Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1276

Comparing Machine Learning Methods for Estimating Heterogeneous Treatment Effects in Randomized Trials: A Comprehensive Simulation Study

Luke Miratrix
Harvard University

Polina Polskaia
MDRC

Richard Dorsett
University of Westminster

Pei Zhu
MDRC

Nicholas Commins
MDRC

J. David Selby
MDRC

This study compares 18 machine learning methods for estimating heterogeneous treatment effects in randomized controlled trials, using simulations calibrated to two large-scale educational experiments. We evaluate performance across continuous and binary outcomes with diverse and realistic treatment effect heterogeneity patterns, varying sample sizes, covariate complexities, and effect magnitudes. Bayesian Additive Regression Trees with S-learner (BART S) outperforms alternatives on average. While no method predicts individual effects with high accuracy, some show promise in identifying who benefits most or least. An empirical application illustrates how ML methods can reveal heterogeneity patterns beyond conventional subgroup analysis. These findings highlight both the potential and the limitations of ML, offering evidence-based practical guidance for analyzing treatment effect variation in experimental evaluations.

VERSION: September 2025

Suggested citation: Miratrix, Luke, Polina Polskaia, Richard Dorsett, Pei Zhu, Nicholas Commins, and J.David Selby. (2025). Comparing Machine Learning Methods for Estimating Heterogeneous Treatment Effects in Randomized Trials: A Comprehensive Simulation Study. (EdWorkingPaper: 25-1276). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/qdkn-z470

Comparing Machine Learning Methods for Estimating Heterogeneous Treatment Effects in Randomized Controlled Trials: A Comprehensive Simulation Study

Luke Miratrix[a]
Polina Polskaia[b]
Richard Dorsett[c]
Pei Zhu[b]
Nicholas Commins[b]
J.David Selby[b]

[a] Harvard Graduate School of Education, [b] MDRC, [c] University of Westminster

Authors' notes: This study is funded by the Institute of Education Sciences research grant R305D220028. Correspondence concerning this article should be addressed to Pei Zhu, MDRC email: Pei.Zhu@mdrc.org

Abstract

This study compares 18 machine learning methods for estimating heterogeneous treatment effects in randomized controlled trials, using simulations calibrated to two large-scale educational experiments. We evaluate performance across continuous and binary outcomes with diverse and realistic treatment effect heterogeneity patterns, varying sample sizes, covariate complexities, and effect magnitudes. Bayesian Additive Regression Trees with S-learner (BART S) outperforms alternatives on average. While no method predicts individual effects with high accuracy, some show promise in identifying who benefits most or least. An empirical application illustrates how ML methods can reveal heterogeneity patterns beyond conventional subgroup analysis. These findings highlight both the potential and the limitations of ML, offering evidence-based practical guidance for analyzing treatment effect variation in experimental evaluations.

Key Words: Machine Learning; Heterogeneous Treatment Effect; Randomized Controlled Trials; Simulation Study; Educational Research

1. Introduction

Understanding how the effects of educational and behavioral interventions vary across individuals is critical for optimizing program design and informing policy decisions, as it enables policymakers to target resources more effectively, improve program effectiveness, and design more equitable interventions. For example, a randomized controlled trial (RCT) of Career Academies revealed that long-term earnings gains were concentrated among men in the sample, suggesting the need for alternative strategies to benefit women (Kemple & Willner, 2008). In contrast, the Accelerated Study in Associate Programs (ASAP) increased degree completion across diverse demographic groups, demonstrating its potential for broad applicability (Miller & Weiss, 2021).

Researchers investigating treatment effect variation typically pursue three distinct but often conflated goals: (a) *individual-level prediction*—estimating how much specific individuals will benefit from treatment based on their characteristics; (b) *subgroup estimation*—determining whether people with certain attributes benefit more or less than others, typically through pre-specified interactions or stratification; and (c) *feature importance analysis*—identifying which characteristics most strongly predict treatment response (Athey & Imbens, 2019; Bloom & Michalopoulos, 2013; Chernozhukov et al., 2018). While these goals are related, they address fundamentally different research questions and can require different methodological tools.

We focus on the first goal: individual-level prediction of conditional average treatment effects (CATEs). While subgroup estimation and feature importance analysis are important in their own right, our simulations are not designed to evaluate these goals directly. Instead, we ask: to what extent can off-the-shelf ML methods predict CATEs in realistic RCT contexts? Subgroup comparisons and feature-based insights appear only in our empirical application as

illustrations of how individual-level predictions may complement conventional subgroup analysis. Individual-level effect estimation serves several important purposes. First, accurate predictions can guide practitioners in targeting interventions to those most likely to benefit, improving cost-effectiveness. Second, they provide a direct measure of how much treatment response varies across individuals, beyond what subgroup averages reveal. Third, they can be flexibly aggregated into group-level effects, allowing subgroup analyses that build on the full distribution of individual-level heterogeneity rather than relying on pre-specified categories.

The use of ML for estimating heterogeneous treatment effects (HTEs) is relatively recent (e.g., Hill, 2011; Athey & Imbens, 2019; Chernozhukov et al., 2018), but significant methodological advances have quickly established a wide range of ML methods for estimating individual effects. The plethora of available methods presents a challenge for researchers: How does a researcher know which tool to use?

This paper provides guidance through a comprehensive simulation study calibrated to data from two large-scale RCTs: Career Academies and ASAP. Building on prior work such as Knaus et al. (2021), we systematically evaluate the performance of 18 ML methods across diverse data-generating processes (DGPs), offering a robust test of ML method performance in predicting individual treatment effects against realistic heterogeneity patterns in RCT settings.

Our study contributes to the literature in three key ways. First, **we focus exclusively on RCTs**, **where confounding is not a concern.** Most existing methods in the literature were designed for observational settings where confounding increases the challenge of understanding impact heterogeneity. However, the relative performance of different methods in observational studies may not carry over to the experimental case, and the question of which methods would be most appropriate for experimental studies is rarely directly addressed.

Second, **we evaluate methods across a wide range of contexts, varying sample size, the number of effect moderators, and the form of treatment effect heterogeneity.** This design extends beyond Knaus et al. (2021)—the largest simulation exercise we found to date—by expanding both the set of methods and the range of simulation scenarios considered.

Third, **we introduce a novel "multiple-queens" framework to generate simulated heterogeneity.** Unlike earlier studies that generated impacts using specified functional forms (e.g., Knaus et al., 2021) or relying on single-model-generated effects (Wendling et al., 2018), we fit multiple ML models ("queens") to real RCT data and use their predictions to create diverse, empirically grounded treatment effect patterns. This approach preserves the structure of the original data while enhancing generalizability, reducing dependence on any one data-generating process, and testing method robustness under varied heterogeneity structures.

We also recognize that researchers often ultimately aim to identify meaningful subgroups with differential treatment responses. We apply the best-performing ML method to real trial data to examine whether the patterns of heterogeneity identified by ML-based estimates are consistent with those revealed by conventional subgroup analysis.

Practically, our study aims to guide researchers in selecting reliable ML methods for HTE analysis in RCTs. We prioritize "off-the-shelf" algorithms, relying on tools with existing implementations either in publicly available R packages or shared codebases. We thus attempt to capture how a typical user might apply these methods in practice, enabling us to offer actionable insights into best practices for modeling HTE in the educational and behavioral sciences.

Among the evaluated methods, Bayesian Additive Regression Trees with the S-learner framework (BART S) outperformed other methods on average across scenarios. In contrast, some other methods produced predictions that were, on average, less accurate than simply

assigning the overall average treatment effect (ATE) to all individuals (similar to results reported in Tipton and Mamakos, 2025). Although none of the methods had strong overall predictive performance, several nonetheless provide useful indication of who will benefit most or least from the treatment. These tools can uncover patterns of treatment effect variation that both mirror and extend those identified by conventional subgroup analyses, offering additional insights into heterogeneity. However, researchers should maintain realistic expectations regarding the precision of individual-level predictions in typical educational and behavioral research contexts.

The remainder of this paper proceeds as follows: Section 2 reviews the ML methods evaluated, Section 3 details our simulation design featuring the multiple-queens framework, Section 4 presents simulation results, Section 5 applies the best-performing method to empirical data, and Section 6 concludes with implications for educational research.

2. Machine Learning Methods Under Evaluation

We use the potential outcomes framework (Neyman, 1923; Rubin, 1974). For a two-armed RCT, each individual i has two potential outcomes: $Y_i(1)$ under treatment ($T_i = 1$) and $Y_i(0)$ under control ($T_i = 0$). For unit i, the individual treatment effect is defined as $\beta_i = Y_i(1) - Y_i(0)$, but the fundamental problem of causal inference (Holland, 1986) prevents direct observation of both outcomes since only one is realized: $Y_i = Y_i(0) + \beta_i T_i$. Randomization ensures that treatment assignment T_i is independent of the potential outcomes, enabling unbiased estimation of the average treatment effect (ATE), $\tau = E(Y|T = 1) - E(Y|T = 0)$, via a simple difference in group means. Random assignment distinguishes RCTs from observational studies, which require unverifiable assumptions to control for selection into treatment (Rosenbaum, 2017; Ding, 2024).

To explore HTE, we focus on the Conditional Average Treatment Effect, or CATE, defined as $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$, where X_i is a vector of baseline covariates. The CATE

captures how treatment impacts vary across subpopulations as defined by X_i. ML methods are well-suited to CATE estimation due to their flexibility in modeling complex relationships, their ability to handle high-dimensional covariate spaces, and their use of regularization to mitigate overfitting (Qian & Murphy, 2011; Hill, 2011; Belloni et al., 2014; Tian et al., 2014; Chen et al., 2017; Wager & Athey, 2018; Nie & Wager, 2021; Athey & Imbens, 2019). Importantly, the CATE is not the same as the individual effects; even given a set of covariates, we may have idiosyncratic variation beyond what those covariates can explain (Ding, Feller & Miratrix, 2019). In this work we assume the best-case scenario that there is no such idiosyncratic variation or, equivalently, that the CATE is the true object of interest.

We examine 18 selected ML methods—more than any prior simulation study of this type that we know of excluding competitions such as the American Causal Inference Challenge (ACIC; Dorie et al, 2019)—across diverse data-generating processes. We use the term 'ML method' to refer to a unique combination of base models and meta-learner. We attempted to select methods representing dominant approaches in the causal ML literature, spanning diverse theoretical properties from double robustness to nonparametric flexibility. Our included methods are listed below, categorized into meta-learners, modified ML models, and reference cases.[1]

Meta-learners are general frameworks that use base models to estimate CATEs. Using taxonomy initially discussed in Künzel et al. (2019), we evaluate the following:

- **S-Learner**: Estimates a single model of the outcome and includes treatment indicator T as a covariate. The CATE is then estimated from the difference in predicted outcomes under alternative values of T. We implement this with Bayesian Additive Regression Trees (BART S) and with ordinary least squares with treatment-covariate interactions (OLS S, a standard practice that can be used as a reference point) and OLS with all pairwise covariate

interactions (OLS S INT). OLS S is simple and interpretable; OLS S INT significantly increases model complexity by including all possible interactions; and BART S can capture non-linear relationships and interactions while offering Bayesian regularization to protect against overfitting.

- **T-Learner**: Estimates $Y(1)$ and $Y(0)$ separately for the treatment and control groups, then takes the difference to estimate CATE. We use LASSO (with and without covariate pairwise interactions), Random Forests (RF), and BART as base models (LASSO T, LASSO T INT, RF T, BART T*)*. LASSO T is suitable for regularizing high-dimensional data when the true DGP may be sparse; RF T can capture non-linear relationships and interactions. We also partially implemented an ensemble approach, Super Learner T (SL T), that uses four base models (OLS, LASSO, RF, Conditional Inference Forest). It has shown good performance in causal inference competitions (Thal and Finucane, 2023). Due to excessive computation time, this method was included in only a limited number of simulations.
- **R-Learner**: Designed to directly estimate the treatment effect by focusing on residualized outcomes and treatment indicators, the R-learner can perform well even if the nuisance parameters (like propensity scores or outcome models) are not perfectly estimated (Robinson, 1988; Nie & Wager, 2021; Zhou, Zhang, & Tu, 2022). We use LASSO for both nuisance parameter estimation and CATE prediction (LASSO R).

Modified ML Models are adaptations of ML models specifically designed for causal inference. Many can be viewed as variants of S-, T- and R- learners:

- **Modified Outcome Methods (MOM)**: These methods transform the outcome variable and then use ML algorithms to model this transformed outcome. Inverse probability weighting (IPW; Robins, 1994) is computationally efficient but sensitive to propensity score

misspecification; doubly robust methods (DR; Robins, Rotnitzky, & Zhao, 1994) offer robustness against misspecification in either the propensity score or outcome model. We evaluate four variants of this approach (LASSO MOM IPW, RF MOM IPW, LASSO MOM DR, RF MOM DR).

- **Modified Covariate Methods (MCM)**: These methods reparameterize covariates to directly model treatment effect heterogeneity by including treatment-by-covariate interaction terms (Tian et al., 2014). MCM provides interpretable effect modification estimates while balancing sparsity. We evaluate both the standard MCM and its efficiency-augmented version (MCM EA; Zhang et al., 2023), using LASSO (LASSO MCM, LASSO MCM EA).
- **Causal Forests (CF)**: These methods extend the random forest framework to estimate CATE nonparametrically (Wager and Athey, 2018). Causal Forests share conceptual foundations with the R-learner (Nie & Wager, 2021), as both employ residualization to isolate treatment effect heterogeneity from prognostic effects. We evaluate a standard version (CF; Athey et al., 2019) and a variant with local centering (CF LC; Athey et al., 2019).
- **Double Machine Learning**: This approach extends the R-learner framework by leveraging cross-fitting and orthogonalization to reduce regularization bias and overfitting in high-dimensional settings (Chernozhukov et al., 2018). We implement the normalized DR-learner proposed in Knaus (2022) that combines ordinary least squares (OLS) for propensity score estimation with random forests (RF) for outcome regression, balancing parametric efficiency with nonparametric flexibility (CDML).

We also include three additional **Reference Cases** to serve as benchmarks:

- **Average Treatment Effect (ATE):** This method simply assigns the estimated ATE to all units, calculated as the difference in mean outcomes between treatment and control groups.

- **Infeasible Learners (LASSO INF and RF INF)**: These methods use known treatment effects as the prediction target and represent 'performance ceilings' for similar methods.

Many prior simulation studies have evaluated subsets of the above methods. Table 1 summarizes several studies along with their key findings. While results generally align with theoretical expectations, the lack of consensus on a superior method highlights the need for further research. Our current study (bottom row), which evaluates a wide range of methods, seeks to fill this gap.

3. Simulation Framework

We use an empirical Monte Carlo simulation framework (Huber et al., 2013) based on real data to generate synthetic datasets with realistic patterns of treatment effects, ensuring empirical plausibility.

We base our simulations on two educational RCTs: Career Academies (CA) and Accelerated Study in Associate Programs (ASAP). These studies were selected based on policy relevance, moderate-to-large sample sizes (CA: N = 1,764; ASAP: N = 2,397), and rich covariate sets.[2] Each study provides one continuous outcome (average monthly earnings for CA; college credits for ASAP) and one binary outcome (full-time employment for CA; degree attainment for ASAP). Hence, the generalizability of our results is enhanced relative to studies using single datasets or outcome types.

3.1 Data Generating Process (DGP)

We generate data in three stages:

Study	Data	S -learner	T-learner	R-learner	Modified ML	Other	Summary of results
Lu et al. (2018)	Synthetic	RF*, RF-I, BART	RF, SF		BI, CF		SF T-learner gives lowest RMSE, especially with larger sample size; The RF-I S-learner and RF/BART T-learners outperform others, including CF.
Powers et al. (2018)	Synthetic	Boost	RF		CF, MARS, MOM-IPW(RF)**		No single approach dominates. MARS gave the lowest MSE in three of the eight simulation scenarios but highest RMSE in the no-heterogeneity case.
Wendling et al. (2018)	Empirical	BART, Boost***			CF-LC, MARS***		With little heterogeneity, all methods had low RMSE. With moderate heterogeneity, BART and causal boosting outperformed MARS and CF. With complex heterogeneity, the differences between the approaches reduced.
Künzel et al. (2019)	Synthetic	RF, BART	RF, BART			X(RF), X(BART)	No method dominates. Base learner choice affects prediction accuracy.
Knaus et al. (2021)	Empirical		RF, LASSO	LASSO	CF, CF-LC, MOM-IPW(RF), MOM-IPW(LASSO), MOM-DR(RF), MOM-DR(LASSO), MCM(LASSO), MCM-EA(LASSO)		No method dominates. MOM-DR (RF), MCM-EA(LASSO), R-learner(LASSO) and CF-LC perform reliably. MOM-DR (LASSO) may be a good choice for large samples. MOM-IPW

							(LASSO), CF and T-learners performed weakly.
Caron et al. (2022)	Empirical	RF, BART	RF, BART	LASSO, Boost	CF	X(RF), X(BART), CMGP, NSGP, BCF	With group-specific potential outcomes and complex heterogeneity, multitask learners and BCF perform best and T-learners outperform R-, S- and X-learners. With simpler heterogeneity, X-learner (RF) and the BCF perform best, only slightly trailed by CF.
Current paper	Empirical	*OLS, OLS INT,* BART,	LASSO, LASSO INT, *RF, BART, SL T*	*LASSO*	*CF*, CF LC, *MOM IPW(RF)*, MOM IPW(LASSO), MOM DR(RF), MOM DR(LASSO), MCM(LASSO), MCM EA(LASSO), *CDML*		*See main findings in results section*

Table 1 *Summary of Simulation Studies Evaluating ML Methods for Treatment Heterogeneity Estimation*
Notes. BI - bivariate imputation (Lu et al. 2018); RF - random forest; RF-I - RF with treatment interactions; CF - causal forest; CF-LC - CF with local centering; BART - Bayesian additive regression tree; SF - synthetic forest (Ishwaran and Malley 2014); LASSO - Least Absolute Shrinkage and Selection Operator; BOOST - Boosted Regression; SL-T - ensemble of multiple T-learners; MOM-IPW - Modified outcome method (MOM) with Inverse Probability Weighting (IPW) and specified base learner; MOM-DR - MOM with Doubly Robust (DR) estimation and specified base learner; MCM - Modified covariate method (MCM) with specified base learner; MCM-EA - MCM with efficiency augmentation and specified base learner; X – X-learner with specified base learner; MARS - Multivariate Adaptive Regression Spines; RLR - Regularized logistic regression (LASSO + Ridge) (Wendling et al., 2018); CMGP - Causal multitask Gaussian process; NSGP - Non-stationary Gaussian process; BCF - Bayesian Causal Forests (BART)
** Lu et al call this approach "Virtual Twins"*
*** Powers et al. (2018) refer to this as "pollinated transformed outcome forest"*
**** Wendling et al. (2018) also include variants of BART, Boost and MARS that include the propensity score as a covariate.*

- **Covariate Generation**. Covariates, X, are generated using the R package *synthpop* (Nowok et al., 2016), which preserves the marginal and conditional distributions of the original data and avoids duplicate profiles. For each source dataset, we create a synthetic population of N=100,000.
- **Untreated Potential Outcome Generation**. We use a random forest fitted to the empirical data control group to predict untreated potential outcomes for all synthetic units. For continuous outcomes, we then use a copula approach to transform the actual outcomes conditional on the predicted values to $Y(0)$. The copula works by sampling implied Z-scores of the ranks of the predicted untreated outcomes. The cumulative distribution function of these Z-scores mapped through the empirical distribution of the original controls yields $Y(0)$. The copula preserves distributional features while allowing the generated outcomes to be correlated with the covariates to the same degree as in the source data. It also allows us to scale up the initial (actual) dataset to much larger samples with a consistent structure, while avoiding duplicate observations. More details of this approach can be found in Supplement Section B. For binary outcomes, we use the random forest to generate predicted probabilities, then convert these to 0s and 1s with Bernoulli draws.
- **Treatment Effect Generation**. Rather than imposing arbitrary functional forms for treatment effect variation, we extend a data-driven approach inspired by Wendling et al. (2018) and generate HTEs using a *multiple-queens* framework that creates realistic patterns of effect heterogeneity.

To implement this *multiple-queens* approach, we designate nine ML methods (highlighted in italics in Table 1) as queens. We generate a distinct pattern of heterogeneous treatment effects for each queen as follows:

1. Train each selected queen on the actual RCT data using a specified covariate set to estimate CATEs.
2. Apply the trained model to the synthetic population to generate predicted treatment effects, τ_i, for each unit i. These predictions reflect plausible patterns of effect heterogeneity consistent with the original data.
3. Scale τ_i to a fixed magnitude (0.2 standard deviations for continuous outcomes and 10 percentage points for binary outcomes).
4. Calculate treated potential outcomes as $Y_i(1) = Y_i(0) + \tau_i$. For continuous outcomes, this calculation is straightforward. For binary outcomes, we first winsorize any values outside [0,1] to ensure valid probabilities, and then generate binary outcomes by drawing from Bernoulli distributions with these winsorized probabilities as success rates.

The *multiple-queens* approach does not claim to recover the "true" pattern of impacts from the original RCTs, as individual-level ground truth is fundamentally unknowable. Instead, each queen creates its own “ground truth”—a synthetic but plausible heterogeneity pattern against which we can objectively measure all estimation methods' performances. Any subgroup heterogeneity emerges naturally from how each queen learns patterns from the original data. We do not manually specify which subgroups should experience larger or smaller effects. The *multiple-queens* approach preserves realistic covariate relationships while creating varied heterogeneity patterns for testing estimation methods. It also avoids unfairly advantaging methods similar to those used for generating impacts (as noted by Gao et al., 2020). For example, LASSO-based queens generate sparse heterogeneity from a small number of covariates, which may favor LASSO-based estimation methods. Similarly, random forest queens produce fine-grained patterns with complex interactions, potentially favoring RF type estimation methods. We

selected queens that span a range of complexity to ensure a robust evaluation. Supplement Section E illustrates such complexity.

3.2 Simulation Procedure

Once we have generated the synthetic population of 100,000 units, we evaluate the ML methods' performance through a systematic simulation process for each queen.

3.2.1 Core Simulation Framework

For each scenario, we run 100 iterations (S = 100) through 21 ML methods (J = 21, 18 methods under evaluation plus 3 reference cases). A fixed test set of 10,000 units (n_{test} = 10,000) is sampled from the synthetic population for each queen, *k*. Each test unit has a specific covariate profile and a known "true" treatment effect generated by the queen. For each iteration (s), a sample with n units ($n = 1{,}000, 2{,}000, or\ 5{,}000$) is randomly drawn from the remaining synthetic population to represent a hypothetical experiment. We then randomly assign units to treatment (T=1) or control (T=0) in equal proportions and calculate observed outcomes as $Y^{obs} = TY(1) + (1 - T)Y(0)$. Each ML method is trained on this generated "experiment" using off-the-shelf R implementations and then used to predict treatment effects for all test units ($\hat{\tau}_{ijs}$). This process yields an n_{test} × J × S array of predictions for each queen.

3.2.2 Simulation Parameters

We systematically vary several key parameters:

1. **Outcome type**: Each method is tested on two continuous and two binary outcomes (from CA and ASAP) to assess performance differences by outcome type.
2. **Covariate set size:** We vary the number of available covariates (see Table 2). Only the small set is used to generate treatment effects, so additional covariates introduce noise, testing each method's ability to handle irrelevant information.

Data Set	Number of Covariates in Small Set	Medium Set	Large Set
Career Academies	6	6 + 8 = 14	6 + 8 + 13 = 27
Accelerated Studies in Associate Program	6	6 + 3 = 9	6 + 3 + 9 = 18

Table 2 *Number of Covariates in Small, Medium, and Large Covariate Sets*

Note. A complete list of these covariates is available in Supplement Section C.

3. **Sample size of the hypothetical experiment**: Sample sizes of 1,000, 2,000, or 5,000 are considered**.**

4. **Impact heterogeneity magnitude**: For a subset of scenarios with continuous outcomes, we increase impact heterogeneity magnitude from 0.2 to 0.4 standard deviations to explore how performance varies with treatment effect variation.

Each unique combination of outcome, covariate set size, sample size of the hypothetical experiment, and impact variation defines a simulation "scenario." This multifactor design enables comprehensive evaluation across diverse conditions.

3.3 Performance Metrics

We use five performance metrics for each ML method *j*:

- Average Root Mean Squared Error (RMSE): $AvgRMSE_j = \sqrt{\frac{1}{n_{test}}\sum_{i=1}^{n_{test}} MSE_{ij}}$
- Average Standard Error: $AvgSE_j = \sqrt{\frac{1}{n_{test}}\sum_{i=1}^{n_{test}} SE_{ij}^2}$
- Average Bias: $AvgBias_j = \sqrt{AvgRMSE_j{}^2 - AvgSE_j{}^2}$
- Average Spearman's Rank Correlation: $\rho_j = \frac{1}{S}\sum_{s=1}^{S} Cor(\, rank(\, \hat{\tau}_{ijs}\,), rank(\tau_i))$
- Average R^2 measure: $R_j^2 = \frac{1}{S}\sum_{s=1}^{S} Cor(\, \hat{\tau}_{ijs}, \tau_i\,)^2$

The first three measures average individual test-point performance, as described in Supplement Section D. The average RMSE represents the typical error when predicting the CATE. Average SE measures how much our predictions vary across iterations for the same scenario and queen. Average bias measures how systematically biased predictions are for a randomly chosen test point.[3]

The average RMSE, bias, and SE may not fully capture performance when ML methods are used simply to identify which units have larger treatment effects. We thus also calculate the average Spearman rank correlation (ρ) and average treatment R^2 to assess how well the estimated impacts predict the true impacts. These metrics are calculated for each iteration and then aggregated at the method level across iterations.

In initial explorations, we found that some methods occasionally produced treatment effect predictions greatly exceeding $\pm 1\sigma$, a large threshold in our context. To mitigate extreme outliers, we winsorized all predictions at $\pm 1\sigma$. All performance metrics are calculated using these winsorized predictions. We record the proportion of winsorized predictions for each method as an indicator of method stability.

4. Simulation Results

We first assess method stability when there is no treatment variation, then illustrate performance for a continuous outcome in a single scenario. We subsequently examine performance across all outcomes and scenarios and explore how results vary by covariate set size, hypothetical experiment sample size, and magnitude of treatment effect heterogeneity.[4]

4.1 General Stability

Figure 1 displays the average standard error (SE) across scenarios with no treatment variation (the ATE queen). T-learners, such as LASSO-T, tend to be less stable (higher average standard

error) due to estimating two separate models. Regularized S-learners, by contrast, jointly regularize the two surfaces, which increases overall stability; see, for example, Hahn et al. (2018) for further discussion on stability and regularization-induced confounding.

All methods become more stable with increased hypothetical experiment sample size, with notable improvements for CDML and OLS S. CF and CF LC are less sensitive to sample size.

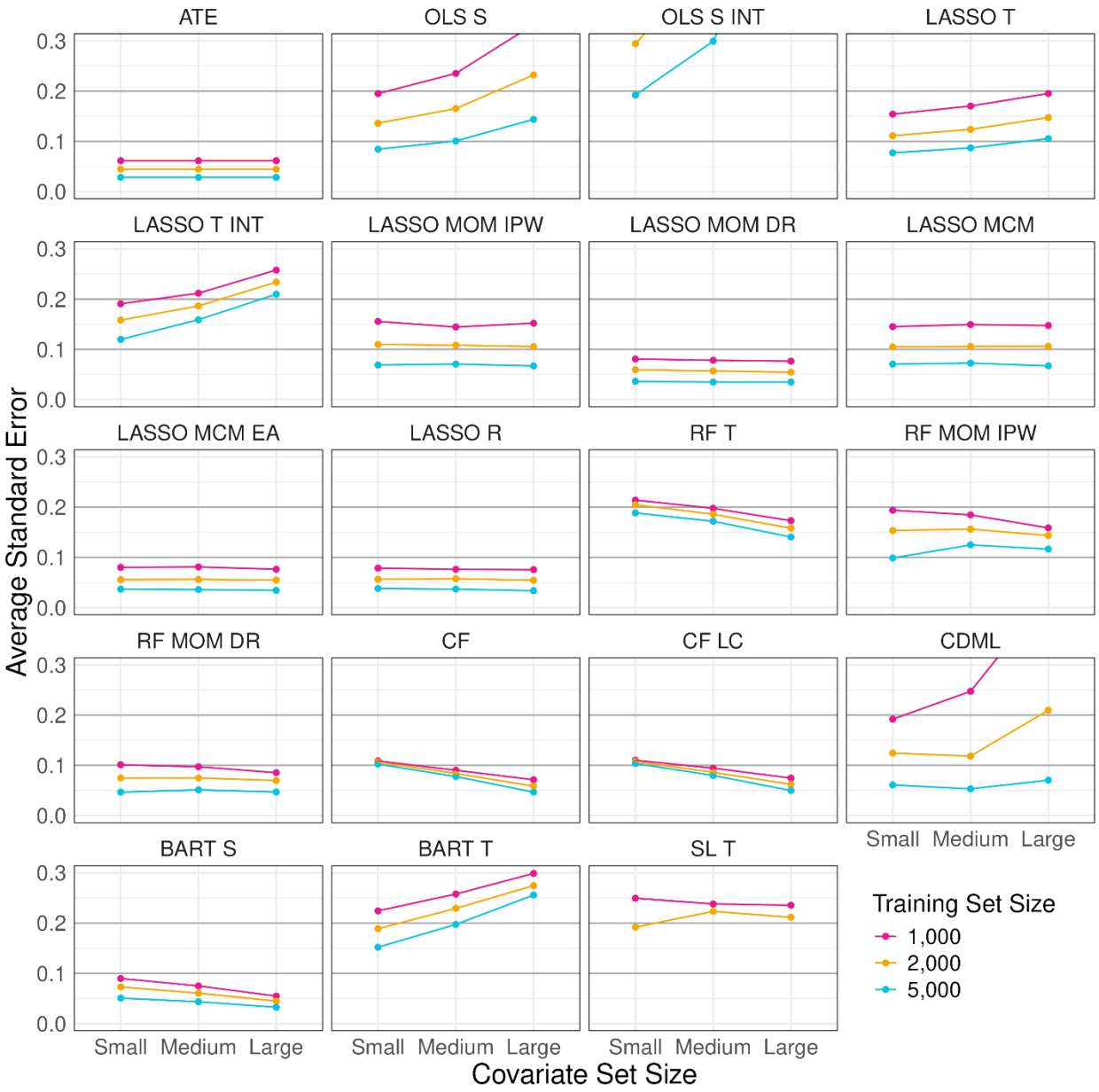

Figure 1 *Average SE Across All Test Points for Various Scenario Types*

Note. SE values represent prediction stability, averaged across 10,000 test units and two continuous outcomes from CA/ASAP datasets. Points grouped by covariate set size (small, medium, and large), with color indicating hypothetical experiment sample size (pink = 1,000, yellow = 2,000, blue = 5,000). Y-axis truncated at 0.30σ for readability, excluding extreme values from OLS S, OLS S INT, and CDML. Values for SL T are based on 7 queens.

The results show that the LASSO variants using MOM, MCM, and R-learner show consistent stability across covariate sets. In contrast, LASSO T deteriorates with additional covariates. Interestingly, the opposite pattern occurs for RF T. Similar to LASSO, RF MOM DR outperforms RF MOM IPW, but the difference is more pronounced. Overall, T-learners do not appear to be a good choice in terms of average precision; compared to both the S-learner (with BART as the base learner in either case) and the R-learner (with LASSO as the base learner in either case), they perform worse. When interpreting these results, we note that the off-the-shelf nature of the implementations means that the degree of automatic tuning varies across learners.

Increasing the covariate set size increases the average SE for OLS S, OLS S INT, and, especially with small hypothetical experiment sample sizes, CDML. This is also the case for LASSO T, LASSO T INT, and BART T. Other methods appear to cope better with larger covariate sets and, in some cases, even show reduced average SE when there is extraneous information (RF T, CF, CF LC, BART S).

4.2 Illustration of Method Performance for One Outcome

We illustrate method performance for the average monthly earnings outcome in the CA dataset,[5] using the small covariate set, a hypothetical experiment with 1,000 units, and treatment heterogeneity of 0.2σ. Figure 2 plots average SE (x-axis) against absolute average bias (y-axis), with contour lines for equal average RMSE. Methods closer to the origin perform better.

There is substantial variation in both bias and standard error.BART S achieves a low overall RMSE by balancing relatively low bias with moderate standard error. The T-learners (LASSO T, RF T, BART T) tend to have higher standard errors, reflecting their two independent estimation processes. The ATE method, which simply assigns the average treatment effect to all units, has high bias but low standard error, placing it on a moderate RMSE contour that outperforms most other ML methods.

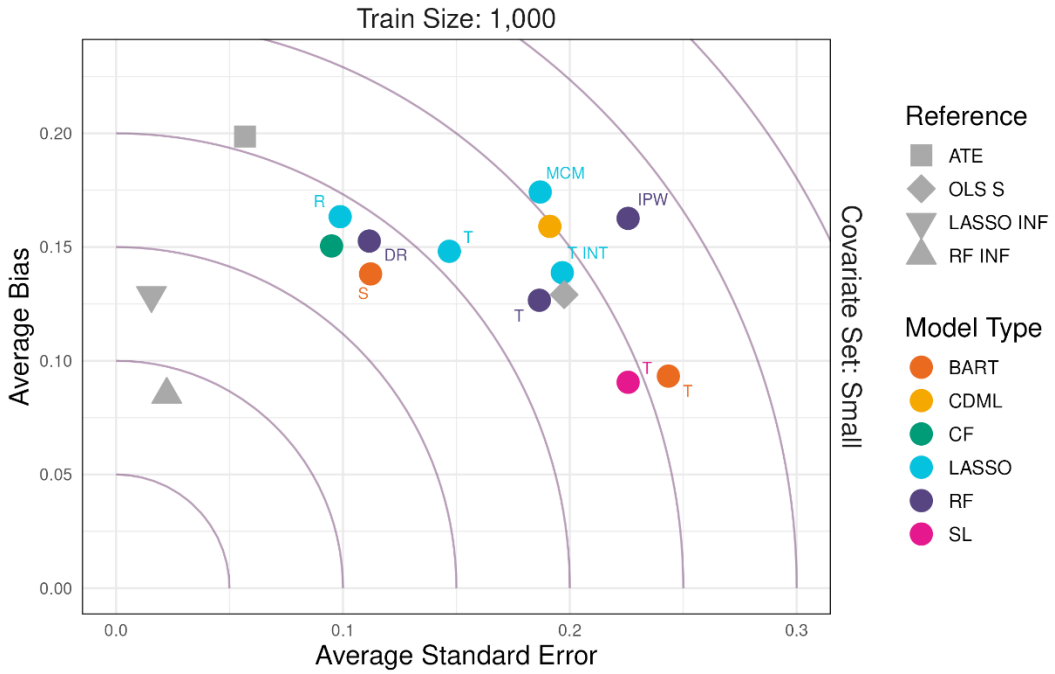

Figure 2. *Average Bias and SE for All Methods in a Single Scenario*

Note. Results for small covariate set and small sample size ($n = 1{,}000$), averaged across all queens for the continuous outcome from the CA dataset with 0.20σ treatment variation. Contour lines denote equal average RMSE. Some methods (LASSO MCM EA, LASSO MOM DR, CF LC, LASSO MOM IPW, OLS S INT) cannot be seen in the plot because of overlap with others or out of range values. Values for SL T are based on 7 queens.

The oracle methods (LASSO INF and RF INF) use the true treatment effects during training and thus establish a performance ceiling. Among the feasible methods, BART S and CF emerge as the strongest performers in this scenario, with LASSO R and RF MOM DR also showing

competitive performance. BART T and SL T have the lowest bias but this comes at the cost of high standard errors.

We also evaluate how well methods sort individuals by responsiveness using R^2 and Spearman's Rank Correlation (ρ) measures: the first captures whether the estimated effects have a strong linear correlation with the truth, and the second captures whether the rank ordering of the estimates generally corresponds with the truth. For this specific scenario, the Spearman's rank correlations range from 0.10 (LASSO MCM) to 0.48 (SL T), and the R^2 measures range from 0.06 (LASSO MCM) to 0.28 (SL T) for the non-reference methods (not shown in Figure 2).

4.3 Overall Performance Across All Scenarios

We now examine method performance across all simulation scenarios. We expect performance to vary by the data-generating process (queen) and scenario characteristics. Linear methods should perform well when the true model for impacts is linear, while flexible methods like BART S and CF might perform well across diverse settings. As an initial exploration, we examine the four best methods (LASSO R, BART S, CF, and RF MOM DR) and two low-bias methods (BART T and RF T) from the single-scenario analysis and reproduce the contour plot.

We plot the performance by individual queen and include our three levels of covariate set size and hypothetical experiment sample size and show all four of our outcomes (continuous and binary from CA and ASAP datasets). Plots for other methods are available in Supplement Section F. The top and bottom panels in Figure 3 display performance for the continuous and binary outcomes, respectively. Each plot represents a different combination of method (columns) and hypothetical experiment sample size (rows), with individual points in each plot representing the performance of a specific method against a specific queen using a specific dataset and covariate set. The overall average across all scenarios within each plot is marked '×'.

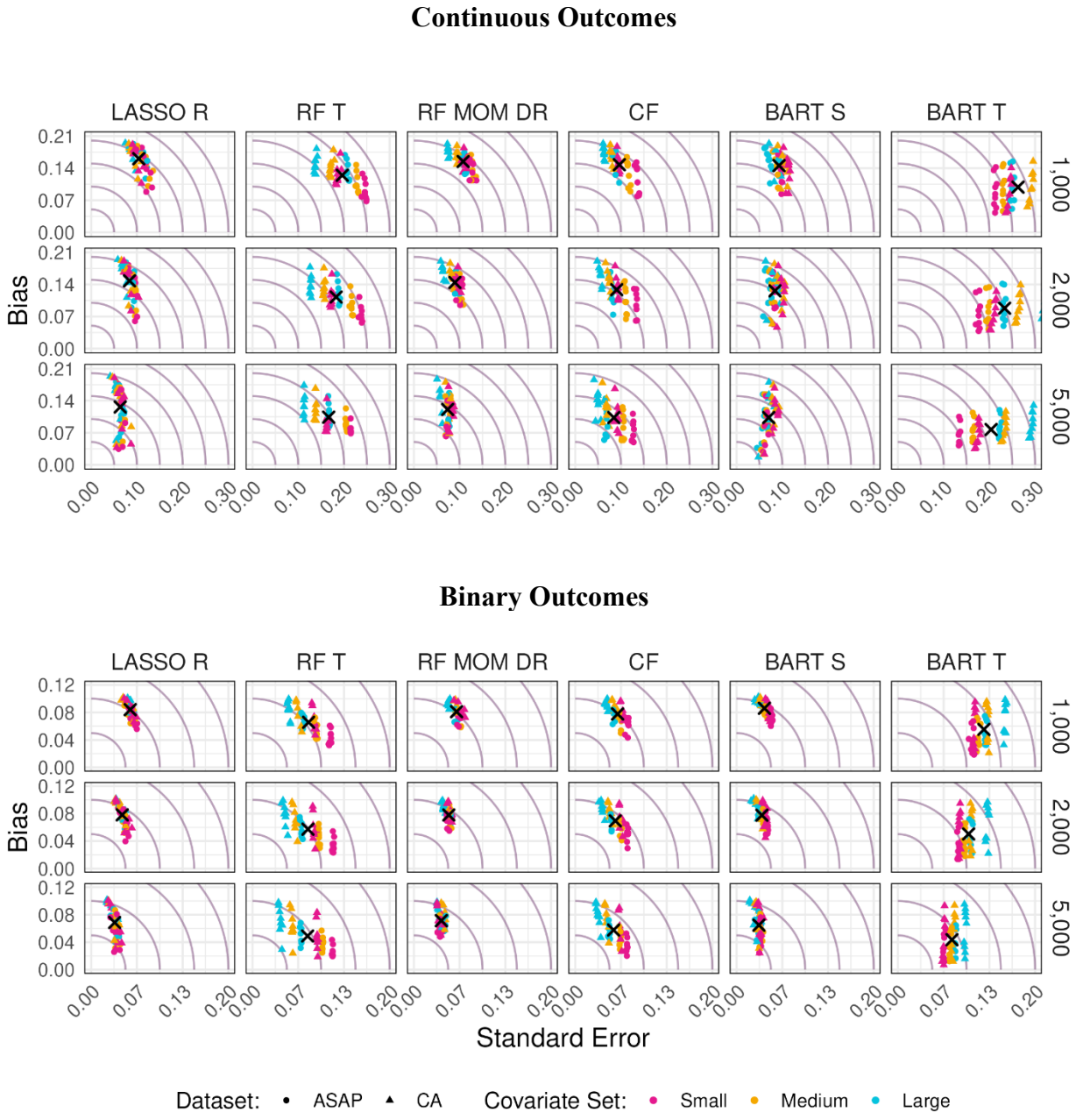

Figure 3 *Performance of Selected ML Methods by Queen Across Simulation Scenarios, by Outcome Type*

Note. Each point represents the performance of a machine learning method against a specific queen (excluding the ATE queen) in each scenario, averaged across two outcomes of the same type. "×" represents the overall average across all scenarios within each facet. Some data points for BART T are truncated due to values outside the range displayed.

Distinct patterns emerge. First, the vertical alignment of points for each covariate set size suggests that, for most methods, standard errors are consistently similar across queens within each sample size, covariate set size, and outcome, indicating that prediction stability depends more on the method than on the heterogeneity pattern. Second, greater vertical dispersion of points for methods like LASSO R (particularly with larger experiment samples) reveals substantial performance variation across queens, with much less bias against some data-generating processes than others. Such variation confirms that methodological relationship between the data-generating process and estimation approach can create advantages, underscoring the value of our multiple-queens framework for comprehensive evaluation.

Comparisons between the two datasets (Career Academies and ASAP) reveal broadly similar patterns, suggesting that our findings generalize across different educational contexts rather than being dataset specific. For binary outcomes, performance characteristics largely mirror those for continuous outcomes, with some exceptions. For instance, BART T demonstrates more stable performance with binary outcomes as covariate complexity increases. For binary outcomes, the relative ranking of methods by average RMSE is similar to that observed with continuous outcomes (see Supplement Figure F3).

We next assess overall performance and performance consistency. To do so, within each outcome, scenario, and queen, we calculate each method's RMSE, bias, and standard error relative to the median performance among all methods and then average these ratios across all outcomes, scenarios, and queens (excluding the ATE queen). Table 3 summarizes these metrics, ordering methods by increasing relative RMSE. BART S emerges as the top performer, with an average relative RMSE of 83% (17% better than median), outperforming the simple average treatment effect (ATE) method in 97% of scenarios. It achieves this through low standard error

(72% of median) while maintaining reasonable bias (93% of median). The next tier of methods—including CF, LASSO R, CF LC, LASSO MCM EA, LASSO MOM DR, RF MOM DR, and LASSO T—achieve approximately 7–10% RMSE reductions compared to the median.

	Relative to Group Median (=100)			Percent of Time (%) RMSE lower than		Average		Percentage
Model	RMSE	BIAS	SE	Median	ATE	R^2	Spearman's ρ	Winsorized
BART S	83	93	72	85	97	0.41	0.55	0
CF	90	90	98	83	95	0.35	0.52	0
LASSO R	90	101	80	68	86	0.32	0.43	0
CF LC	90	91	98	83	94	0.34	0.51	0
LASSO MCM EA	90	101	80	67	86	0.32	0.43	0
LASSO MOM DR	90	101	80	67	85	0.32	0.43	0
RF MOM DR	92	104	88	76	92	0.31	0.49	0
LASSO T	93	99	90	61	76	0.33	0.46	0
LASSO T INT	105	95	120	45	59	0.30	0.44	0.2
LASSO MCM	108	121	104	25	42	0.15	0.24	0.1
LASSO MOM IPW	108	121	105	26	41	0.15	0.24	0.1
ATE	**109**	**149**	**40**	**0**	**0**	**NA**	**NA**	**0**
RF T	111	78	159	28	42	0.30	0.48	0.1
RF MOM IPW	112	117	126	22	33	0.16	0.34	0
OLS S	123	72	185	33	37	0.32	0.48	0.7
BART T	124	61	202	23	28	0.33	0.5	0.4
CDML	131	110	170	34	39	0.25	0.42	2.6
OLS S INT	283	88	495	2	3	0.12	0.27	25.2

Table 3 *Overall Performance of ML Methods Across All Simulation Scenarios and Queens*

Note. NA=not applicable. Relative RMSE, bias, and SE are reported as percentages of the group median (median=100; values below 100 indicate better-than-median performance). "Percent Better Than Median" and "Percent Better Than ATE" indicate the percentage of all scenario and queen (excluding the ATE queen) combinations in which the method outperformed the group median and the ATE approach, respectively. Average R^2 and Spearman's ρ measure the correlation between predicted and true treatment effects. The "Winsorized" column reports the percentage of predictions requiring truncation due to extreme values. Rows are sorted by relative RMSE.

The coefficient of determination (R^2) and Spearman's rank correlation (ρ) assess the strength of the association between predicted and true treatment effects. These measures are of particular interest since the ability to identify those most affected by a policy is perhaps the key aim of understanding heterogeneous treatment effects.

The results reveal limited predictive ability on average, with the best performer, BART S, being the only method to exceed a R^2 of 0.40. For Spearman's ρ, four methods—BART S, CF, CF LC, and BART T—reach an average of 0.5 or higher. Because it is a direct measure of how well methods rank individuals correctly, subsequent presentations focus on Spearman's ρ, with R^2 results available in the supplement.

In our context, a ρ of 0.5 means roughly 55–65% of individuals predicted to be in the top response quintile truly belong there; the remainder are misclassified, often falling into adjacent quintiles. Supplement Section G provides illustrations of cross-quintile agreement.

Finally, most methods (16 of 18) produced stable predictions with negligible winsorization (≤0.7%), indicating rare extreme outlier predictions. However, CDML and OLS S INT exhibited higher winsorization rates (2.6% and 25.2%, respectively), raising concerns about their stability.

The first three columns in Figure 4 visualize the distribution of relative performance metrics (RMSE, bias, and standard error) across all outcomes, scenarios, and queens. Values below 1.0 indicate better-than-median performance. BART S shows the best average performance but with some variation. The ATE approach exhibits consistently high bias but very low variance, resulting in moderate RMSE. Wider boxplots for methods like CDML, BART T, and OLS S INT indicate less consistent performance, suggesting sensitivity to specific data conditions. The last column shows the distribution of Spearman's ρ and reveals that several methods (CF, CF LC, RF MOM DR) are comparable to BART S in terms of performance on this measure.

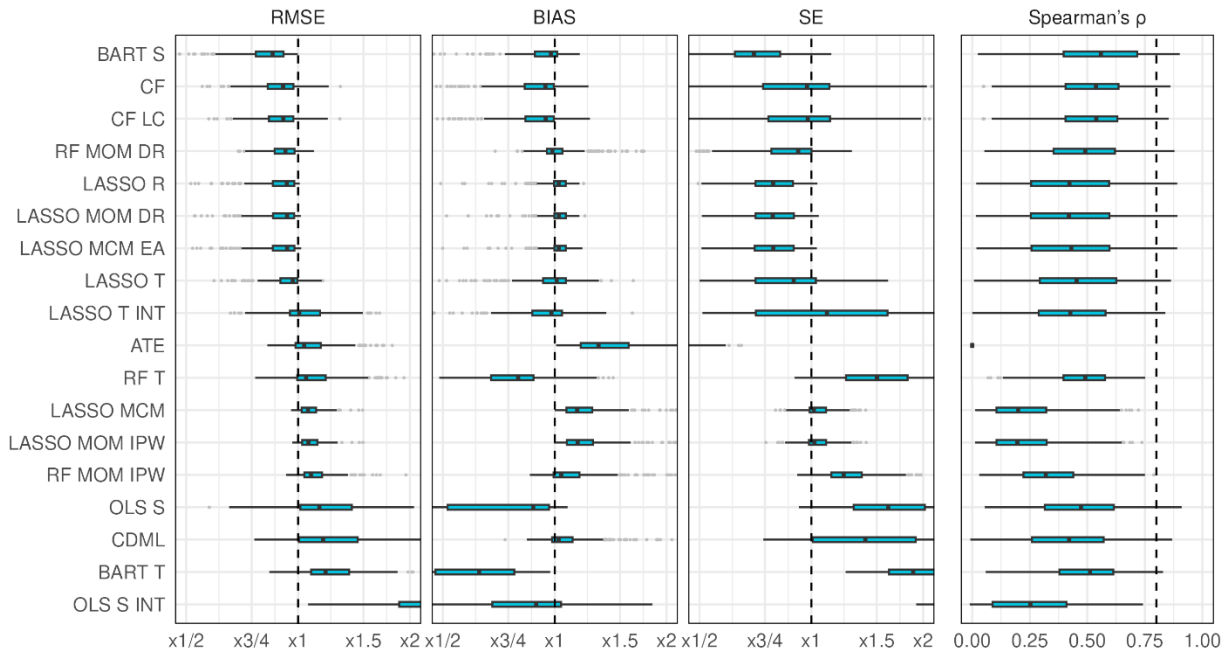

Figure 4 *Boxplots of Relative Performances Metrics by ML Method*

Note. Each boxplot summarizes 300+ simulations per method, spanning all outcomes, scenarios and queens. Relative performance for the average RMSE, Bias, and SE with values below 1.0 indicates better-than-median performance. Narrower boxes reflect greater consistency in performance across simulation conditions.

4.4 Performance Variation Across Key Dimensions

We next assess how performance changes as the number of covariates increases, as the size of the hypothetical experiment increases, and as the amount of treatment variation increases. We fit random effect models (one each for log RMSE, log absolute bias, log SE, and the average Spearman's ρ) to study the impact of such changes. Details of these models are described in Supplement Section H.

4.4.1 Impact of Including More Covariates

A central advantage of ML is the ability to handle high-dimensional data by selecting the most relevant covariates. We test this by examining how performance changes as the number of covariates increases. In all scenarios, the true impact heterogeneity is fully explained by the

"small" covariate set, making any additional covariates noise. However, due to the DGP they may still be correlated with the outcome and other covariates.

Table 4 presents the impact of increasing the number of covariates on the performance metrics, with each row showing the percentage change in performance (absolute change for Spearman's ρ) relative to the small covariate set (baseline).

ML Method	RMSE			BIAS		SE		Spearman's ρ		
	Baseline	Med	Lar	Med	Lar	Med	Lar	Baseline	Med	Lar
BART S	0.11	2%	6%	9%	20%	-9%	-22%	0.59	-0.04	-0.09
LASSO R	0.12	3%	8%	8%	19%	-4%	-7%	0.48	-0.05	-0.11
LASSO MCM EA	0.13	3%	8%	8%	18%	-4%	-7%	0.48	-0.05	-0.11
LASSO MOM DR	0.13	3%	8%	8%	18%	-4%	-7%	0.48	-0.05	-0.11
RF MOM DR	0.13	3%	5%	7%	13%	-3%	-13%	0.54	-0.05	-0.11
CF	0.13	-2%	-4%	13%	30%	-18%	-38%	0.54	-0.03	-0.04
CF LC	0.13	-2%	-3%	13%	30%	-17%	-36%	0.54	-0.03	-0.05
LASSO T	0.13	2%	7%	0%	-5%	7%	22%	0.48	-0.02	-0.04
LASSO T INT	0.14	7%	19%	4%	6%	11%	29%	0.49	-0.05	-0.10
OLS S	0.14	17%	54%	10%	19%	24%	78%	0.55	-0.07	-0.15
LASSO MCM	0.15	3%	7%	6%	12%	-2%	-3%	0.30	-0.05	-0.12
LASSO MOM IPW	0.15	3%	7%	6%	12%	-2%	-2%	0.30	-0.05	-0.11
BART T	0.15	15%	36%	15%	35%	16%	38%	0.57	-0.06	-0.15
ATE	**0.15**	**0%**	**0%**	**1%**	**0%**	**-1%**	**-1%**	**NA**	**NA**	**NA**
CDML	0.16	11%	49%	7%	16%	11%	72%	0.51	-0.08	-0.19
RF MOM IPW	0.16	3%	2%	5%	9%	2%	-5%	0.39	-0.05	-0.11
RF T	0.17	-6%	-14%	18%	46%	-15%	-34%	0.50	-0.02	-0.04
OLS S INT	0.24	51%	149%	34%	103%	52%	155%	0.44	-0.15	-0.35
OVERALL	0.15	6%	15%	9%	21%	1%	4%	0.46	-0.05	-0.11

Table 4 *Estimated Percent Change in Average RMSE, Bias, and SE, and Change in Spearman's ρ, as Number of Covariates Increases, by ML Method*

Note. Baseline = small covariate set; Med = medium covariate set; Lar = large covariate set. Results are from multilevel regression model with random effects for the ML method. Values represent the average percent change in average RMSE, Bias, and SE and a decrease in Spearman's ρ when moving from the small covariate set to medium or large set. Rows are sorted by baseline RMSE.

Table 4 shows that performance generally worsens as covariate complexity increases, but the extent varies by method. The LASSO variants show relatively modest degradation, reflecting their built-in variable selection capabilities, which help filter out irrelevant covariates. In

contrast, methods like CDML and the unregularized OLS variants show much larger performance declines.

Methods like RF T, CF, and CF LC exhibit a trade-off, with standard error improving while bias worsens. This suggests that additional covariates invoke a greater degree of regularization, which helps these methods become more stable, but at the cost of less accurately capturing heterogeneity patterns. In some cases, this trade-off results in a net improvement in RMSE.

4.4.2 Impact of Increasing Experiment Sample Size

We next examine how performance changes as our hypothetical experiment size increases from 1,000 to 2,000 and 5,000, using the same approach as described above. While larger samples are expected to improve performance, some methods may benefit more than others.

Table 5 shows that all methods benefit substantially from larger experiment samples, with average RMSE reductions of 13% for 2,000 versus 1,000 observations (baseline) and 28% for 5,000 versus 1,000. The improvements are more pronounced for standard error (39% reduction) than for bias (21% reduction), indicating that larger samples primarily help by stabilizing predictions. Method-specific gains vary: CDML shows the largest improvements, with a 53% RMSE reduction at n=5,000, followed closely by OLS S with 45%. RF T shows the smallest gains (13% RMSE reduction), suggesting it is less able to leverage additional data compared to other methods. Predictive power as captured by Spearman's ρ increases substantially with larger experiment sample sizes.

ML Method	RMSE			BIAS		SE		Spearman's ρ		
	Baseline	2,000	5,000	2,000	5,000	2,000	5,000	Baseline	2,000	5,000
BART S	0.13	-14%	-29%	-13%	-32%	-12%	-22%	0.45	0.10	0.21
CF	0.14	-10%	-21%	-13%	-31%	-7%	-13%	0.42	0.09	0.20
CF LC	0.14	-10%	-21%	-13%	-31%	-7%	-14%	0.41	0.10	0.20
RF MOM DR	0.15	-11%	-24%	-8%	-20%	-17%	-33%	0.38	0.10	0.23

LASSO R	0.15	-13%	-28%	-10%	-26%	-20%	-39%	0.29	0.13	0.29
LASSO MCM EA	0.15	-13%	-28%	-10%	-26%	-20%	-39%	0.29	0.14	0.29
LASSO MOM DR	0.15	-13%	-28%	-10%	-26%	-20%	-39%	0.28	0.14	0.29
ATE	**0.16**	**0%**	**-4%**	**1%**	**-1%**	**-30%**	**-54%**	**NA**	**NA**	**NA**
LASSO T	0.16	-12%	-26%	-7%	-17%	-21%	-40%	0.35	0.11	0.21
RF T	0.17	-5%	-13%	-11%	-24%	-5%	-11%	0.41	0.07	0.18
LASSO MCM	0.18	-10%	-23%	-4%	-12%	-23%	-44%	0.15	0.07	0.22
LASSO MOM IPW	0.18	-11%	-23%	-4%	-12%	-23%	-45%	0.15	0.07	0.22
LASSO T INT	0.18	-11%	-23%	-9%	-21%	-13%	-27%	0.32	0.12	0.23
RF MOM IPW	0.18	-9%	-20%	-2%	-7%	-18%	-37%	0.25	0.07	0.18
BART T	0.21	-14%	-29%	-14%	-27%	-14%	-31%	0.41	0.08	0.18
OLS S	0.23	-24%	-45%	-9%	-16%	-29%	-56%	0.38	0.10	0.20
CDML	0.27	-29%	-53%	-6%	-17%	-42%	-72%	0.26	0.14	0.32
OLS S INT	0.48	-19%	-41%	-12%	-27%	-20%	-43%	0.17	0.08	0.21
OVERALL	0.19	-13%	-28%	-9%	-21%	-19%	-39%	0.30	0.09	0.21

Table 5 *Estimated Percent Change in Average RMSE, Bias, and SE, and Change in Spearman's ρ, as Experiment Sample Size Increases, by ML Method*

Note. Baseline = 1,000 observations in the hypothetical experiment. Results are from a multilevel regression model with random effects for the ML method. Values represent the average percentage change in average RMSE, Bias, SE, and increase in Spearman's ρ when increasing experiment sample size from 1,000 to 2,000 or 5,000 observations. Estimates are based on performance metrics for each ML method across around 300+ simulation runs across scenarios and queens. Rows are sorted by baseline RMSE.

4.4.3 Impact of Treatment Effect Heterogeneity Magnitude

When true impact heterogeneity is large, bias can become a more significant source of error, potentially favoring low-bias methods and affecting their relative performance. To explore this, for a subset of scenarios using the continuous outcome from CA, we increased the magnitude of impact heterogeneity from 0.2σ to 0.4σ. We view 0.4σ as substantial variation, the upper limit of what we might see in practice—though we acknowledge this is a question that should be answered empirically.

	Baseline	Percent Increase in			Spearman's ρ	
ML Method	RMSE	RMSE	BIAS	SE	Baseline	Increase
BART S	0.15	68%	86%	33%	0.46	0.08
CF LC	0.15	74%	90%	20%	0.46	0.11
CF	0.15	74%	90%	20%	0.47	0.11
RF MOM DR	0.16	67%	78%	29%	0.43	0.13
LASSO MOM DR	0.16	75%	88%	28%	0.34	0.10

LASSO MCM EA	0.16	75%	88%	29%	0.34	0.10
LASSO R	0.16	76%	88%	30%	0.35	0.10
LASSO T	0.18	66%	104%	5%	0.40	0.04
RF T	0.19	43%	89%	4%	0.46	0.11
ATE	**0.19**	**94%**	**100%**	**2%**	**NA**	**NA**
LASSO MCM	0.21	64%	93%	14%	0.12	0.09
LASSO MOM IPW	0.21	65%	93%	16%	0.12	0.10
LASSO T INT	0.22	39%	99%	0%	0.38	0.09
RF MOM IPW	0.22	57%	93%	10%	0.21	0.13
OLS S	0.23	38%	112%	1%	0.38	0.07
CDML	0.27	39%	83%	14%	0.36	0.10
BART T	0.28	13%	86%	1%	0.40	0.15
OLS S INT	0.55	9%	86%	0%	0.21	0.08
OVERALL	0.20	54%	94%	10%	0.37	0.08

Table 6 *Changes in Average Performance of ML Methods When Treatment Variation Increases From 0.2σ to 0.4σ, by ML Method*

*Note. Baseline =0.2σ magnitude for impact heterogeneity. For average RMSE, Bias, and Standard Error, the table reports the percent increase in each measure when the true treatment effect variation increases from 0.2σ to 0.4σ. For Spearman's ρ, the table reports change in value. Results are based on averages across all queens for the continuous outcome in the CA dataset, Scenarios vary by covariate set size (small vs. large) and experiment sample size (*n= *1,000 vs* n= *5,000). Rows are sorted by baseline RMSE.*

As shown in Table 6, with larger true heterogeneity, bias becomes a more dominant component of overall RMSE. Consequently, lower-bias methods, such as the RF variants, show improved relative performance. Conversely, the ATE model, which ignores heterogeneity entirely, now performs worse than most other methods. BART S maintains strong performance, but with a smaller relative advantage over other flexible methods. Spearman's ρ increases in all cases, with BART T having the largest increase of 0.15. With more impact variation, it is easier to sort units by responsiveness.

5. Empirical Application: Applying ML Method to Real RCT Data

The simulation study demonstrates that the choice of ML method is consequential, with BART S consistently performing well across a range of scenarios. To illustrate the practical implications of these findings, we apply BART S to the average monthly earnings outcome in the Career

Academies dataset. This empirical application aims to: (a) compare heterogeneity detected by BART S to conventional subgroup analysis; (b) illustrate insights enabled by ML-based CATE estimation; and (c) assess the sensitivity of empirical findings to the choice of ML estimator.

5.1 Data and Analytical Approach

This analysis uses 1,254 observations with non-missing continuous outcome values from the CA data. We restrict the analysis to the "small" covariate set (ethnicity, gender, and baseline math and reading scores). We estimate CATEs using BART S with BART S using five-fold cross-fitting and aggregate these estimates into group average treatment effects (GATEs) for subgroups defined by the covariate set. For comparison, we also compute conventional subgroup average treatment effects. Supplement Section I provides implementation details.

5.2 Comparison to Conventional Subgroup Analysis

Table 7 presents ATEs and GATEs for key subgroups. For larger subgroups (e.g., Hispanic students), BART S estimates align closely with conventional subgroup ATEs, providing reassurance about unbiasedness. Differences are more pronounced for smaller subgroups (e.g., Asian/Native American students), reflecting greater estimation uncertainty. This correspondence provides reassurance that BART S recovers true patterns of heterogeneity (that estimates within subgroups are unbiased) while leveraging greater modeling flexibility to capture how heterogeneity varies across combinations of characteristics.

Subgroup	ATE ($)	GATE ($)	Characteristics in Impact Prediction Quintile						
			Q1	Q2	Q3	Q4	Q5	Q5- Q1	N
Race/Ethnicity									
Hispanic	130	129	74%	55%	58%	60%	34%	-39%	704
Black	126	149	12%	36%	33%	34%	27%	15%	355
White	84	78	14%	8%	4%	3%	4%	-10%	82
Asian/Native American	560	314	0%	1%	3%	2%	28%	28%	85
Ethnicity missing	304	295	0%	1%	2%	1%	7%	6%	28
Baseline Reading Score									
≥75% NPR	-19	179	2%	10%	8%	8%	9%	7%	90
50–74% NPR	297	233	0%	6%	13%	20%	30%	30%	173
25–49% NPR	122	156	7%	37%	42%	23%	16%	9%	311
≤24% NPR	233	208	0%	13%	22%	48%	40%	40%	311
Missing	24	52	91%	34%	16%	1%	5%	-85%	369
Baseline Math Score									
≥75% NPR	422	226	0%	5%	6%	7%	12%	12%	76
50–74% NPR	193	230	1%	11%	12%	14%	35%	34%	183
25–49% NPR	81	146	8%	39%	32%	23%	10%	3%	280
≤24% NPR	241	201	0%	10%	35%	54%	38%	37%	342
Missing	5	53	91%	39%	16%	2%	6%	-85%	373
Gender									
Male	230	203	18%	28%	34%	45%	79%	62%	506
Female	73	115	82%	72%	68%	55%	21%	-62%	748

Table 7 *Impact Heterogeneity Estimated by BART S and Sample Characteristics by Impact Prediction Quintile*

Note. NPR = National Percentile Rank; ATE = Average Treatment Effect; GATE = Group Average Treatment Effect. ATE is estimated as the difference in average outcomes between treatment and control units within each subgroup. GATE is calculated as the average individual treatment effect (estimated by BART S) within each subgroup. Results are for the continuous outcome of average earnings from the Career Academy study (N = 1,254).

5.3 Insights from ML-Based Heterogeneity Estimation

A key advantage of ML-based CATE estimation is the ability to uncover heterogeneity across the joint distribution of covariates, rather than being limited to marginal subgroup comparisons. To explore this, we break the sample into quintiles based on the BART S-estimated CATES and examine the sample characteristics of quintiles 1 through 5. We also examine the difference in group composition between the top and the bottom effect quintile to see which demographics are over- (or under-) represented at the top relative to the bottom of the impact distribution (see middle columns in Table 7). Notably, certain ethnic groups and baseline academic performance categories are differentially represented across the CATE distribution, and gender differences are

especially pronounced: the proportion of males is higher in higher CATE quintiles, while females are more prevalent in lower quintiles.

To provide more insight, Figure 5 visualizes the full CATE distribution by gender, revealing that males are more likely to benefit substantially from the intervention, while a notable subset of females has negative estimated CATEs, suggesting potential harm. These findings, which extend beyond what is accessible through standard subgroup analysis, underscore the value of flexible ML approaches for exploring complex treatment effect heterogeneity.

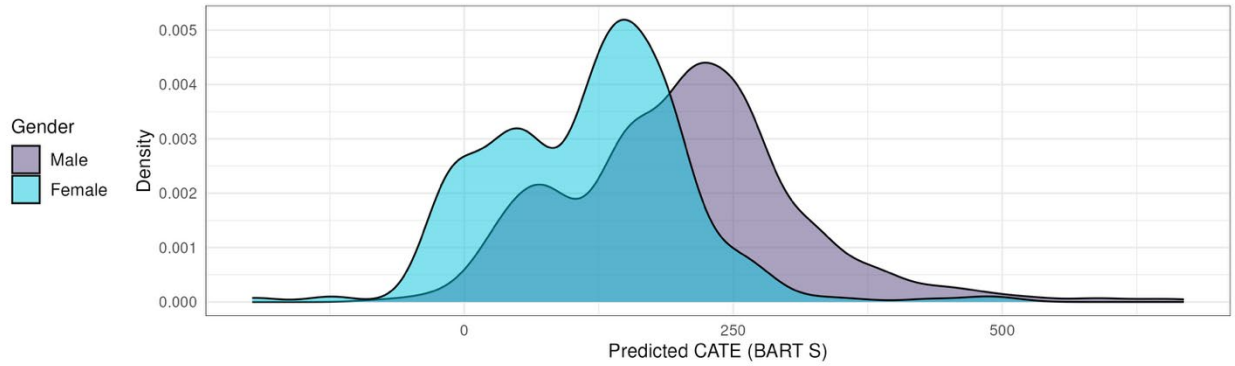

Figure 5. *Distribution of CATEs Predicted by BART S, by Gender*

Note. Results are for the continuous outcome of average earnings from the Career Academy study (N = 1,254).

5.4 Sensitivity to Choice of ML Estimator

To assess robustness, we compare BART S CATEs with those from other well-performing methods (e.g., CF, RF MOM DR, LASSO R—the better performers for scenarios with small sample size and small covariate set from Figure 2). Figure 6 shows how the differential representation of key characteristics between the top and bottom CATE quintiles (Q5 – Q1) varies across ML methods. While the extent of characteristic sorting varies by method, the qualitative patterns are consistent. In particular, all methods suggest stronger impacts for males than females. Such consistency indicates that main conclusions are not artifacts of model choice.

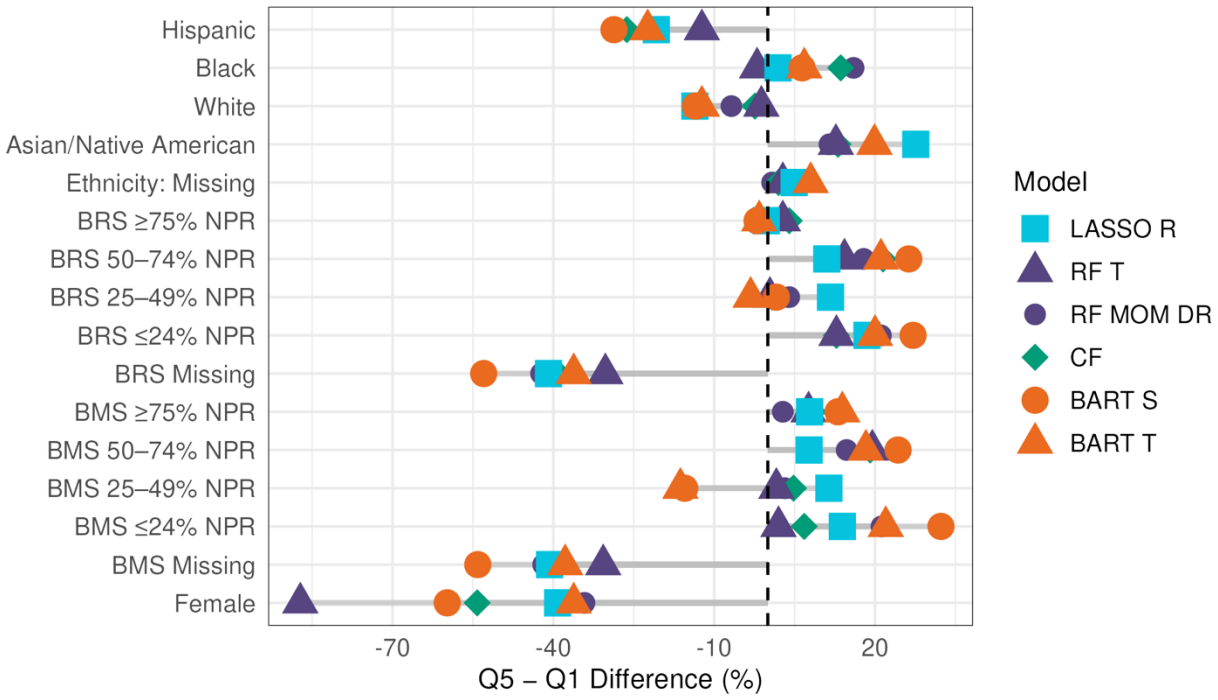

Figure 6 *Difference in Demographic Characteristics Between Top and Bottom Predicted treatment Effect Quintiles, by ML Method*

Note. BRS=Baseline Reading Score; BMS=Baseline Math Score. NPR = National Percentile Rank. Results are for the continuous outcome of average earnings from the Career Academy study (N = 1,254). Each shape corresponds to a different ML model. The x-axis displays the difference in the proportion of students with that characteristic between the top and bottom CATE quintiles.

6. Discussion

The proliferation of ML methods for estimating heterogeneous treatment effects offers potentially important new tools for researchers but also presents the challenge of selecting an appropriate approach. This study conducted a comprehensive simulation study of ML methods for detecting impact heterogeneity in the context of RCTs. By systematically comparing 18 ML methods across a wide range of data-generating processes, sample sizes, covariate complexities, and impact heterogeneity magnitudes, we provide practical guidance to applied educational researchers on which methods are reliable and what they can realistically achieve.

6.1 Key Findings

Our simulation study yielded five key findings. First, **the choice of ML method matters substantially.** We found wide variation in performance, with some popular methods performing worse, in many scenarios, than simply using the overall ATE as the prediction for every individual. Second, **BART S emerged as the most reliable method on average across 300+ scenarios.** Its superior performance was driven primarily by its stability (low standard error), which resulted in the lowest average RMSE. Causal Forests and several R-learner and modified-outcome variants also performed consistently well. Third, **none of these methods achieve high prediction accuracy.** Even BART S only yielded a Spearman's rank correlation of 0.55. Fourth, despite this, **ML methods can provide some guidance for prioritizing individuals for interventions that are useful for policymakers**. Fifth, **performance depends critically on study characteristics including sample size, the number of potential moderators, and the magnitude of true heterogeneity.** Larger sample sizes improve the performance of all methods, particularly by reducing variance. Conversely, adding extraneous covariates increased RMSE for most methods, with regularized learners (e.g., LASSO variants) least affected. When the magnitude of treatment effect heterogeneity increases, the overall ability to identify those with the biggest impacts improves and bias becomes a more dominant component of error, advantaging lower-bias methods such as RF variants.

Our findings align with and extend previous simulation studies in this domain. Similar to these earlier studies (e.g., Wendling et al., 2018; Künzel et al., 2019; Knaus et al., 2021), we found that no single method dominates across all scenarios. Unlike earlier studies that relied on specific functional forms (Knaus et al., 2021) or single-model-generated effects (Wendling et al.,

2018), our *multiple-queens* approach enhanced generalizability by testing method robustness against diverse heterogeneity structures.

The empirical application to the Career Academies dataset further showed that ML methods can provide insights not easily detected through conventional subgroup analysis alone. Encouragingly, these patterns of heterogeneity were generally robust across the top-performing methods.

6.2 Limitations

We note several limitations that should be considered when interpreting our findings. First, our simulation framework, while more comprehensive than previous efforts, cannot capture all possible patterns of heterogeneity that might occur in practice. For example, our focus on two educational datasets, while providing realistic contexts for simulation, may limit generalizability to other educational interventions or different outcome types. Replication across broader contexts would strengthen confidence in our findings.

Second, we assessed methods using default implementations rather than custom tuning. This reflects typical applied research practice but may understate the performance of methods that benefit from hyperparameter optimization. Our evaluation also excludes some newer approaches and methods lacking stable R implementations. We prioritized approaches feasible within the resource constraints of applied research, but future extensions of our framework could incorporate additional methods as software matures, potentially altering comparative performance.

Third, we focused on CATE performance metrics rather than inference or group average treatment effect (GATE), leaving questions about uncertainty estimation and statistical properties of aggregated effects for future research.

Finally, while we cannot share the original RCT data due to confidentiality, we provide synthetic datasets and full simulation code to enable replication and application of the multiple-queens framework.

6.3 Practical Implications for Educational Researchers

For educational researchers considering ML approaches to analyze treatment effect heterogeneity, our findings suggest several practical recommendations:

- **Start with a reliable method**: Our simulation results consistently identify BART S as the most reliable method across diverse conditions, making it a reasonable default choice for educational researchers, while CF and LASSO R form a consistent second tier.
- **Consider study characteristics:** Method selection should account for sample size, number of potential moderators, and expected effect magnitude. Regularized methods may be preferable with limited samples and many covariates, while more flexible approaches may be appropriate with larger samples (e.g., CDML).
- **Be realistic about predictive power:** The best-performing method achieved a Spearman's ρ of 0.5 on average, meaning that about 60% of individuals in the top predicted quintile of responsiveness are truly in the top quintile. This level of precision may be informative for exploring subgroup patterns or broadly prioritizing participants but may not be sufficient for high-stakes individual targeting.
- **Conduct sensitivity analyses:** Because method choice matters, we recommend comparing results across multiple well-performing methods. Consistency of findings across conceptually distinct approaches provides greater confidence in the substantive conclusions.

Endnotes

[1] This categorization is largely for expositional purposes. The differences between the categories are not rigid. For instance, the modified ML models can be seen as special cases of the meta-learners. Note that the selection is not exhaustive, given the rapidly evolving landscape of causal inference methods and practical limitations imposed by computational resources and the availability of off-the-shelf software implementations. We were unable to include all variants of many methods considered due to computational constraints or instability; see Supplement Section A on what methods we did not include and why.

[2] These studies were selected based on the following criteria: (a) The studies found moderate to large impacts for the full sample of participants on at least one outcome. (b) Due to the nature of the intervention and logic model, it is reasonable to assume that impacts could vary across individuals with different baseline characteristics. (c) All of the interventions tested in the studies were implemented with high fidelity. (d) The studies have moderate to large sample sizes. (e) The available datasets include many baseline covariates. Supplementary material lists all available covariates for each study. Detailed description of the studies and their findings can be found in Kemple and Willner (2008) and Miller and Weiss (2021).

[3] We averaged in the squared space to preserve the bias-variance decomposition of MSE, expressed as ($MSE = Bias^2 + Variance$), and subsequently take the square root to return to the scale of the outcome. The average bias approach avoids over inflation of near-zero biases due to Monte Carlo simulation uncertainty.

[4] Complete simulation results for all methods are provided in a master results file in supplementary material. In addition, we also provide the simulation code and framework so the overall simulation approach we promote in this paper can be carried to new datasets and contexts. While for data security concerns, we cannot provide the raw data used in this work, we provide a purely synthetic dataset and demo code for running the simulations.

[5] We selected this outcome essentially arbitrarily—the results are similar in character to the other continuous covariate. The results for all scenarios and outcomes are in the supplement Part F and are included in our regression analysis below.

References

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50. https://doi.org/10.1257/jep.28.2.29

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science*, *14*(2), 179–188. https://doi.org/10.1007/s11121-011-0222-z

Caron, A., Baio, G., & Manolopoulou, I. (2022). Estimating individual treatment effects using Bayesian additive regression trees and non-parametric methods for longitudinal data. *Statistics in Medicine*, *41*(4), 692–712. https://doi.org/10.1002/sim.9238

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*(4), 1199–1209. https://doi.org/10.1111/biom.12676

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1093/ectj/utx023

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2022). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *Journal of Econometrics*, *229*(1), 393–427. https://economics.mit.edu/sites/default/files/2022-08/2020.12%20HeterogTE-RE-v26.pdf

Ding, P. (2024). *A first course in causal inference*. Chapman & Hall/CRC Press.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317. https://doi.org/10.1080/01621459.2018.1438926

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43–68. https://doi.org/10.1214/17-STS613

Gao, X., Wang, Y., Zhou, H., & Li, J. (2020). Random forest-based causal inference for heterogeneous treatment effect estimation with observational data. *Journal of Machine Learning Research*, *21*(37), 1–37. https://jmlr.org/papers/v21/19-058.html

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, *13*(1), 163–182. https://doi.org/10.1214/16-BA1044

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*(1), 1–21. https://doi.org/10.1016/j.jeconom.2013.02.006

Ishwaran, H., & Malley, J. D. (2014). Synthetic forests and the extraction of heterogeneous treatment effects. *arXiv*. https://doi.org/10.48550/arXiv.1408.1672

Jacob, R. T., Zhu, P., Somers, M.-A., & Bloom, H. S. (2019). Practical guidance for estimating subgroup effects in experimental evaluations of social programs: A systematic review of methods and recommendations for future research. *Evaluation Review*, *43*(5), 283–325. https://doi.org/10.1177/0193841X19878825

Kemple, J. J., & Willner, C. J. (2008). *Career Academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthood*. MDRC. https://www.mdrc.org/publication/career-academies-long-term-impacts-work-education-and-transitions-adulthood

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627. https://doi.org/10.1093/ectj/utac015

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence with the causal random forest and double machine learning approaches. *Statistics in Medicine*, *40*(21), 4597–4619. https://doi.org/10.1002/sim.9071

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Li, M., Jiang, Z., & Athey, S. (2022). Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *arXiv*. https://arxiv.org/abs/2203.14511

Lu, M., Sadiq, S., Feaster, D. J., & Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, *27*(1), 209–219. https://doi.org/10.1080/10618600.2017.1356325

Miller, C., & Weiss, M. J. (2021). *Accelerated Study in Associate Programs: Two-year impact report on degree completion and transfer rates at Bronx Community College*. MDRC. https://www.mdrc.org/publication/accelerated-study-associate-programs-asap

Miratrix, L. (2023, June 5). Using copulas for making calibrated data generating processes (DGPs) for simulation. *CARES Blog*. https://cares-blog.gse.harvard.edu/post/copulas-for-simulation/

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. (Original work published 1923)

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Polley, E. C., Rose, S., & van der Laan, M. J. (2011). Super Learning. In M. J. van der Laan & S. Rose (Eds.), *Targeted learning: Causal inference for observational and experimental data* (pp. 43–66). Springer. https://doi.org/10.1007/978-1-4419-9782-1_3

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high-dimensional data. *Statistics in Medicine*, *37*(23), 3309–3324. https://doi.org/10.1002/sim.7820

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, *39*(2), 1180–1210. https://doi.org/10.1214/10-AOS864

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954. https://doi.org/10.2307/1912705

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, *23*(8), 2379–2412. https://doi.org/10.1080/03610929408831393

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., & Fresques, H. (2015). *Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students*. MDRC. https://www.mdrc.org/sites/default/files/doubling_graduation_rates_fr.pdf

Thal, D. R. C., & Finucane, M. M. (2023). Causal methods madness: Lessons learned from the 2022 ACIC competition to estimate health policy impacts. *Observational Studies*, *9*(3), 3–27. https://doi.org/10.1353/obs.2023.0023

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Tipton, E., Hallberg, K., & McCaffrey, D. F. (2022). The limits of covariate-driven heterogeneity estimation in randomized trials. *Journal of Educational and Behavioral Statistics*, *47*(2), 135–167. https://doi.org/10.3102/10769986211066080

Tipton, E., & Mamakos, M. (2025). Designing randomized experiments to predict unit-specific treatment effects. *Statistics and Public Policy*, *12*(1), 1–35. https://doi.org/10.1080/2330443X.2024.2306202

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, *37*(23), 3309–3324. https://doi.org/10.1002/sim.7820

Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*(4), 1010–1018. https://doi.org/10.1111/j.1541-0420.2012.01777.x

Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2023). Efficiency augmentation for covariate-adjusted estimators. *Biometrics*, *79*(2), 1258–1269. https://doi.org/10.1111/biom.13618

Zhou, Z., Zhang, S., & Tu, W. (2022). R-learner: A machine learning approach to estimate individualized treatment rules from observational data. *Statistics in Medicine*, *41*(8), 1446–1464. https://doi.org/10.1002/sim.9280

Supplementary Materials

Table of Contents

Section A. Evaluated ML Methods and Their Selection

In reviewing the literature, we identified a wide range of methods with the potential for estimating heterogeneous treatment effects. For each method found in the key papers we surveyed, we sought to locate a working R package or reliable code implementation. For most approaches—namely, LASSO R, LASSO MOM IPW, RF MOM IPW, LASSO MOM DR, RF MOM DR, LASSO MCM, and LASSO MCM EA—we adapted code from the publicly available CATE estimation methods repository (Knaus, Lechner, & Strittmatter, 2023) to streamline workflows for our simulation.

Our primary drivers in method selection were (a) their prevalence in recent literature and (b) practical considerations about the availability and robustness of implementations. During initial simulation efforts, we attempted to include as many of the identified methods as possible, but not all methods proved workable. Some methods were excluded due to the absence of reliable or computationally feasible implementations. For example, we were unable to find a version of adaptive LASSO that was both robust and efficient for our setting. Other packages presented reproducibility issues: certain models, such as XBART and XGBoost, would override the simulation's random seed, undermining the integrity of repeated trials. In some cases, packages ran too slowly to be feasible for inclusion in all scenarios; for instance, the Super Learner T method was only deployed in a subset of simulations due to its substantial runtime. To keep the computational burden manageable, we further limited the inclusion of method variants, which is why not all possible combinations appear in our results. Table A1 lists the selected ML methods and notes the R packages we used to implement them.

Method	Reference	Our Implementation
ATE	-	**ATE**
INFEASIBLE	-	LASSO INF, RF INF
S-LEARNER	Foster et al. (2011); Hill (2011)	**OLS S**, **OLS S INT,** BART S,
T-LEARNER	Künzel et al. (2019)	LASSO T, LASSO T INT, **RF T**, **BART T**
ENSEMBLE	Polley et al. (2011)	**SL T *(OLS, LASSO, RF, CIF)***
R-LEARNER	Nie and Wager (2021)	**LASSO R**
MOM	Zhang et al. (2012)	LASSO MOM IPW, **RF MOM IPW**, LASSO MOM DR, RF MOM DR
MCM	Tian et al. (2014)	LASSO MCM, LASSO MCM EA
CF	Wager and Athey (2018)	**CF**, CF LC
DML	Knaus (2022)	**CDML *(OLS, RF, RF)***

Table A1 *Selected Machine Learning Methods*

Note. We used the following R packages to implement these methods: dbarts for BART, causalDML for CDML, grf for CF and RF, glmnet for LASSO, stats for OLS, and SuperLearner for the ENSEMBLE method. Methods designated as queens are bolded. Italicized models in parentheses indicate base learners used within more complex methods like ENSEMBLE. CIF stands for Conditional Inference Forest.

Overall, we found significant variation in how R packages handled the repeated and intensive use required for our simulation study. Occasionally, even rare numerical errors—if they caused R to crash despite our error trapping—made continued use of some packages impractical without extensive rewriting of our simulation framework. This experience highlighted the broader need for making packages more accessible and robust for applied researchers.

The following methods were excluded from our simulation study for the reasons detailed below:

- **Super Learner:** This method was incorporated in some simulation runs but was prohibitively computationally expensive for wider use. For instance, running 100 iterations with n=5,000 exceeded the limits of our available infrastructure.
- **LASSO S:** Implementing LASSO S required manually generating all treatment-by-covariate interactions and running LASSO on the expanded covariate set. However, in

practice, LASSO S frequently pruned all interaction terms, effectively reducing the model to the ATE. Given this outcome, we deemed it impractical for use in our study.

- **Adaptive LASSO:** Rakovic, Ronde, and Verweij (2014) proposed an alternative to LASSO S in which treatment-by-covariate interactions are regularized separately, helping to prevent main effects from overshadowing true impact heterogeneity. Unfortunately, the available package for this approach performed a computationally intensive, two-dimensional grid search for tuning parameters using cross-validation, making it too slow for our use. It also failed to handle certain covariate settings, so we were unable to obtain consistently reliable results across scenarios. We believe future research would benefit from including this method, as it allows modeling of complex heterogeneity, including three-way interactions.
- **Additional RF variants:** While LASSO methods were generally computationally efficient, we could not include all the variants of Random Forests due to their computational demands. As a result, we prioritized those variants most prominent and best-performing in the literature.
- **XGBoost:** Some studies advocate the use of XGBoost for causal effect estimation. However, in our experience, the R xgboost package sporadically produced implausible results, especially in scenarios with no treatment effect variation (the "ATE queen"), in which some units were assigned unreasonably large, predicted impacts. Concerned that this reflected underlying package issues or misuse, we excluded XGBoost from consideration to avoid reporting misleading results.

We also considered additional approaches, including X-learners (Künzel et al., 2019) and neural network–based CATE estimators (e.g., TARNet, DragonNet; Hartford et al., 2017; Curth

& van der Schaar, 2021). However, at the time of our study, stable and computationally feasible R implementations were not available. Pilot tests revealed convergence failures and excessive runtimes. While these approaches are promising, especially given recent advances in deep learning for causal inference, they could not be reliably incorporated into our framework.

This process of adaptation and vetting underscores the practical challenges that applied researchers face when implementing state-of-the-art methods and points to the importance of continued software development for robust, user-friendly causal inference tools.

Computational complexity

In terms of running time, 9 of the 18 non-benchmark methods, including LASSO variants, OLS S, and RF T, finished in under one minute per iteration, making them efficient for large-scale or rapid analyses. More complex methods, such as those involving ensemble learning (BART variants, CF variants), cross-fitting (CDML), or extensive interaction modeling (OLS S INT), required moderately longer runtimes, ranging from 1 to 8 minutes. The most computationally intensive approach, SL T, took up to 12 minutes per iteration for an experiment size of 1,000; it was substantially longer as the experiment size grew. These results demonstrate a trade-off between model complexity and computational speed: while simpler models offer near-instantaneous execution and stable predictions, more flexible or robust methods may incur longer runtimes.

Method	Running time per iteration (minutes)
ATE	<1
LASSO INF	<1
RF INF	<1
OLS S	<1
LASSO T	<1
LASSO T INT	<1
LASSO MOM IPW	<1
LASSO MOM DR	<1
LASSO MCM	<1
LASSO MCM EA	<1
LASSO R	<1
RF T	<1
CF	1
BART T	2
CF LC	4
CDML	4
BART S	5
RF MOM DR	5
RF MOM IPW	5
OLS S INT	8
SL T	12

Table A2. *Running Time Per Iteration by ML Method*
Note. Numbers represent the amount of time it takes to run a given ML method once on a training set with 1,000 units and with the large covariate set for the continuous outcome from the CA dataset, averaged across 100 iterations.

Section B. Further Details of the Data Generation Process

In this section, we describe the data-generating process (DGP) in greater detail. Our goal is to generate synthetic data that closely mirrors the real data from our two RCT studies. We aim to create empirically plausible datasets to examine how machine learning methods—typically designed for large observational studies—perform in real-world randomized evaluations with smaller sample sizes.

For each study, we initiated the process by generating a dataset of 100,000 units, from which we subsequently subsampled data for the actual simulation. The DGP encompasses three steps:

(1) generating covariates, (2) generating untreated potential outcomes, and (3) generating "true" treatment impacts. We provide details of these steps below.

Generate the Covariates (X_i)

We generated covariates using the "synthpop" package in R (Nowok et al., 2016). This package is designed to generate synthetic versions of original datasets that have identical marginal distributions and similar conditional distributions as the original data. It works by synthesizing variables sequentially, with each variable generated based on a conditional distribution that depends on previously synthesized variables. The first variable is synthesized by randomly sampling from its observed values, while subsequent variables are modeled based on earlier ones. This method preserves the relationships and dependencies between variables, ensuring that the covariance structure of the original data is maintained. Importantly, the synthpop approach prevents duplicates of individual covariate profiles, which would occur with a bootstrapping method.

Generate the Untreated Potential Outcomes ($Y_i(0)$)

To generate the untreated potential outcomes ($Y_i(0)$), we employed a predictive modeling approach combined with a copula method to replicate the covariate-outcome relationship observed in the actual data while preserving the empirical marginal distribution of the outcome. We first estimated the outcome variable ($Y_i(0)$) from the control group of the source data using a random forest model with the large covariate set to capture the underlying relationships that influence the outcome. Once the model was trained, we predicted untreated outcomes, denoted as $\hat{Y}_i(0)$, for our full set of synthetically generated covariates.

To ensure the distribution of the generated untreated potential outcomes aligns with the empirical distribution of the real data, we transformed the predicted untreated outcomes (the

$\hat{Y}_i(0)$) using a copula. A copula is a statistical approach that models the relationships between different random variables. In this case, we mapped the predicted and observed outcomes in the original data to a bivariate normal space through the copula and then used this to draw untreated outcomes (the $Y_i(0)$) conditional on the predicted outcomes (the $\hat{Y}_i(0)$) for the synthetic data. This process effectively incorporated residuals into the predicted outcomes while preserving the shape and bounds of the actual outcome distribution. The following steps outline this process:

- Rescale the marginal distribution of $\hat{Y}_i(0)$ and $Y_i(0)$ so that they are standard normal, giving *z-value* pairs $\left(z\left(\hat{Y}_i(0)\right), z(Y_i(0))\right)$. Note that these z-values are not traditional z-scores but rather are non-linear (but monotonic) transforms of the original data. We calculated the z-value of each empirical observation by first calculating its percentile, and then mapping that percentile to a z-score corresponding to that percentile in a standard normal distribution:

$$z_i = g(X_i) := \Phi^{-1}\left(\frac{rank(X_i)}{K} - \frac{1}{2K}\right), \qquad (1)$$

  where $\Phi^{-1}(x)$ is the inverse of the cumulative normal distribution function and K is the number of observations in the original data. We call this z-value transformation function $g(x)$.

- Next calculate the resulting correlation coefficient (ρ) of the transformed $\hat{Y}_i$ and Y_i within the original empirical data. The correlation ρ is a measure of how predictive the covariates are of the control-side potential outcome.

- Use the calculated ρ to define a copula: a bivariate normal distribution F ~ $N((0,0), \Sigma)$, where $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

- Generate observed outcomes for synthetic units. For each new synthetic unit *i*, we generated an observed outcome by sampling a $z(Y_i(0))$, conditional on $z(\hat{Y}_i) = a$, using

$$z(Y_i(0)) \mid z(\hat{Y}_i) = a) \sim N(\rho a, 1 - \rho^2).$$

  In other words, we drew a normal variable for each unit *i* centered at $\rho z(\hat{Y}_i)$ with variance $1 - \rho^2$. This distribution is simply the conditional distribution of one side of a bivariate normal given a value for the other side.

- Transform back to the original scale. We converted the new z-values, $z(Y_i(0))$, back into the original scale of the observed outcomes (Y_i) by inverting $g(x)$ as defined in Equation (1):

$$Y_i(0) = g^{-1}(z(Y_i(0)))$$

  This step gave us new untreated outcomes, $Y_i(0)$, for all generated units in the synthetic dataset. These new outcomes maintained the same overall distribution as the original observed outcomes.

Through this copula approach, we have generated new outcomes ($Y_i(0)$) that reflected the relationships between the predicted and observed values from the original data while still incorporating a level of uncertainty in $Y_i(0)$ given the X_i. This means that the generated untreated outcomes were realistic and consistent with the patterns in the original data, while allowing for some variation. More general discussions of this approach and R code for its implementation can be found on the CARES Blog (Miratrix, 2025).

We finally generated $Y_i(1)$ for each unit by first using the designated queen to predict τ_i, and then calculating $Y_i(1) = Y_i(0) + \tau_i$.

Section C. List of Covariates in the Small, Medium, and Large Sets

To make this study's findings as relevant to reality as possible, we generated data that captured the contexts of the two selected RCT studies, Career Academies (CA) and Accelerated Studies in Associated Programs (ASAP).

- *Career Academies (CA)* is a widely used high school reform initiative that uses small learning communities and partnerships with local employers to provide work-based learning opportunities. Kemple & Willner (2008) found that students randomly assigned to Career Academies stayed in school longer and had 11% higher earnings over 8 years after scheduled high school graduation. The study data contained 1,764 observations with 29 individual-level covariates that could be used for subgroup analysis of impacts. The simulation focused on two long-term labor-market outcomes: a continuous variable that measured the average monthly earnings during the 8 years after a participant's scheduled high school graduation and a dichotomous variable that indicated whether a participant was employed full-time for longer than the sample median full-time employment duration in the 5-8 years after the scheduled high school graduation.
- *The Accelerated Studies in Associated Programs (ASAP)* aimed to increase community college graduation rates with comprehensive support like advising, tuition assistance, and requiring full-time enrollment and regular meetings with advisors. Randomized field trials were conducted at CUNY community colleges (Scrivner et al., 2015) and in Ohio (Miller et al., 2021). Both studies found that ASAP increased 3-year graduation rates by around 16 percentage points. The CUNY trial had 896 students across 3 colleges, while the Ohio trial had 1,501 students across 4 locations at 3 colleges. Both collected similar baseline data that could be pooled for the analysis of differential impacts across student

subgroups. The simulation focused on two outcome measures: a continuous measure of the total number of credits earned and a dichotomous indicator of whether a student earned any college degree.

Table C1 provides the list of variables included in the small-, medium-, and large-covariate sets.

Study	Small Covariate Set	Medium Covariate Set	Large Covariate Set
Career Academies (CA)	Ethnicity Gender Age at baseline Attendance rate at baseline 8th Grade Math state test score (baseline) 8th Grade Reading state test score (baseline)	**Small Set +** Has sibling that dropped out of High School Credit earned in baseline year GPA in baseline year LEP status at baseline Overage for grade level at baseline Household structure at baseline Transferred for 2+ times prior to study Family receives welfare or food stamp at baseline	**Medium set +** Father education level Mother education level Hours per day watching TV Hours per day unsupervised Hours per week on homework Feeling unsafe at school Parents work for pay Neither parent has diploma Sent to office at baseline Single parent household Postsecondary expectations at baseline # of times family moved in past 2 years Student ever worked for pay
Accelerated Studies in Associate Program (ASAP-CUNY+Ohio)	Race/Ethnicity Age Gender Employed Has children Parents pay more than 50% of educational expenses	**Small Set +** First in immediate family to attend college Marital status and living situation Diplomas and degrees earned	**Medium Set +** Date earned high school diploma Highest grade completed Speak a language other than English at home Highest degree planned Number of children Age of youngest child Diplomas and degrees earned

Table C1 *Variables Included in Small-, Medium-, and Large-Covariate Sets*

Section D. Unit-Level Performance Metrics

The primary performance metric for our analysis is the ability of each machine learning model to accurately predict the CATE function. To facilitate this evaluation, we established a static test set that mirrors the distribution of our target population and calculated the true CATE for each entry.

In our simulation, we repeatedly generated new sets of synthetic experimental data and then used the fitted models to predict treatment impacts for each element in the set-aside static test set.

Maintaining a static test set provided significant advantages. It allowed us to assess predictive performance locally and explore how well different ML methods estimated the CATE across various regions of the covariate distribution. Specifically, for a test point characterized by covariates ($X_i = x$), we could evaluate the performance of each model at that precise point.

We computed unit-level performance metrics, including mean squared error (MSE), bias, and standard error (SE), by averaging across iterations. Let $CATE_i = CATE(X_i)$ denote the true $CATE$ for unit i in the test set, and let $\hat{\tau}_{ijs}$ denote model $j's$ prediction of CATE_i for unit i during iteration s. The performance metrics for each unit k and model j across S iterations were defined as follows:

- $MSE_{ij} = \frac{1}{S}\sum_{s=1}^{S}(\hat{\tau}_{ijs} - CATE_i)^2$
- $Bias_{ij} = \frac{1}{S}\sum_{s=1}^{S}\hat{\tau}_{ijs} - CATE_i$
- $SE_{ij}^2 = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{\tau}_{ijs} - \frac{1}{S}\sum_{s=1}^{S}\hat{\tau}_{ijs})^2$

These unit-level performance metrics were then aggregated to the model level by averaging across all test units, as presented in Section 3.3 of the main paper.

Section E. Distributions of True Impact Generated by Selected Queens

The *"multiple-queens" approach* we used in the data generating process helps prevent unfairly advantaging methods similar to those used for generating impacts (as noted by Gao et al., 2020). By generating a diverse set of treatment effects, the multiple-queens approach ensured our

evaluation did not systematically favor any particular class of estimation methods. As shown in Figure E1, the resulting distributions of the generated treatment effects varied substantially across queens. Some generated spiky distributions (where many individuals have identical effects based on shared categorical characteristics), while others produced smoother distributions with more gradual variation. This diversity ensured our evaluation did not systematically favor any particular class of estimation methods and allowed us to evaluate method performance across a range of realistic heterogeneity scenarios rather than relying on a single arbitrary specification.

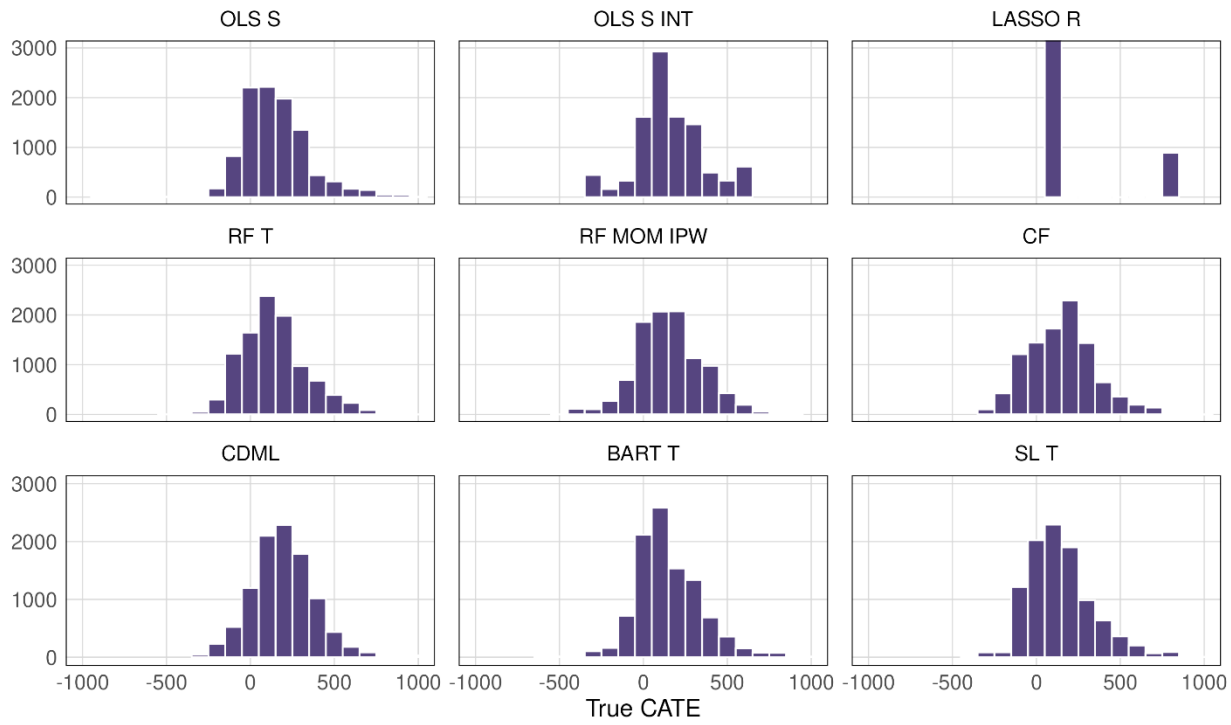

Figure E1. Distribution of Generated Treatment Effects by Queen.

Note. Based on the continuous outcome in the CA dataset across 10,000 test units. The standard deviation of the generated effects is set to 0.20σ for all queens. The figure shows how different ML methods ("queens") generate different patterns of treatment effect heterogeneity, from sparse, spike-like distributions to more continuous, tailed distributions. The x-axis and y-axis are truncated for clarity. Specifically, the x-axis is limited to the range [-1000, 1000] and the y-axis to [0, 3000] to enhance comparability across panels and reduce the influence of extreme values.

Section F. Additional Method Performance Plots by Queen and Outcome Type

Figures F1 and F2 below display the performances of those ML methods not included in Figure 3 of the main paper for the continuous and binary outcomes, respectively. As in Figure 3 of the paper, each plot represents a different combination of method (columns) and training set size (rows), with individual points in each plot representing the performance of a specific method against a specific queen (excluding the ATE queen) using a specific dataset and covariate set, averaged across two outcomes of the same type. As before, "×" represents the overall average across all scenarios within each plot. Some data points for CDML and BART T are truncated due to values outside the range displayed. Note that the reference methods (ATE, LASSO INF, and RF INF) are not included, nor are SL T (for too few scenarios run) and OLS S INT (too many out-of-range values).

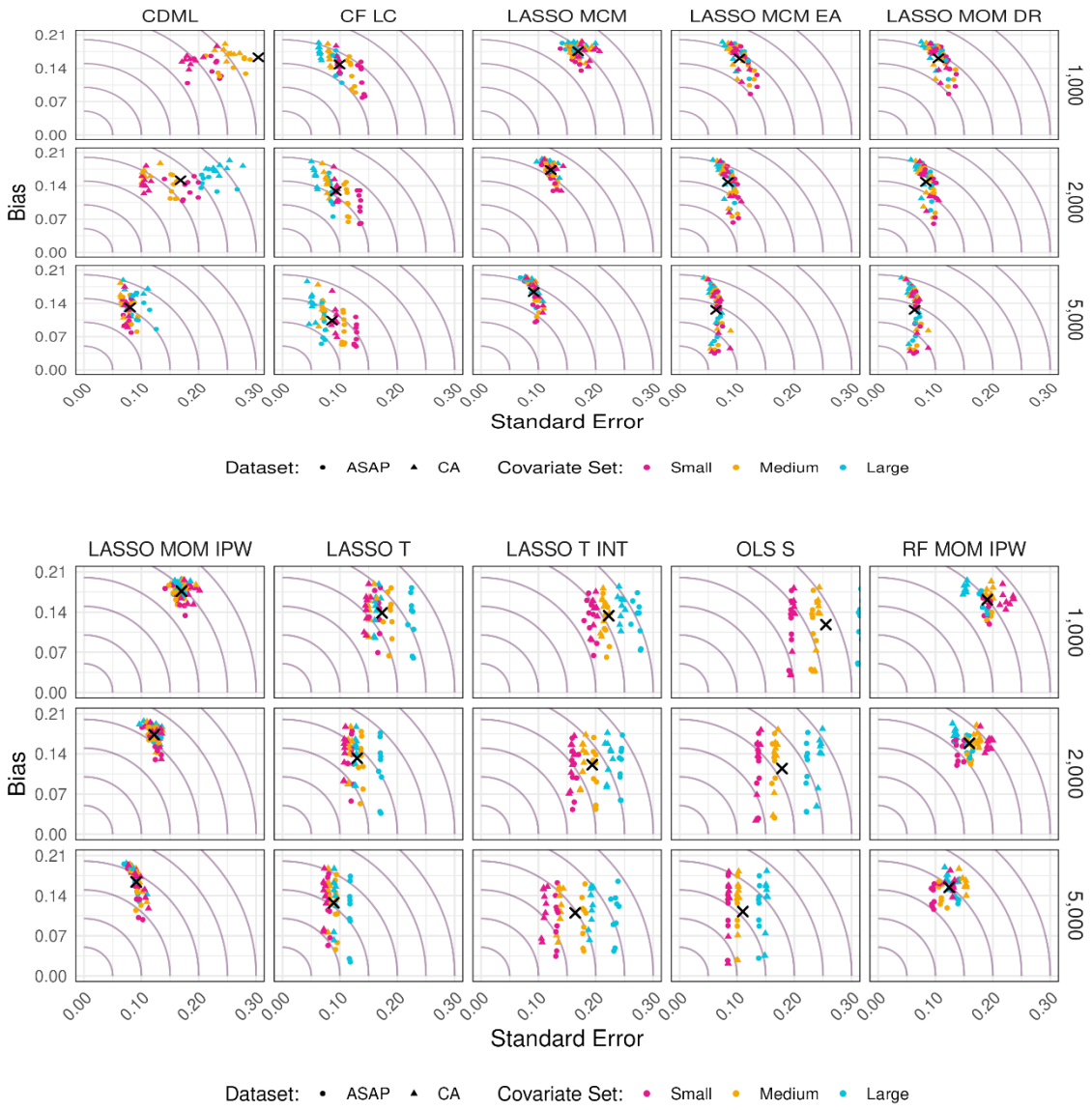

Figure F1 Performance of Additional Selected ML Methods by Queen Across Simulation Scenarios, Continuous Outcomes

Note. Each point represents the performance of a machine learning method against a specific queen (excluding the ATE queen) in each scenario, averaged across two outcomes of the same type. "×" represents the overall average across all scenarios within each plot. The x-axis is limited to a maximum of 0.21 (bias) and the y-axis to 0.3 (standard error). Some data points beyond these limits are truncated for clarity and to improve visual comparability across plots.

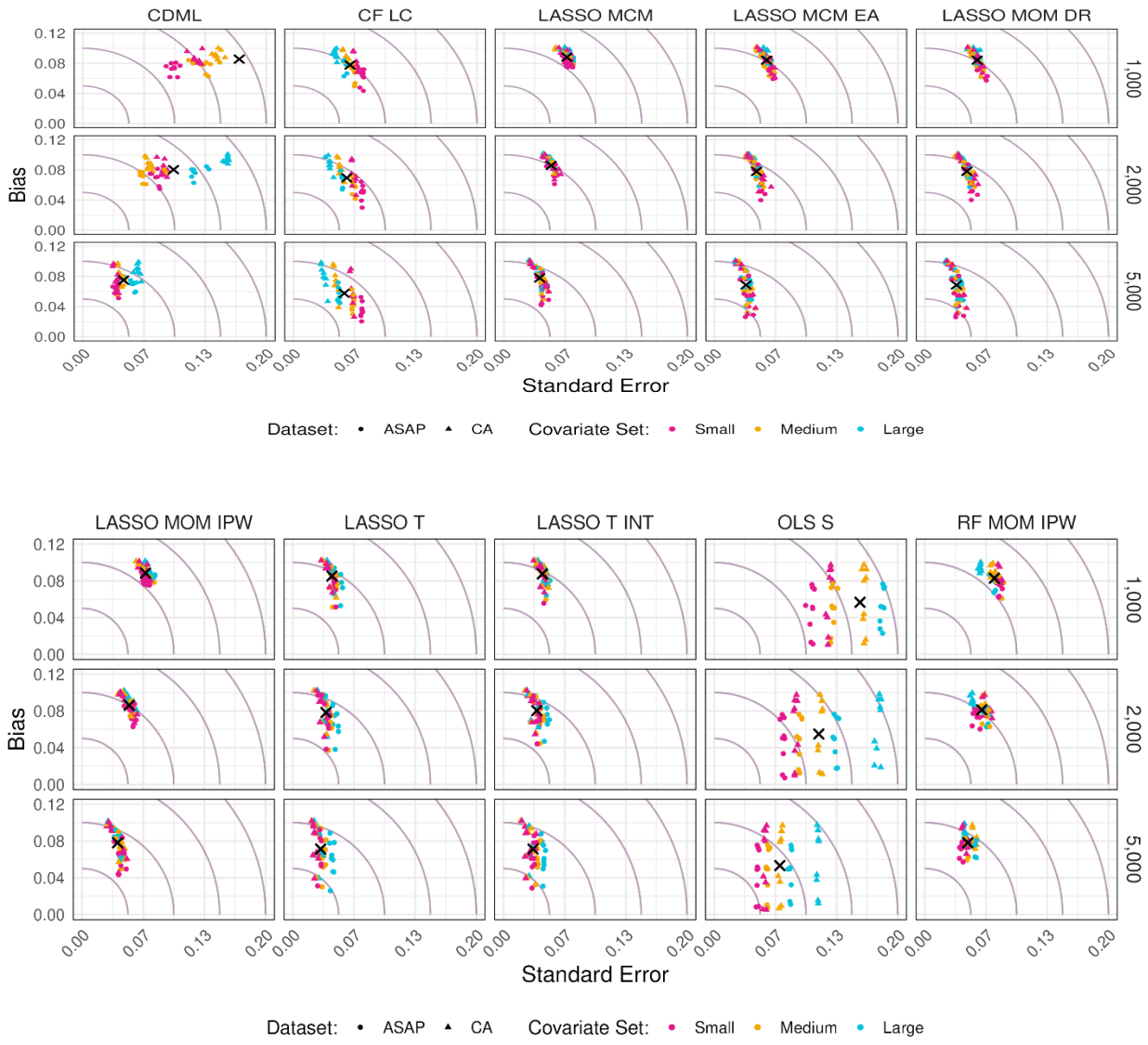

Figure F2: Performance of Additional Selected ML Methods by Queen Across Simulation Scenarios, Binary Outcomes

Note. Each point represents the performance of a machine learning method against a specific queen (excluding the ATE queen) in each scenario, averaged across two outcomes of the same type. "×" represents the overall average across all scenarios within each plot. The x-axis is limited to a maximum of 0.12 (bias) and the y-axis to 0.2 (standard error). Some data points beyond these limits are truncated for clarity and to improve visual comparability across plots.

Compare rankings of methods by outcome type

Figure F3 presents the relationship between ML method performances (as measured by average RMSE) across outcome types and datasets. The two plots on the top show the performance

correlation between continuous and binary outcomes for the Career Academies and ASAP data separately, while the bottom two plots show the performance correlation between the Career Academies and ASAP data for each outcome type separately. These plots illustrate that the relative performance ranking of methods remained similar across outcome types and data sources evaluated in the paper.

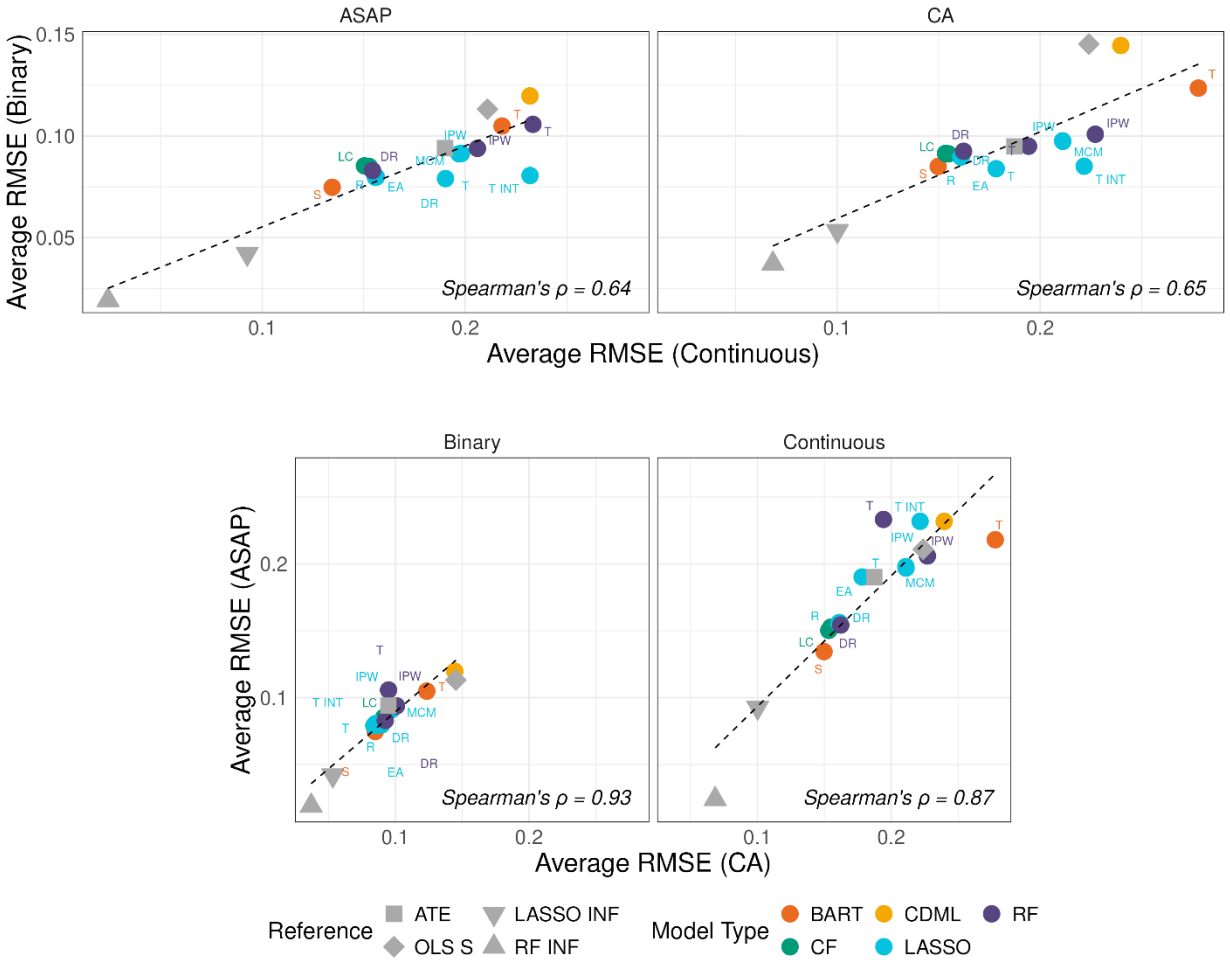

Figure F3 *Relationship of Average RMSE Across ML Methods Between Outcome Types and Data Sources*

Note. *Each point represents a method's average RMSE across scenarios for continuous (x-axis) and binary (y-axis) outcomes. Spearman's rank correlation between the two is 0.6 when excluding reference methods (gray shapes) and 0.7 when including them. This suggests a moderate positive association between method performance across outcome types. SL T is not included because it was not run for the binary outcome, and OLS S INT is excluded due to being*

an extreme outlier. A higher Spearman's ρ indicates that methods ranking well for continuous outcomes also tend to perform well for binary outcomes.

Section G. Interpretation of Spearman's ρ

Spearman's ρ is the established, standard nonparametric statistic for quantifying the strength and direction of monotonic rank relationships between two variables. It is appropriate for ordinal, non-normal, or otherwise non-linearly related data, widely used across sciences for quantifying rank agreement. In the main paper, we used Spearman's ρ to measure the rank order correlation between the actual and predicted unit treatment effects. Put simply, it captured how well the predicted treatment effects correctly order the test units by their actual treatment effects.

Figure G1 illustrates the agreement between predicted and actual individual treatment effects by displaying the distribution of predicted quintiles compared to the "true" effect quintiles. The plot is shown for BART S across three simulation scenarios with varying Spearman's ρ values to demonstrate how predictive ranking quality changes as ρ improves.

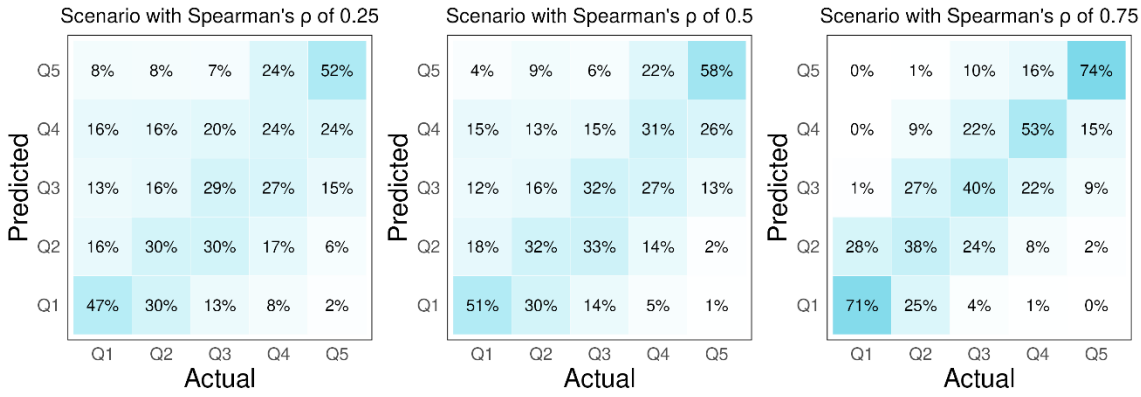

Figure G1. Illustration of Cross-Quintile Agreement Between Actual CATE and Those Predicted by BART S, For Three Scenarios with Varying Spearman's ρ Values

Note. *The number in each cell represents the percentage of individuals in a given quintile of the actual CATE distribution that are correctly predicted to be in that quintile by BART S. Perfect prediction would show 100% along the main diagonal and 0% elsewhere.* Each quintile contains 2,000 observations (based on a test set of 10,000). All three scenarios use Career Academies data with a continuous outcome and use BART S as the estimation model. The first scenario (mean Spearman's $\rho \approx 0.25$) uses the large covariate set, 1,000 units in the hypothetical experiment, and the CDML queen. The second scenario (mean Spearman's $\rho \approx 0.5$) uses the medium covariate

set, experimental sample with 5,000 units, and the CDML queen. The third scenario (mean Spearman's $\rho \approx 0.75$) uses the small covariate set, experimental sample with 1,000 units, and the RF MOM IPW queen.

When interpreting this figure, note that perfect concordance would concentrate all counts along the diagonal, indicating that predicted quintile and true quintile always match. In a simulation scenario where $\rho \approx 0.25$ (left panel), 52% of the top quintile individuals are still correctly identified as being in the top quintile. With $\rho \approx 0.5$ (middle panel), 58% of Q5 individuals are correctly placed in Q5, and the concentration along the diagonal indicates better overall rank ordering. The strongest performance ($\rho \approx 0.75$, right panel) has 74% of the Q5 individuals correctly being identified as high-impact, Q5, individuals. Most misclassifications occurred in adjacent groups rather than at extremes: for example, an individual in the true top quintile was much more likely to be predicted to be in the top or second-highest group than to be misplaced in the lowest.

This cross-quintile illustration provided a transparent, intuitive sense of what predicting "individual treatment effects" means in real applications. While methods like BART S sorted individuals better than random allocation, considerable uncertainty remained. This suggested that, in typical randomized trial settings, ML-based targeting may be informative for policy or exploratory subgroup discovery, but was insufficiently precise for high-stakes, individual-level targeting.

Section H. Assessing Performance Variation by Key Dimensions

We evaluated changes in performance metrics as either the number of covariates increased (Table 4 in the main paper) or the sample size of the hypothetical experiment increased (Table 5 in the main paper). To assess these changes, we fitted four separate models—one for each

outcome: log RMSE, log absolute bias, log SE, and Spearman's ρ—of the following form, where method *j* was evaluated against queen q in scenario s:

$$Y_{jqs} = \beta_{0\mathrm{j}} + \beta_{1\mathrm{j}}\mathrm{X}_{1s} + \beta_{2\mathrm{j}}\mathrm{X}_{2s} + v_q + w_s + r_{qs} + \varepsilon_{jqs}$$

$$\text{with } \begin{pmatrix}\beta_{0\mathrm{j}}\\ \beta_{1\mathrm{j}}\\ \beta_{2\mathrm{j}}\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}, \Sigma\right)$$

Here, Y_{jqs} denotes the outcome measure for method *j* with queen q in scenario s. X_{1s} is an indicator for either a medium covariate set or a sample size of 2,000, and X_{2s} is an indicator for either a large covariate set or a sample size of 5,000. Coefficients $\beta_{1\mathrm{j}}$ and $\beta_{2\mathrm{j}}$ represent the change in performance relative to the small covariate set or baseline sample size. The terms v_q, w_s, and r_{qs} are random effects that account for clustering in the simulation structure, allowing queens and scenarios to vary systematically in performance. Σ is a 3×3 covariance matrix for the method-level random effects.

For the log RMSE, log absolute bias, and log SE outcomes, we exponentiated the empirical Bayes estimates $\hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ to express them as multiplicative factors. We then subtracted 100 from these multiplicative factors to report the percentage change relative to the baseline group. For Spearman's ρ, $\hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ represented changes in the absolute value of ρ relative to its baseline value. Each estimate therefore reflected the typical increase or reduction in performance when moving from the small to the medium or large covariate sets or sample sizes.
Tables 4 and 5 in the main paper presented these results. Each row reported the estimated percentage point or absolute change for a given method, and the final row showed the corresponding fixed-effect estimates.

Section I. Further Notes on the Empirical Application

Implementation Details

For the empirical application presented in Section 5 of the main paper, we estimated CATE values using the code developed for our simulation, with a slight modification. In particular, we used cross-fitting to guarantee that regardless of how the methods tuned themselves internally, we would have true out-of-sample predictions of individual effects. We did this by randomly dividing our data into five chunks and fitting our designated model, in turn, to four of the chunks to predict CATE values on the left-out chunk. Over the course of fitting 5 times, we could obtain individual CATE estimates for all individuals in our dataset. We treated the ML process itself as a black box, ignoring any tuning it was doing internally under this scheme.

To generate profiles by quintile, we took the collection of all predicted CATEs and sorted units into five equally-sized groups by the size of predicted CATE. We could then examine, for example, the proportion of males within each of the five groups. These proportions were reported in Table 7 of the main paper. This allowed us to assess questions such as whether those who had a higher predicted response to treatment tended to be of one demographic group over another. All codes are provided for general use and will be stored at **[LOCATION TO BE PROVIDED UPON PUBLICATION]**.

Sensitivity Checks

Section 5.4 of the main paper illustrated that the main substantive patterns of heterogeneity were likely to be robust to the choice of ML estimator when using well-performing methods appropriate for the context. However, even if overall patterns were stable, there might be sensitivity for the unit-level CATE predictions. Figure I1 plots the CATEs from a selection of alternative approaches against BART S's CATEs. OLS S has the highest correlation with BART

S (as shown by the R^2), but it also has a high proportion of winsorized CATEs, as evidenced by the plateau above $1,000. LASSO T and, to a lesser extent, LASSO R give a smaller range of CATEs compared to BART S, reflecting the regularization inherent in those methods. They also have a fairly low correlation with BART. The remaining methods—RF MOM DR and CF—have CATEs that are somewhat correlated with the BART S CATEs and the lack of vertical scatter of the points suggest that their distributions are more compressed.

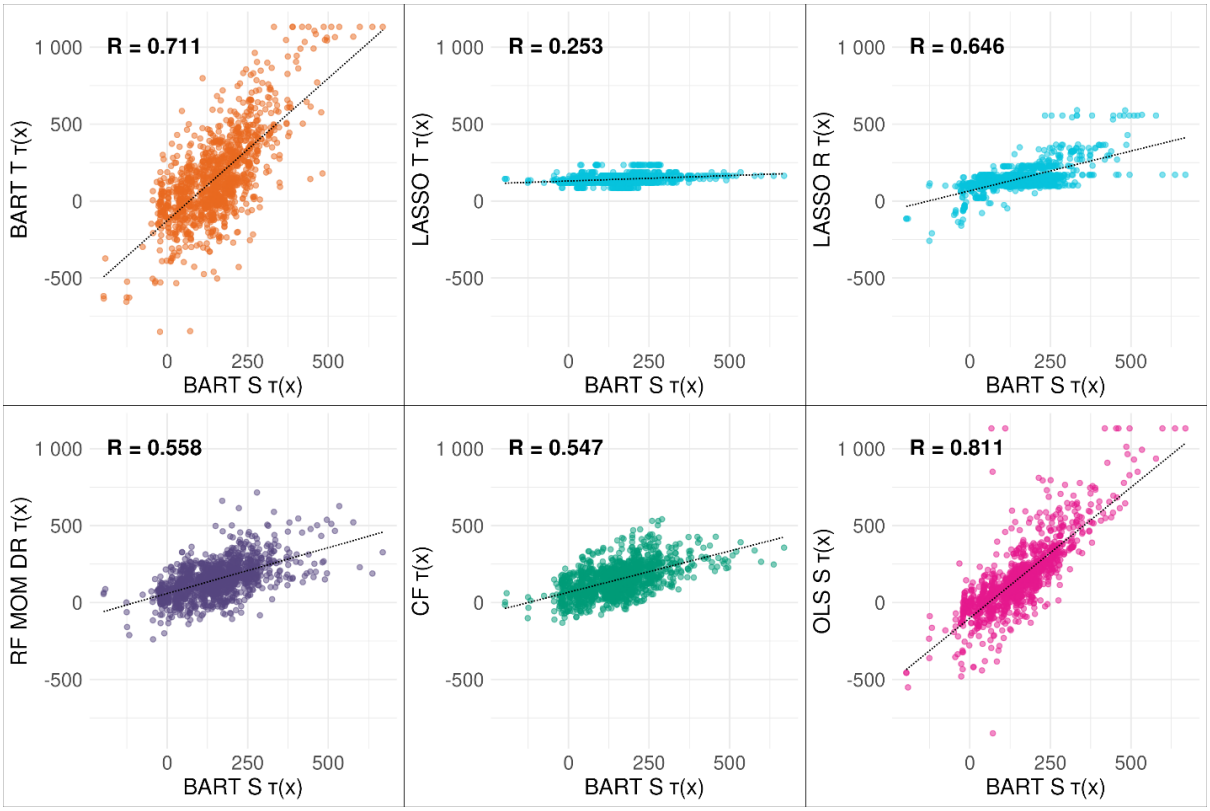

Figure I1. *Correlation Between CATEs Predicted by BART S and Those by Other ML Methods*

Note. Results are for the continuous outcome of average earnings from the Career Academy study (N = 1,254).

Section J. Simulation Result Data and Documentation

For transparency, we provide all performance measures for fitted models across outcomes, queens, and scenarios in an accompanying Excel file, along with its documentation (see below). While the raw RCT data cannot be shared due to confidentiality, we release a fully synthetic dataset with the same structure and distributional properties as the originals. Along with the dataset, we provide the full simulation code and documentation in an open repository [**LOCATION TO BE PROVIDED UPON PUBLICATION**]. These resources allow researchers to replicate every result in this paper and adapt our framework to new datasets.

References

Curth, A., & van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 130, 1810–1818. http://proceedings.mlr.press/v130/curth21a.html

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867–2880. https://doi.org/10.1002/sim.4322

Gao, X., Wang, Y., Zhou, H., & Li, J. (2020). Random forest-based causal inference for heterogeneous treatment effect estimation with observational data. *Journal of Machine Learning Research*, *21*(37), 1–37. https://jmlr.org/papers/v21/19-058.html

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1414–1423. http://proceedings.mlr.press/v70/hartford17a.html

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Kemple, J. J., & Willner, C. J. (2008). *Career Academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthood*. MDRC. https://www.mdrc.org/publication/career-academies-long-term-impacts-work-education-and-transitions-adulthood

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627. https://doi.org/10.1093/ectj/utac015

Knaus, M. C., Lechner, M., & Strittmatter, A. (2023). CATE estimation methods (Version 1.0) [Computer software]. GitHub. https://github.com/MCKnaus/CATEs

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Miller, C., & Weiss, M. J. (2021). *Accelerated Study in Associate Programs: Two-year impact report on degree completion and transfer rates at Bronx Community College*. MDRC. https://www.mdrc.org/publication/accelerated-study-associate-programs-asap

Miratrix, L. (2023, June 5). Using copulas for making calibrated data generating processes (DGPs) for simulation. *CARES Blog*. https://cares-blog.gse.harvard.edu/post/copulas-for-simulation/

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Polley, E. C., Rose, S., & van der Laan, M. J. (2011). Super Learning. In M. J. van der Laan & S. Rose (Eds.), *Targeted learning: Causal inference for observational and experimental data* (pp. 43–66). Springer. https://doi.org/10.1007/978-1-4419-9782-1_3

Rakovic, M., Ronde, H., & Verweij, P. A. (2014). A modified adaptive lasso for variable and interaction selection in Cox models. *Statistical Methods in Medical Research*, *23*(7), 641–660. https://doi.org/10.1177/0962280213515828

Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., & Fresques, H. (2015). *Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students*. MDRC. https://www.mdrc.org/sites/default/files/doubling_graduation_rates_fr.pdf

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*(4), 1010–1018. https://doi.org/10.1111/j.1541-0420.2012.01777.x