Gamer	0.129 (0.335)	0.104 (0.306)	0.077	0.113 (0.317)	0.104 (0.305)	0.030
AppCategoryCohort_General	0.106 (0.308)	0.112 (0.316)	-0.021	0.110 (0.313)	0.112 (0.316)	-0.008
AppCategoryCohort_Media	0.019 (0.136)	0.026 (0.160)	-0.050	0.026 (0.159)	0.026 (0.160)	-0.003
AppCategoryCohort_Productivity	0.137 (0.344)	0.172 (0.377)	-0.097	0.160 (0.366)	0.172 (0.377)	-0.033
Manufacturer_1	0.123 (0.329)	0.115 (0.320)	0.024	0.116 (0.321)	0.115 (0.319)	0.005
Manufacturer_2	0.141 (0.348)	0.136 (0.343)	0.013	0.138 (0.345)	0.136 (0.342)	0.007
Manufacturer_3	0.218 (0.413)	0.212 (0.409)	0.016	0.203 (0.402)	0.211 (0.408)	-0.018
Manufacturer_4	0.173 (0.378)	0.166 (0.372)	0.020	0.162 (0.368)	0.166 (0.373)	-0.013
Manufacturer_5	0.045 (0.207)	0.076 (0.264)	**-0.129**	0.069 (0.254)	0.076 (0.265)	-0.025
Manufacturer_6	0.214 (0.410)	0.219 (0.414)	-0.012	0.230 (0.421)	0.221 (0.415)	0.022
Device_Spec_A	0.183 (0.387)	0.247 (0.431)	**-0.156**	0.233 (0.422)	0.247 (0.431)	-0.033
Device_Spec_B_1	0.656 (0.475)	0.625 (0.484)	0.064	0.641 (0.480)	0.625 (0.484)	0.035
Device_Spec_B_2	0.135 (0.342)	0.188 (0.391)	**-0.144**	0.179 (0.383)	0.188 (0.390)	-0.023
Device_Spec_C_1	0.028 (0.164)	0.041 (0.198)	-0.073	0.043 (0.203)	0.042 (0.200)	0.007
Device_Spec_C_2	0.671 (0.470)	0.511 (0.500)	**0.331**	0.537 (0.499)	0.510 (0.500)	0.055
Device_Spec_C_3	0.004 (0.066)	0.007 (0.085)	-0.037	0.008 (0.089)	0.007 (0.084)	0.009

B Propensity Estimation and Tuning

Here we provide the propensity score distributions across a range of different flexible and inflexible models. We compare the "untuned" or default sklearn implementations, as well as an AutoML-tuned version.[16] For the neural network, we distinguish between a *fully-tuned* neural network and a *weakly-tuned* baseline. Both are optimized with `Optuna` (another Bayesian optimization AutoML framework), but the search space for the fully-tuned network is much deeper and more expressive. Details are given in Appendix Table B.1.

Table B.1: Hyper-parameter search spaces for the fully-tuned and the weakly-tuned neural networks.

Model	**Fully-tuned**	**Weakly-tuned**
Hidden layers n_{layers}	$\{2, \ldots, 7\}$	$\{1, 2, 3\}$
Units per layer $n_{\text{units}}^{(l)}$	$\{32, 64, \ldots, 1024\}$	$\{16, 32, \ldots, 128\}$
Activation functions	`relu`, `tanh`, `gelu`, `selu`, `leaky_relu`, `elu`	`relu`, `tanh`, `gelu`
Dropout rate	$[0, 0.5]$	$[0, 0.3]$
Batch normalization	`True` / `False`	`False` (Fixed)
Skip connections	`True` / `False`	`False` (Fixed)
Learning rate	$[10^{-4}, 5 \times 10^{-2}]$	$[10^{-4}, 5 \times 10^{-2}]$
Weight decay (ℓ_2)	$[0, 5 \times 10^{-2}]$	0 (Fixed)

[16] Details on default parameters can be found in the API pages for linear models, random forests, and boosted trees.

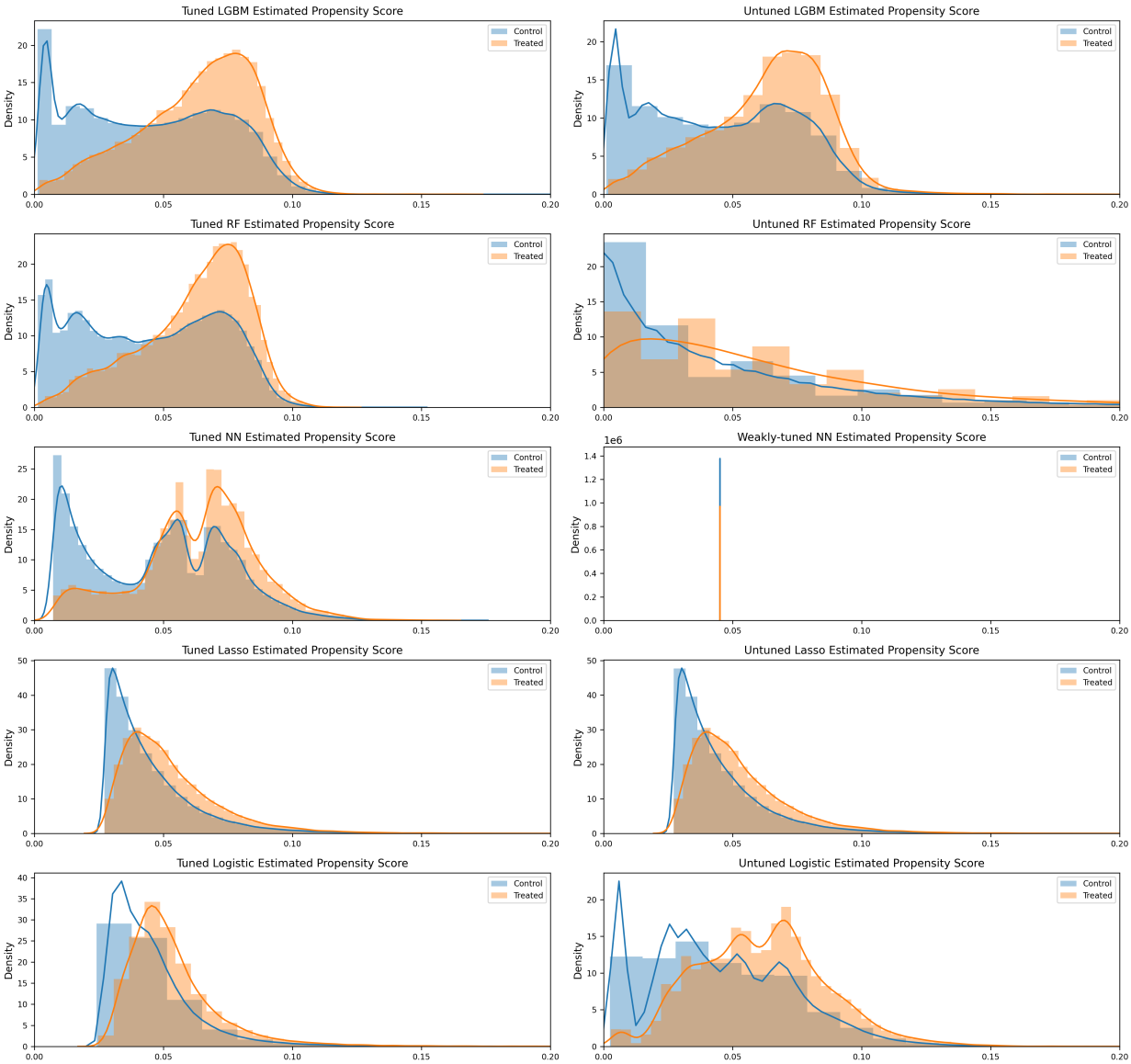


Figure B.1: Untrimmed Sample Propensity Scores: Model Comparison with and without Tuning

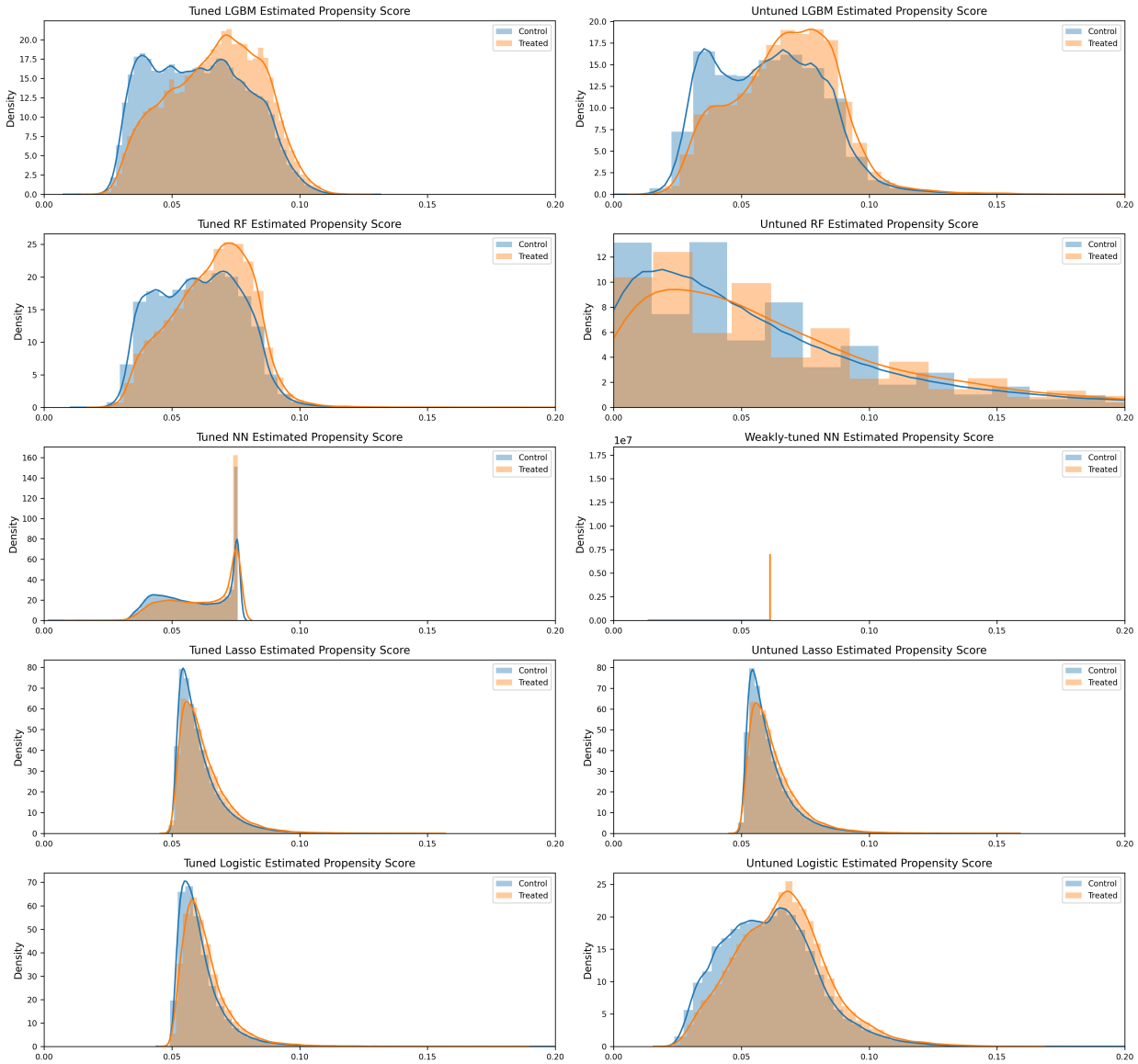
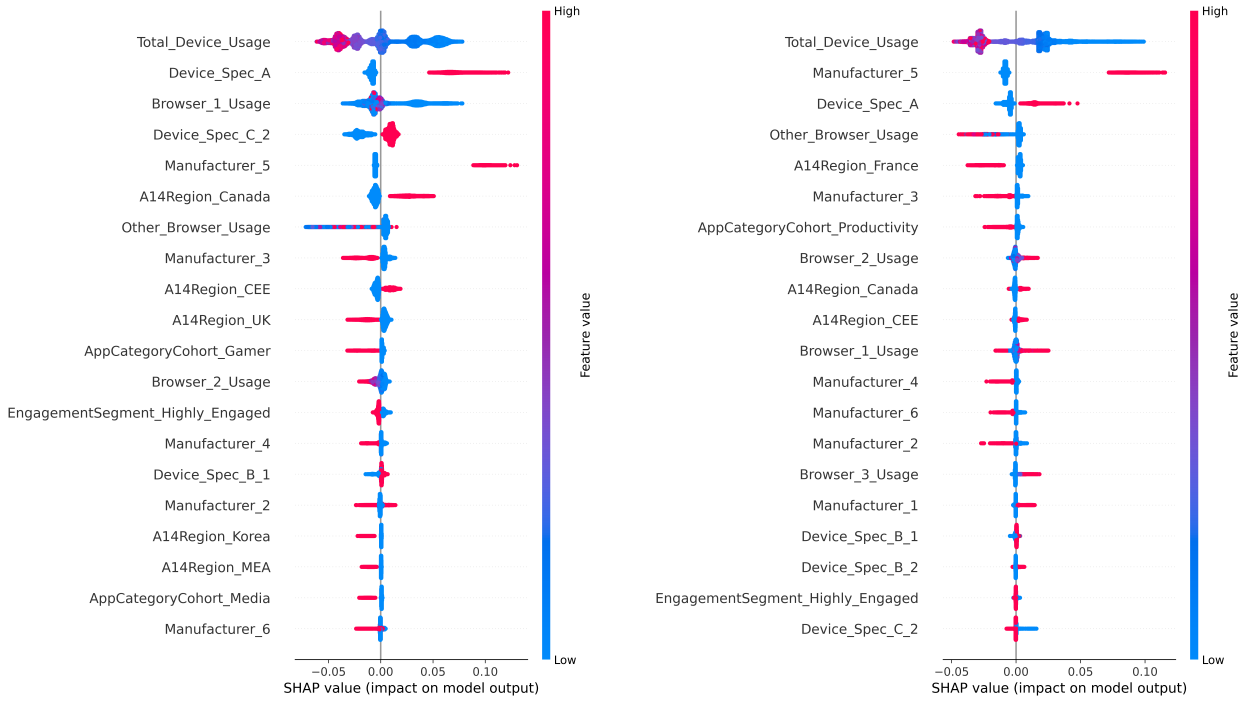


Figure B.2: Trimmed Sample Propensity Scores: Model Comparison with and without Tuning



(a) Experimental Sample: SHAP Summary

(b) Observational Sample: SHAP Summary

Figure B.3: SHAP summary plots of feature importance in Q-aggregated CATE models for the binary outcome. Consistent signals are observed in both datasets.

C Additional ATE Results on the Untrimmed Samples

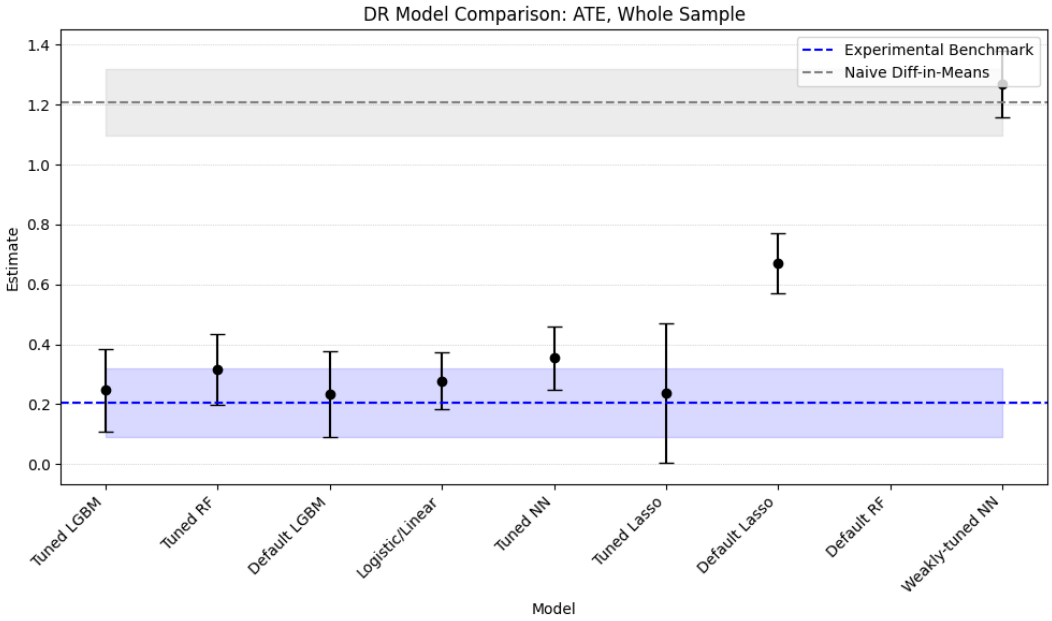
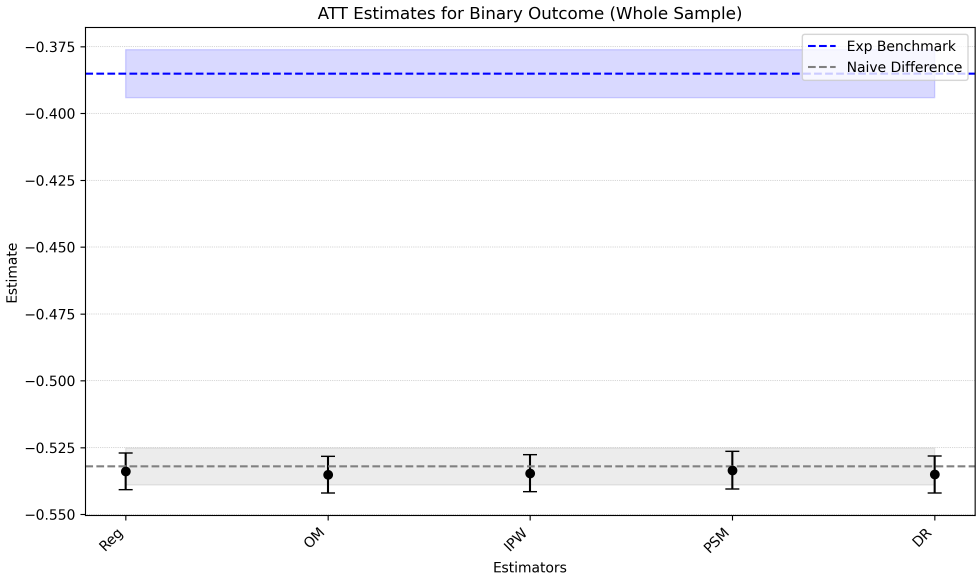


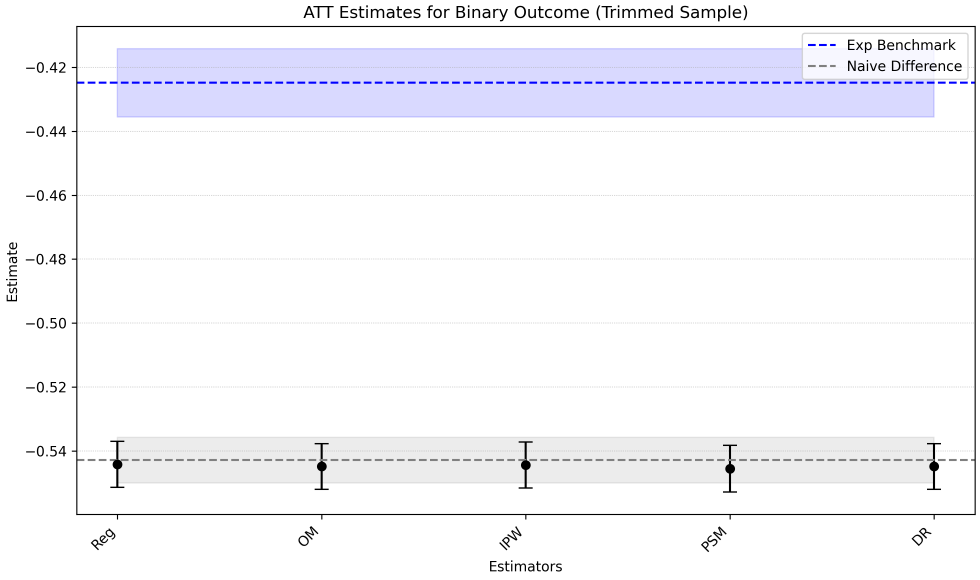
Figure C.1: **DR Estimates Across Different First-Stage Models (Untrimmed Sample):** This figure compares ATE estimates from doubly robust estimators using a range of flexible and rigid first-stage models for outcome and treatment modeling. Both tuned and default sklearn versions are considered. Poorly tuned flexible models, particularly RFs and NNs, exhibit severe bias and instability, highlighting the critical role of hyperparameter tuning in these cases. The default random forest implementation yields an estimate of -93.776 (67.239), driven by very unstable propensity weights, and is thus not pictured in the figure. Trimming is not applied here, amplifying model misspecification and overlap issues.

D ATT Results

We show average treatment effect on the treated results for the binary outcome in Figure D.1, and those for the continuous outcome in Figure D.2.

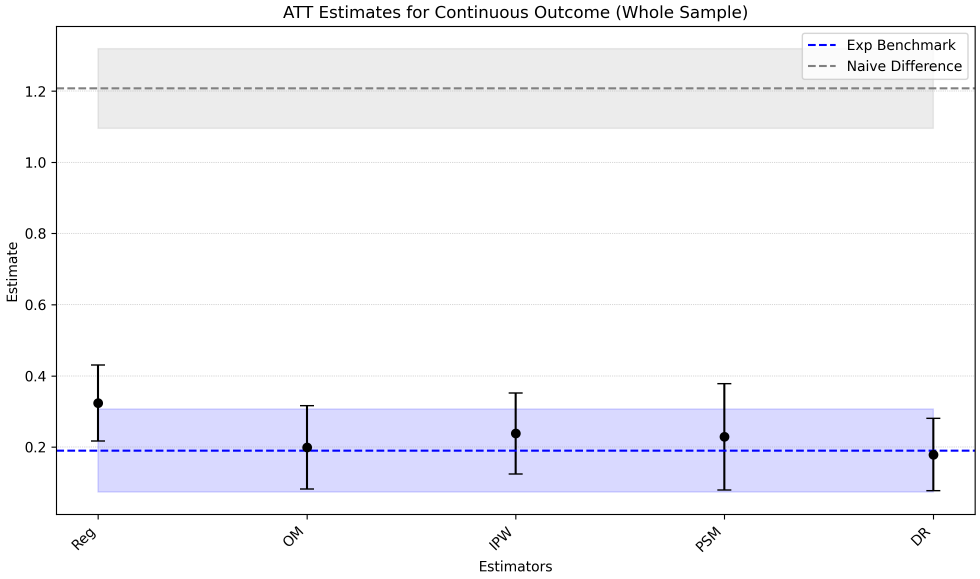


(a) Untrimmed Sample

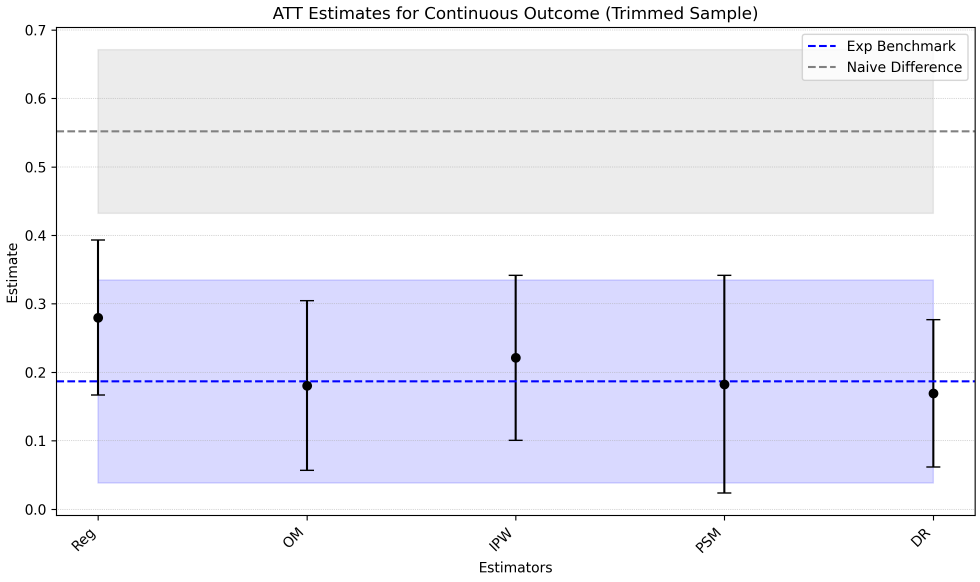


(b) Trimmed Sample

Figure D.1: **ATT Estimates for Binary Outcome:** Panels (a) and (b) display ATT estimates and their 95% confidence intervals for five estimators: regression adjustment (Reg), outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust estimation (DR). Panel (a) shows results from the untrimmed sample, and panel (b) presents estimates after overlap trimming. Across both samples, observational estimators fail to recover the experimental benchmark (blue dashed line), with estimates consistently biased downward and concentrated around the naive difference in means (gray dashed line). This pattern strongly suggests unobserved confounding. Unlike the continuous outcome, where modern econometric advancements help align with the experimental benchmark, these results show that even with doubly robust methods and trimming, we fail to recover the ground truth.



(a) Untrimmed Sample



(b) Trimmed Sample

Figure D.2: **ATT Estimates for Continuous Outcome:** Panels (a) and (b) display ATT estimates and their 95% confidence intervals for five different estimators: regression adjustment (Reg), outcome modeling (OM), inverse probability weighting (IPW), propensity score matching (PSM), and doubly robust estimation (DR). Panel (a) shows the full untrimmed sample; Panel (b) displays estimates from the trimmed sample. In each plot, the black dashed line and shaded band indicate the naive difference in means and its 95% confidence interval, while the blue dashed line and shaded region denote the experimental benchmark and its 95% CI. Trimming improves estimator agreement with the benchmark and reduces variance. These results are consistent with findings in Dehejia and Wahba (1999), Crump et al. (2009), and Imbens and Xu (2025).

E Additional CATE Results

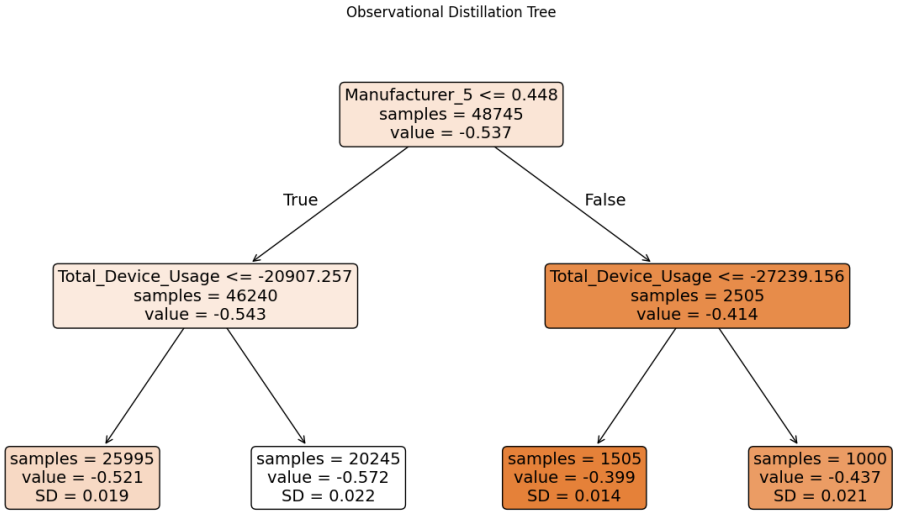


Figure E.1: Distilled CATE tree from a Q-aggregated model for the binary outcome, estimated using only the observational sample. This tree recovers many of the same splits as the experimental tree like Total_Device_Usage and Manufacturer_5. Each final leaf shows the standard deviation of estimated CATE within the group (not the standard error of the estimate).

F Experimental Adjustment

We apply a doubly-robust adjustment on the experiment to get the benchmark.[17] The summary of this is given in Appendix Table F.1.

Table F.1: Treatment Effect Estimates Under Trimming and Estimation Methods

Sample	**Estimator**	**Full Sample**	**Trimmed**	**Trimmed Off-Support**
Experimental	Difference-in-Means	0.433 (0.065)	0.309 (0.082)	0.236 (0.086)
	DR (ATT)	0.190 (0.059)	0.187 (0.076)	0.184 (0.081)
	DR (ATE)	0.205 (0.059)	0.210 (0.077)	0.165 (0.083)
Observational	Difference-in-Means	1.208 (0.057)	0.552 (0.061)	0.605 (0.149)
	DR (ATT)	0.179 (0.052)	0.169 (0.055)	0.317 (0.144)
	DR (ATE)	0.247 (0.061)	0.206 (0.056)	0.414 (0.172)

Notes: Each entry reports the estimated treatment effect with its standard error in parentheses. "Trimmed Off-Support" denotes the excluded region from Crump et al. (2009) trimming.

[17] We note a simple regression adjustment leads to an almost indistinguishable benchmark.

G More on Sensitivity Analysis

G.1 Benchmarking

We report our benchmarking (C_Y, C_D, ρ) for each observed covariate in Appendix Table G.1

(a) Continuous outcome

	C_Y	C_D	ρ
Total_Device_Usage	0.0232	0.013	0.1038
Other_Browser_Usage	0.0111	0.001	0.0675
Manufacturer_2	0.0059	0.0009	0.0519
Browser_2_Usage	0.0053	0	−1
A14Region_UK	0.0033	0.0008	0.0761
A14Region_Latam	0.0029	0.0012	−0.1385
Browser_1_Usage	0.0026	0	−1
AppCategoryCohort_Gamer	0.0015	0	1
Manufacturer_4	0.0015	0.0011	0.0027
Device_Spec_C_2	0.0013	0.0033	0.5336
AppCategoryCohort_Developer	0.0013	0	0.176
Manufacturer_3	0.0012	0.0008	−0.1088
Device_Spec_B_1	0.0008	0.0003	−0.1928
A14Region_Canada	0.0008	0.0011	−0.0921
Manufacturer_6	0.0007	0	−1
Device_Spec_C_1	0.0005	0.0008	−0.0018
Manufacturer_1	0.0004	0	−1
Browser_3_Usage	0.0004	0.0001	0.0821
EngagementSegment_Highly Engaged	0.0004	0.0151	0.2069
Browser_4_Usage	0.0004	0.0006	−0.0241
Device_Spec_A	0.0004	0.0013	0.2863
Manufacturer_5	0.0003	0	−1
A14Region_Greater China	0.0002	0.0006	−0.0563
Device_Spec_C_3	0.0002	0.0012	−0.081
AppCategoryCohort_Media	0.0002	0	1
Device_Spec_B_2	0.0001	0.0005	−0.4256
A14Region_Korea	0.0001	0.0011	−0.3347
EngagementSegment_General	0.0001	0.0014	−0.1758
A14Region_Germany	0.0001	0.0011	−0.0069
A14Region_India	0	0.001	−0.2162
A14Region_MEA	0	0.001	−0.197
A14Region_CEE	0	0.0001	0.3466
EngagementSegment_Inactive	0	0.0009	−0.211
A14Region_APAC	0	0.0014	−1
AppCategoryCohort_Productivity	0	0	−1
A14Region_France	0	0.0009	−1
AppCategoryCohort_General	0	0.0009	−1
EngagementSegment_Low Engaged	0	0.0008	−1
A14Region_Japan	0	0.0008	−1
A14Region_United States	0	0.0008	−1
A14Region_Western Europe	0	0.0011	−1

(b) Binary outcome

	C_Y	C_D	ρ
Manufacturer_5	0.0032	0	−1
Total_Device_Usage	0.0016	0.013	−0.8427
Manufacturer_1	0.001	0	1
Device_Spec_A	0.0006	0.0013	−0.0194
A14Region_Canada	0.0006	0.0011	−0.2316
A14Region_CEE	0.0004	0.0001	−1
Device_Spec_B_1	0.0003	0.0003	−0.0417
A14Region_UK	0.0002	0.0008	−0.6281
AppCategoryCohort_Productivity	0.0002	0	1
Device_Spec_C_2	0.0002	0.0033	−0.5695
Manufacturer_4	0.0002	0.0011	−0.1681
Device_Spec_C_3	0.0002	0.0012	−0.2996
A14Region_Japan	0.0002	0.0008	−0.3434
Device_Spec_C_1	0.0002	0.0008	−0.5457
Manufacturer_3	0.0002	0.0008	−0.132
A14Region_Korea	0.0002	0.0011	−0.1916
Manufacturer_6	0.0001	0	−1
Other_Browser_Usage	0.0001	0.001	−1
Browser_2_Usage	0.0001	0	1
A14Region_Greater China	0.0001	0.0006	−0.4853
EngagementSegment_General	0.0001	0.0014	−0.4892
Device_Spec_B_2	0.0001	0.0005	−0.4331
A14Region_Western Europe	0.0001	0.0011	−0.2027
A14Region_Latam	0.0001	0.0012	−0.5014
A14Region_India	0.0001	0.001	−0.544
A14Region_United States	0.0001	0.0008	−0.5213
AppCategoryCohort_General	0.0001	0.0009	−0.2546
AppCategoryCohort_Media	0.0001	0	−1
AppCategoryCohort_Gamer	0.0001	0	−1
EngagementSegment_Inactive	0.0001	0.0009	−0.4463
A14Region_France	0.0001	0.0009	−1
EngagementSegment_Low Engaged	0.0001	0.0008	−0.4973
A14Region_APAC	0.0001	0.0014	−0.3106
A14Region_MEA	0.0001	0.001	−0.5491
A14Region_Germany	0.0001	0.0011	−0.318
EngagementSegment_Highly Engaged	0.0001	0.0151	−0.6871
AppCategoryCohort_Developer	0.0001	0	−1
Browser_1_Usage	0.0001	0	−1
Browser_4_Usage	0	0.0006	−0.8065
Browser_3_Usage	0	0.0001	−1
Manufacturer_2	0	0.0009	−1

Table G.1: Covariate benchmarking for sensitivity analysis for both outcomes, sorted by C_Y. Each panel reports (C_Y, C_D, ρ) for a single omitted observed covariate, constructed by re-estimating the model without that covariate. These benchmarks show that the *observed* covariates rarely achieve both large C_Y and large C_D, and for those that do, their empirical ρ's are far from ± 0.5.

G.2 Implementation of Sensitivity Analysis

Let X be observed covariates and U a latent confounder, Following Chernozhukov et al. (2022), we define the *long* regression that conditions on $W := (X, U)$ and the *short* regression that conditions on X only. Denote by $\alpha(\cdot)$ the Riesz representer of the doubly robust ATE functional, $\alpha(X) = \frac{D}{e(X)} - \frac{1-D}{1-e(X)}$. For convenience, we also define the long and short outcome regressions:

$$\mu_L(X) = \mu_{1,L}(X)D + \mu_{0,L}(X)(1-D), \qquad \mu_S(X) = \mu_{1,S}(X)D + \mu_{0,S}(X)(1-D)$$

Likewise, $\alpha_L(X)$ and $\alpha_S(X)$ are built from $e_L(X)$ and $e_S(X)$. Our bias bound is

$$|\theta_S - \theta| \;\le\; |\rho|\, C_Y\, C_D\, S, \qquad S^2 := \mathbb{E}\big[(Y-\mu_S)^2\big]\;\mathbb{E}\big[\alpha_S^2\big],$$

with

$$C_Y^2 := \frac{\mathbb{E}\big[(\mu_L - \mu_S)^2\big]}{\mathbb{E}\big[(Y-\mu_S)^2\big]}, \qquad C_D^2 := \frac{\mathbb{E}[\alpha_L^2] - \mathbb{E}[\alpha_S^2]}{\mathbb{E}[\alpha_S^2]}, \qquad \rho := \mathrm{Corr}(\mu_L - \mu_S,\ \alpha_L - \alpha_S).$$

For empirical implementation, we estimate the long and short models with our cross-fit estimates, and use them as plug-in estimators for their population quantities:

$$\hat\sigma_S^2 = \frac{1}{n}\sum_i \big(Y_i - \hat\mu_S(W_i)\big)^2, \qquad \hat\nu_S^2 = \frac{1}{n}\sum_i \hat\alpha_S(W_i)^2, \quad \text{with} \quad \hat\alpha_S(W) = \frac{D}{\hat e_S(X)} - \frac{1-D}{1-\hat e_S(X)}.$$

Analogously obtain $\hat\sigma_L^2, \hat\nu_L^2$ from the long models. We then plug-in these quantities to estimate our benchmark strengths:[18]

$$\widehat{C_Y^2} = \frac{\hat\sigma_S^2 - \hat\sigma_L^2}{\hat\sigma_S^2}, \qquad \widehat{C_D^2} = \frac{\hat\nu_L^2 - \hat\nu_S^2}{\hat\nu_S^2}.$$

[18] Our implementation follows the `DoubleML` benchmarking routine, which reports

$$C^2_{Y,\mathrm{DoubleML}} = \frac{\hat\sigma_S^2 - \hat\sigma_L^2}{\hat\sigma_L^2}.$$

This differs slightly from the parametrization in Chernozhukov et al. (2022), where

$$C_Y^2 = \frac{\hat\sigma_S^2 - \hat\sigma_L^2}{\hat\sigma_S^2}.$$

The two are related by the monotone transform $C_Y^2 = C^2_{Y,\mathrm{DoubleML}}/(1 + C^2_{Y,\mathrm{DoubleML}})$, so the empirical interpretation of these benchmark strengths and robustness values remains unchanged.

We also compute the observed benchmark set bias and its implied adversity:

$$\widehat{\Delta\theta} = \hat{\theta}_S - \hat{\theta}_L, \qquad \hat{\rho} = \operatorname{sign}(\widehat{\Delta\theta}) \left(\frac{|\widehat{\Delta\theta}|}{\sqrt{(\hat{\sigma}_S^2 - \hat{\sigma}_L^2)\,(\hat{\nu}_L^2 - \hat{\nu}_S^2)}} \right).$$

Given a ρ, the minimal joint strength $C_Y C_D$ required to move the point estimate to zero is summarized by the robustness value:

$$\mathrm{RV}(\rho) = \frac{|\hat{\theta}|}{|\rho|\,\hat{S}}, \qquad \hat{S} = \sqrt{\hat{\sigma}_S^2\,\hat{\nu}_S^2}.$$

At confidence level $1-\alpha$, the CI-robustness value is

$$\mathrm{RV}_\alpha(\rho) = \frac{\max\{\,|\hat{\theta}| - z_{1-\alpha}\,\widehat{\mathrm{se}}(\hat{\theta}),\, 0\}}{|\rho|\,\hat{S}}.$$