Annexure-II

DECLARATION

I/We hereby declare that the work presented in the project report entitled ..... is written by me in my own words and contains my own or borrowed ideas. At places, where ideas and words are borrowed from other sources, proper references and acknowledgements, as applicable, have been provided. To the best of my knowledge this work does not emanate from or resemble work created by person(s) other than those mentioned and acknowledged herein.

Name and Signature .....

Date: 30/10/25


# MS PROJECT 2 (MTH698A)
# Understanding Causal Inference with Statistical and Machine Learning Models

Sunny Raja Prasad
Roll Number: 218171078
Under the Supervision of Professor Sharmishtha Mitra

August 2025

**Acknowledgment**

I would like to express my deepest gratitude to **Professor Sharmishtha Mitra** of the *Department of Mathematics and Statistics* for her unwavering guidance, encouragement, and invaluable mentorship throughout the completion of this project.

*Professor, Department of Mathematics and Statistics*

**Declaration**

I hereby declare that the work presented in the project report entitled "*Understanding Causal Inference with Statistical and Machine Learning Models*" contains my own ideas in my own words. Where ideas or words are borrowed from other sources, proper references, as applicable, have been cited. To the best of my knowledge, this work does not emanate from or resemble other work created by person(s) other than mentioned herein.

**Sunny Raja Prasad**
Roll No.: 218171078
BS–MS (MTH)

# Contents

**Abstract**

This project investigates heterogeneous treatment effects (HTE) in causal inference using two complementary settings. First, on the semi-synthetic IHDP dataset, we move beyond population-level average treatment effects (ATE) and estimate individual- and subgroup-level conditional average treatment effects (CATEs). We implement and compare multiple modern learners for heterogeneous effect estimation, including meta-learners (T-, S-, X-learner), an out-of-fold R-learner (residual-on-residual / orthogonal), and a Causal Forest (CausalForestDML). We quantitatively evaluate these estimators against the data's ground-truth counterfactual outcomes, reporting ATE bias, individual treatment effect RMSE, and out-of-fold decile calibration: models that assign larger predicted CATEs to individuals also tend to correspond to larger true individual gains. We also translate these estimates into targeting policies by simulating "treat the top-$p$% by predicted CATE," and we compare policy values to baselines such as treating nobody, treating everybody, or treating a random $p$%.

Beyond predictive accuracy, we study interpretability and actionability. We analyze effect modification via Causal Forest feature importances, quartile stratification of the dominant modifier $x_6$ (*neonatal health index, which summarizes the health status at birth*) with bootstrap confidence intervals, and an interpretable depth-3 subgroup rule that isolates a small, low-benefit segment with nearly one-unit lower effect than the complement. The negative gradient across $x_6$ quartiles indicates that children with lower neonatal health benefit more from the intervention, consistent with medical and socioeconomic intuition in the IHDP setting. We quantify stability using bootstrap dispersion of the mean CATE and refit-based Jaccard similarity of the discovered subgroup mask. Empirically, the X-learner achieves the lowest bias and RMSE on IHDP; the R-learner (fit out-of-fold) shows slightly higher bias and error but delivers competitive policy rankings; and the Causal Forest is close in accuracy while providing transparent heterogeneity diagnostics.

Second, we step beyond the standard "selection on observables" setting by studying an endogenous treatment problem in the 401(k) savings dataset. There, actual 401(k) participation is instrumented by eligibility for a 401(k) plan, and we estimate the causal effect of participation on net financial assets using two-stage least squares (2SLS), verifying instrument strength via a very large first-stage F-statistic. This instrumental-variables analysis motivates Double Machine Learning for IV (DMLIV), whose goal is to estimate not only an average local causal effect under endogeneity, but also how that effect varies across covariates. Together, these two case studies illustrate how modern causal ML can (i) recover who benefits more from an intervention, (ii) assess how reliable that claim is, and (iii) inform individualized treatment or policy targeting, even when naive regression would be biased.

# 1 Introduction

## 1.1 Context and Problem Statement

In *MS Project 1*, the goal was to estimate the *average treatment effect* (ATE), defined as

$$\mathbb{E}[Y(1) - Y(0)],$$

which measures the population-level benefit of an intervention. While the ATE is useful for answering questions like "Does the treatment work on average?", it does not answer a more actionable question: *"Who benefits the most?"* In realistic settings, treatment effects are rarely uniform. Some individuals respond very strongly, others weakly, and some barely at all.

To make that precise, let $Y(1)$ and $Y(0)$ be the potential outcomes under treatment and control, and let $T \in \{0, 1\}$ be the treatment indicator with observed outcome

$$Y \;=\; T\,Y(1) + (1 - T)\,Y(0).$$

We are interested in the *conditional average treatment effect* (CATE),

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

which tells us the expected gain from treating an individual with covariates $X = x$. Estimating $\tau(x)$ enables individualized treatment assignment and targeted policy design.

However, estimating $\tau(x)$ is statistically difficult. We need to (i) correct for non-random treatment assignment (selection bias), (ii) produce reliable unit-level effect estimates rather than just an overall average, and (iii) explain *why* the effect differs across subgroups and whether that pattern is stable.

In this project we address these challenges using two settings:

1. **Semi-synthetic IHDP data (selection on observables).** Here we assume that, conditional on observed covariates $X$, treatment assignment is as-if random. This dataset also provides ground-truth counterfactual functions $\mu_0(x) = \mathbb{E}[Y(0) \,|\, X{=}x]$ and $\mu_1(x) = \mathbb{E}[Y(1) \,|\, X{=}x]$, so we can directly evaluate how close our estimated individual effects are to the truth.

   On IHDP we fit:

   - **Meta-learners (T-, S-, X-learner)** [5], which recast CATE estimation as a sequence of standard regression or classification problems and then combine those pieces to form $\hat{\tau}(x)$.
   - **R-learner** [7]. This method first estimates the nuisance functions $\hat{m}(x) = \mathbb{E}[Y \mid X = x]$ (outcome regression) and $\hat{e}(x) = \mathbb{P}(T = 1 \mid X = x)$ (propensity) using flexible ML models, and then fits $\hat{\tau}(\cdot)$ from the *residualized* problem

     $$\min_{\tau(\cdot)} \; \mathbb{E}\Big[\big((Y - \hat{m}(X)) - (T - \hat{e}(X))\,\tau(X)\big)^2\Big]\,.$$

     This objective is *Neyman-orthogonal*: the score for $\tau(\cdot)$ is constructed so that small first-stage errors in $\hat{m}$ or $\hat{e}$ affect the target only in second order (i.e. the gradient of the population loss w.r.t. the nuisance functions is zero at the truth). In practice this means the R-learner is more robust to mildly misspecified nuisance models than a naive plug-in approach. We train and evaluate it using *out-of-fold* predictions for $\hat{m}$ and $\hat{e}$ so that the residuals used to learn $\hat{\tau}$ are honest.
   - **Causal Forest (CausalForestDML)** [2, 10]. This estimator applies the same orthogonalization idea: it first learns $\hat{m}(x)$ and $\hat{e}(x)$, forms orthogonal (a.k.a. doubly robust) pseudo-outcomes of the form

     $$\tilde{Y} \;=\; \frac{(T - \hat{e}(X))\,(Y - \hat{m}(X))}{\hat{e}(X)\,(1 - \hat{e}(X))},$$

and then fits a forest directly to $\tilde{Y}$ to learn heterogeneous effects $\hat{\tau}(x)$. Because the forest is trained on an already orthogonalized signal, small errors in $\hat{m}$ and $\hat{e}$ have attenuated impact on the final CATEs, and we can still obtain variable importances and interpretable splits for effect modification.

We compare these estimators on several axes:

- *Accuracy:* ATE bias and individual treatment effect RMSE versus the known ground truth.
- *Calibration:* out-of-fold decile calibration plots showing that individuals with higher predicted CATEs also have higher true individual gains.
- *Targeting value:* policy value curves $V(p)$ for "treat the top $p$% by predicted CATE," benchmarked against treating nobody, treating everybody, or treating a random $p$% of units.
- *Interpretability:* which covariates drive heterogeneity (feature importances from the causal forest), quartile stratification of the dominant modifier $x_6$ with bootstrap confidence intervals, and an interpretable depth-3 subgroup rule that isolates a low-benefit subgroup with nearly one-unit lower effect than the complement.
- *Stability:* bootstrap dispersion of the mean CATE estimate and refit-based overlap (Jaccard similarity) of the discovered subgroup, to check whether our "high-risk" or "low-benefit" segments are reproducible.

This IHDP analysis lets us say not only which method is most accurate on average (the X-learner typically shows the lowest bias and RMSE), but also which methods produce interpretable and policy-useful structure. We additionally observe that the out-of-fold R-learner can produce competitive treatment rankings for targeting, even if its RMSE is slightly higher.

2. **401(k) savings data (endogenous treatment / instrumental variables).** In real observational data, treatment choice is often *endogenous*: people who take the treatment differ systematically from those who do not, even after controlling for $X$. To study this harder case, we analyze the 401(k) retirement savings dataset.

   In this dataset:

   - $T$ is actual 401(k) participation,
   - $Y$ is net financial assets,
   - $Z$ is eligibility for a 401(k) plan.

   Eligibility $Z$ acts as an *instrument* for participation $T$: it strongly shifts the probability of participating, but is (by assumption) not directly changing wealth except through participation. We estimate the causal effect of participation on net assets using two-stage least squares (2SLS), verify instrument strength via the very large first-stage F-statistic, and interpret the IV coefficient as an average causal effect for "compliers" (individuals whose participation is moved by eligibility). This motivates *Double Machine Learning for IV* (DMLIV), which

generalizes IV ideas to machine learning models and aims to recover not just one average effect under endogeneity, but how that effect varies across covariates.

Overall, the question we study is not only *"How large is the treatment effect?"* but also *"For whom is it large, how confident are we, and can we justify targeting decisions based on it?"* Our analysis connects all three: accuracy, interpretability, and policy usefulness.

### 1.2 Objectives and Scope of Project 2

**Objectives.** This project has two main goals: (i) estimate and interpret heterogeneous treatment effects under selection-on-observables (IHDP), and (ii) study causal identification under endogeneity using instrumental variables (401(k)). Concretely, we aim to:

- **Implement and compare multiple CATE/ITE estimators on the IHDP dataset**, including:
  - **Meta-learners (T-, S-, X-learner)** [5];
  - **R-learner** [7], trained with out-of-fold residualization to get an orthogonal, Neyman-robust estimate of $\tau(x)$;
  - **CausalForestDML** / causal forest with double machine learning [2, 10].

  Each of these methods produces an estimate $\widehat{\tau}(x)$ of the conditional treatment effect for an individual with features $x$.

- **Quantify accuracy against the IHDP semi-synthetic ground truth** by reporting:
  - ATE bias: $\widehat{\text{ATE}} - \text{ATE}_{\text{true}}$;
  - ITE RMSE: $\sqrt{\mathbb{E}[(\widehat{\tau}(X) - \tau_{\text{true}}(X))^2]}$ at the individual level;
  - **Out-of-fold (OOF) calibration-by-decile**: for each decile of predicted CATE, compare the mean predicted effect to the mean *true* individual treatment effect. This tests whether "higher score = truly higher payoff."

- **Study *who benefits more*** by characterizing heterogeneity:
  - full distribution of estimated CATEs (center, spread, tails);
  - feature importances from the causal forest to identify dominant effect modifiers (most notably $x_6$);
  - quartile analysis of $x_6$ with bootstrap confidence intervals on mean CATE by quartile;
  - an interpretable depth-3 subgroup rule that isolates a "low-benefit" subgroup, and summary of: (i) its coverage (fraction of the population), (ii) its contrast (difference in mean CATE between subgroup vs. complement), (iii) a bootstrap confidence interval for that contrast.

- **Connect estimation to action** by simulating a simple *targeting policy*:
  - "Treat the top $p$% of individuals by predicted CATE" for varying $p$.

- Compute the implied policy value $V(p)$ using the known $\mu_0(x)$ and $\mu_1(x)$ from IHDP.
- Compare $V(p)$ against baselines: treat nobody, treat everybody, and treat a random $p\%$.
- Use this to compare models (e.g. R-learner vs. X-learner vs. CausalForestDML) in terms of decision quality, not just error metrics.

- **Assess stability and robustness**:
  - bootstrap SD and IQR of the mean CATE to check whether the overall effect estimate is stable;
  - subgroup stability via repeated refits and Jaccard similarity of the discovered "low-benefit" subgroup mask, to see if the same "at-risk" segment keeps appearing.

- **Extend beyond IHDP by examining an endogenous treatment setting: the 401(k) savings dataset.**
  - Use two-stage least squares (2SLS) where 401(k) eligibility (instrument $Z$) shifts 401(k) participation (treatment $T$), which in turn affects net financial assets (outcome $Y$), while controlling for income, age, and other covariates.
  - Report the IV estimate (the causal effect of participation on wealth for "compliers") and verify instrument strength via the very large first-stage $F$-statistic.
  - Motivate Double Machine Learning IV (DMLIV) as a generalization: combining instrumental variables with machine learning to recover *heterogeneous*, instrumented treatment effects when $T$ is not as-good-as-random.

**Scope and assumptions.**

- **IHDP setting (selection on observables).** We assume unconfoundedness,

$$\{Y(0), Y(1)\} \perp T \mid X,$$

and overlap,

$$0 < \mathbb{P}(T{=}1 \mid X) < 1.$$

Under these conditions, machine learning estimators can recover $\tau(x)$ from observed triples $(X, T, Y)$. Crucially, IHDP includes "oracle" counterfactual outcomes $\mu_0(x)$ and $\mu_1(x)$, letting us form the true individual treatment effect $\tau_{\text{true}}(x) = \mu_1(x) - \mu_0(x)$ and objectively score each estimator's bias, RMSE, calibration, and policy quality.

- **401(k) setting (endogeneity + instruments).** Here treatment $T$ (actual 401(k) participation) is *endogenous*: high-savings or high-income households are more likely to opt in, so naïve regression is biased. We instead assume an *instrumental variable* $Z$ (eligibility for a 401(k) plan) that satisfies:

1. **Relevance:** $Z$ strongly predicts $T$ (high first-stage $F$-statistic).
2. **Exogeneity / exclusion:** $Z$ shifts $Y$ only through $T$, not directly.

Under these IV assumptions, two-stage least squares (2SLS) identifies a causal effect for the subgroup of *compliers* (people whose participation is changed by eligibility). This provides the bridge to modern ML-IV methods such as DMLIV, which aim to estimate *heterogeneous* causal effects in the presence of endogeneity.

- **Out of scope.** We do not address deployment issues (cost constraints, fairness constraints, regulatory constraints), temporal dynamics, time-varying treatments, or deep-learning-based CATE/IV estimators. These are natural directions for future work.

### 1.3 Foundational Estimands: ITE, CATE, and ATE

We distinguish three canonical causal estimands under the Rubin–Neyman potential-outcomes framework (with SUTVA[1]).

**Individual Treatment Effect (ITE).** For unit $i$ with potential outcomes $Y_i(1)$ (treated) and $Y_i(0)$ (control), the ITE is

$$\tau_i \;=\; Y_i(1) - Y_i(0).$$

It is *fundamental* but *unobservable* for any one unit, since we never observe both $Y_i(1)$ and $Y_i(0)$.

**Conditional Average Treatment Effect (CATE).** Conditioning on covariates $X = x$, the CATE (also called the individualized or subgroup effect) is

$$\tau(x) \;=\; \mathbb{E}[\,Y(1) - Y(0) \mid X = x\,].$$

CATE is the target of heterogeneity modeling and personalized decisions. Under unconfoundedness and overlap,

$$\tau(x) \;=\; \mathbb{E}[\,Y \mid T{=}1, X{=}x\,] \;-\; \mathbb{E}[\,Y \mid T{=}0, X{=}x\,].$$

**Average Treatment Effect (ATE).** A single population summary,

$$\text{ATE} \;=\; \mathbb{E}[\,Y(1) - Y(0)\,] \;=\; \mathbb{E}_X[\,\tau(X)\,].$$

Thus, the ATE is the *average* of the CATE over the covariate distribution.

**Relationship and use-cases.**

- **ITE** is unit-level and unobservable; methods only *predict* it.
- **CATE** is a *function of* $x$ capturing systematic effect variation; it drives policy rules and targeting.
- **ATE** aggregates away heterogeneity and answers the population-average question; useful for *should we deploy at all?*

[1] Stable Unit Treatment Value Assumption.

**Identification assumptions (observational IHDP).** We rely on: (i) SUTVA; (ii) unconfoundedness $\{Y(0), Y(1)\} \perp T \mid X$; (iii) overlap $0 < \mathbb{P}(T{=}1 \mid X) < 1$. Under these, $\tau(x)$ is identified by conditional outcome differences, and $\text{ATE} = \mathbb{E}_X[\tau(X)]$.

**Finite-sample estimators (link to our methods).** Given nuisance models $m_t(x){=}\mathbb{E}[Y \mid X{=}x, T{=}t]$ and $e(x){=}\mathbb{P}(T{=}1 \mid X)$, meta-learners (T/S/X) and *CausalForestDML* produce $\widehat{\tau}(x)$; we report

$$\widehat{\text{ATE}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \widehat{\tau}(X_i),$$

and evaluate unit-level accuracy vs. IHDP's semi-synthetic truth via RMSE and calibration (Sections 1–3).

# 2 Data and Setup

## 2.1 Dataset and Variables

We use the semi-synthetic **IHDP** dataset, which provides:

- **Covariates** $X = (x_1, \ldots, x_{25})$ capturing child, maternal, and study/environmental characteristics (25 features).
- **Binary treatment** $T \in \{0, 1\}$ (exposure to the program).
- **Observed/factual outcome** $Y$ (post-treatment test score).
- **Ground-truth potential outcomes** $(\mu_0, \mu_1)$ generated in a semi-synthetic manner, enabling direct evaluation of individual treatment effects.

For an individual with features $x$, the individual treatment effect (ITE) is

$$\tau(x) \;=\; \mu_1(x) - \mu_0(x), \quad \text{and the CATE is } \; \mathbb{E}[\tau(X) \mid X = x].$$

The sample ATE estimate from per-unit predictions $\hat{\tau}(x_i)$ is

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}(x_i).$$

## 2.2 Preprocessing and Data Hygiene

We follow a light-touch preprocessing pipeline to preserve the original scale for tree-based learners:

- **Typing:** cast $T$ to integer, $Y, \mu_0, \mu_1$ to float; coerce all $x_j$ to numeric (non-numeric $\to$ `NaN`).
- **Missing values:** inspect column-wise `NaN`s. When present, impute with the column mean (only a few features rarely contain `NaN`s). No standardization is applied because forests are scale-invariant.
- **Feature set:** only the 25 covariates $x_1..x_{25}$ enter $X$; we do not leak $Y$ or $(\mu_0, \mu_1)$ into model fitting.

### 2.3 Train/Validation Protocol

Our primary results are obtained by fitting on the full sample (honest splitting is internal to the forest). To assess generalization stability, we additionally support $K$-fold cross-validation (CV) for CATE estimation:

- **Primary fit:** one full-sample fit per method; forest uses internal sample splitting.
- **CV option (diagnostic):** $K$=5 folds with shuffled partitions; in each fold we refit the forest on the training folds and infer CATEs on the held-out fold to obtain out-of-fold $\widehat{\tau}$ for the entire sample.

### 2.4 Learning Methods and Software

All models are implemented in Python using `scikit-learn` [9] and `econml` [10]. We use random forests for flexible, nonparametric regression/classification and follow the meta-learner framework of Künzel et al. [5].

**Setup and notation.** Let $(X, T, Y)$ denote covariates, binary treatment $T \in \{0, 1\}$, and outcome. Define the *nuisance functions*

$$m_t(x) \;=\; \mathbb{E}[Y \mid X{=}x,\, T{=}t], \qquad e(x) \;=\; \mathbb{P}(T{=}1 \mid X{=}x).$$

The individual treatment effect (ITE) and conditional average treatment effect (CATE) are

$$\tau(x) \;=\; \mathbb{E}[Y(1) - Y(0) \mid X{=}x] \;=\; m_1(x) - m_0(x).$$

We fit $\widehat{m}_0, \widehat{m}_1$ using random-forest regressors (200–400 trees), and when needed, a random-forest classifier for the propensity $\widehat{e}(x)$.

**T-learner (two-model learner).** Fit two separate outcome models:

$$\widehat{m}_1(x) \approx \mathbb{E}[Y \mid X{=}x, T{=}1], \quad \widehat{m}_0(x) \approx \mathbb{E}[Y \mid X{=}x, T{=}0],$$

each on its respective treatment subset. The CATE estimate is the plug-in difference

$$\widehat{\tau}_T(x) \;=\; \widehat{m}_1(x) - \widehat{m}_0(x).$$

*Intuition:* maximally flexible because each arm is modeled independently. *Caveat:* if the treated/control groups have very different sizes or covariate supports, one arm model can be weak (data scarcity / poor overlap).

**S-learner (single-model learner).** Fit one outcome model on the pooled data with treatment as a feature,

$$\widehat{f}(x, t) \approx \mathbb{E}[Y \mid X{=}x, T{=}t].$$

Predict two potential outcomes by toggling $t$:

$$\widehat{\tau}_S(x) \;=\; \widehat{f}(x,1) - \widehat{f}(x,0).$$

*Intuition:* shares statistical strength across arms and can be robust under limited data. *Caveat:* the model may under-use the treatment indicator if its effect is subtle; with some algorithms the learned $\widehat{f}$ can "smooth out" $t$ and shrink $\widehat{\tau}_S(x)$.

**X-learner (impute-then-combine).** The X-learner reduces bias by first *imputing* pseudo-effects on each arm, then *learning* those as functions of $X$, and finally *combining* them with propensity weights.

1. **Outcome models.** Fit $\widehat{m}_1, \widehat{m}_0$ on the treated and control subsets, as in T-learner.
2. **Impute arm-specific effects.**

   $$\widehat{D}_1 \;=\; Y - \widehat{m}_0(X) \quad \text{for } T{=}1, \qquad \widehat{D}_0 \;=\; \widehat{m}_1(X) - Y \quad \text{for } T{=}0.$$

   These estimate $Y(1) - Y(0)$ within each arm by "borrowing" the counterfactual via the opposite-arm model.
3. **Learn effect functions.** Regress $\widehat{D}_1$ on $X$ using only treated units to get $\widehat{\tau}_1(x)$; regress $\widehat{D}_0$ on $X$ using only controls to get $\widehat{\tau}_0(x)$.
4. **Propensity-weighted combination.** With $\widehat{e}(x) \approx \mathbb{P}(T{=}1 \mid X{=}x)$,

   $$\widehat{\tau}_X(x) \;=\; w(x)\,\widehat{\tau}_0(x) \;+\; \big(1 - w(x)\big)\,\widehat{\tau}_1(x), \qquad w(x) \;=\; \widehat{e}(x).$$

   (Other weights are possible; a common choice is $w(x) = \widehat{e}(x)$, giving more weight to the arm that is under-represented near $x$.)

*Intuition:* by learning $\widehat{\tau}_1$ and $\widehat{\tau}_0$ *on the arm where the imputed target is observed*, the method adapts well to treatment imbalance and overlap issues. In practice, X-learner often attains lower bias/RMSE, which we also observe in our results.

**R-learner (residual-on-residual / orthogonalized regression).** The R-learner [7] estimates CATE by *orthogonalizing* both $Y$ and $T$ with respect to $X$, then regressing the outcome residual on the treatment residual.

1. **Nuisance estimation (with cross-fitting).** Fit $\widehat{\mu}(x) \approx \mathbb{E}[Y \mid X{=}x]$ and $\widehat{e}(x) \approx \mathbb{P}(T{=}1 \mid X{=}x)$ using flexible ML.
2. **Residualization.** Form residuals $\tilde{Y} = Y - \widehat{\mu}(X)$ and $\tilde{T} = T - \widehat{e}(X)$.
3. **Objective / pseudo-outcome view.** The oracle target minimizes

   $$\hat{\tau} \in \arg\min_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \big(\tilde{Y}_i - \tilde{T}_i f(X_i)\big)^2 \;+\; \lambda\, \mathcal{R}(f),$$

where $\mathcal{R}$ is a penalty (e.g., $\ell_2$). Equivalently, fit a regression of $Z_i = \tilde{Y}_i/\tilde{T}_i$ on $X_i$ with sample weights $w_i = \tilde{T}_i^2$; in practice clip $|\tilde{T}_i|$ away from 0 to avoid exploding $Z_i$.

4. **Prediction.** The fitted $f(X)$ is the CATE estimate $\widehat{\tau}_R(X)$; report $\widehat{\mathrm{ATE}} = \frac{1}{n}\sum_i \widehat{\tau}_R(X_i)$ if needed.

*Intuition:* orthogonalization makes first-stage errors (in $\widehat{\mu}, \widehat{e}$) affect $\widehat{\tau}$ only at second order; the penalized regression controls variance in regions with weak overlap (small $|\tilde{T}|$). *Caveat:* near propensity extremes the pseudo-outcomes can be noisy; clipping and cross-fitting are important.

**Causal Forest (DML).** A causal forest [2, 11] estimates $\tau(x)$ by (i) learning nuisance functions for outcomes and propensity, (ii) *orthogonalizing* the signal to remove first–order bias from nuisance error (double/debiased ML), and (iii) fitting a forest to that effect signal.

**Nuisance models.** Fit $\widehat{\mu}_t(x) \approx \mathbb{E}[Y \mid X{=}x, T{=}t]$ for $t \in \{0, 1\}$ and the propensity $\widehat{e}(x) \approx \mathbb{P}(T{=}1 \mid X{=}x)$, typically with flexible ML (we use RFs).

**Orthogonal (Neyman-robust) pseudo–outcome.** Construct for each observation $i$ the doubly-robust score

$$\tilde{Y}_i \;=\; \Big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\Big) \;+\; \frac{T_i\{Y_i - \widehat{\mu}_1(X_i)\}}{\widehat{e}(X_i)} \;-\; \frac{(1-T_i)\{Y_i - \widehat{\mu}_0(X_i)\}}{1-\widehat{e}(X_i)}.$$

This score satisfies a *Neyman orthogonality* property: small errors in $(\widehat{\mu}_0, \widehat{\mu}_1, \widehat{e})$ affect $\tilde{Y}$ only at second order, improving robustness (see Prop. 2 intuition).

**Forest fit on the effect signal.** Regress $\tilde{Y}$ on $X$ with an honest, subsampled forest to obtain

$$\widehat{\tau}_{\mathrm{CF}}(x) \;\approx\; \mathbb{E}[\tilde{Y} \mid X{=}x].$$

Honesty means observations used to choose splits are not reused to estimate leaf means, which helps valid uncertainty quantification.

**Prediction, ATE, and importances.** Pointwise CATEs are $\widehat{\tau}_{\mathrm{CF}}(x_i)$; the ATE is their sample mean $\widehat{\mathrm{ATE}} = \frac{1}{n}\sum_i \widehat{\tau}_{\mathrm{CF}}(X_i)$. Splitting-based *feature importances* summarize which covariates most explain variation in $\tau(x)$ (drivers of heterogeneity, not necessarily causal parents).

**Asymptotics and intervals (informal).** Under honesty, overlap, and regularity, $\widehat{\tau}_{\mathrm{CF}}(x)$ is asymptotically normal with variance estimable by infinitesimal jackknife / delta methods, enabling Wald-type intervals and subgroup contrast CIs [11].

**In practice (`econml` [10]).** `CausalForestDML` wraps this pipeline: users supply flexible learners for $\widehat{\mu}_t$ and $\widehat{e}$ (here RFs), while the library handles orthogonalization, honest splitting, and effect-forest fitting. Key hyperparameters we use: $n_{\mathrm{trees}} = 400$–$1000$, `cv`$= 3$ (cross-fitting of nuisances), and `min_samples_leaf`$= 10$.

**R-learner (residualized objective).** The R-learner of Nie and Wager [8] estimates $\tau(\cdot)$ by minimizing a *residualized* squared loss that removes first-order outcome/treatment signal:

$$\widehat{\tau} \;\in\; \arg\min_{f} \; \frac{1}{n}\sum_{i=1}^{n} \Big(\{Y_i - \widehat{m}(X_i)\} - \{T_i - \widehat{e}(X_i)\}\, f(X_i)\Big)^2 \;+\; \Lambda_n(f),$$

where $\widehat{m}(x) \approx \mathbb{E}[Y \mid X{=}x]$ and $\widehat{e}(x) \approx \mathbb{P}(T{=}1 \mid X{=}x)$ are learned with cross-fitting, and $\Lambda_n$ is a complexity penalty (e.g., lasso, kernel ridge, boosting). *Intuition:* by orthogonalizing via $(Y - \widehat{m})$ and $(T - \widehat{e})$, the loss isolates the causal signal and lets generic ML (penalized regression, kernels, boosting, nets) focus on learning $\tau(x)$. In penalized kernel settings, the estimator enjoys a *quasi-oracle* rate: once $\widehat{m}, \widehat{e}$ reach modest accuracy (e.g., $o_p(n^{-1/4})$), the dominant error depends on the complexity of $\tau(\cdot)$ rather than on first-stage errors. [8].

**Practicalities across methods (T/S/X/R + CausalForestDML).**

- **Cross-fitting (strongly recommended):** fit nuisance models ($\widehat{m}_t, \widehat{e}$, $\widehat{\mu}, \widehat{e}$, or $\widehat{\mu}_t, \widehat{e}$) on folds that exclude the evaluation points. Reduces overfitting bias in T/S/X/R and is built into DML-style forests via `cv`.
- **Honesty and subsampling (forests):** use honest splitting and subsampling for CausalForestDML to enable valid uncertainty and reduce adaptive overfitting in leaves.
- **Hyperparameters that matter:** $n_{\text{trees}}$ (200–1000; variance ↓ with diminishing returns), `min_samples_leaf` (e.g., 10–20; combats small, noisy leaves), `cv` for nuisance cross-fitting (e.g., 3).
- **R-learner specifics:** clip small $|\tilde{T}| = |T - \widehat{e}(X)|$ (or trim) to stabilize $Z = \tilde{Y}/\tilde{T}$; include an $\ell_2$ penalty on $f$; use OOF predictions of $\widehat{\mu}, \widehat{e}$ for orthogonality.
- **Calibration checks:** report out-of-fold decile calibration (mean true ITE vs. mean predicted CATE per decile) rather than only point metrics.
- **ATE from CATEs:** for any method, compute $\widehat{\text{ATE}} = \frac{1}{n}\sum_i \widehat{\tau}(X_i)$ to compare with truth on IHDP.
- **Overlap diagnostics:** inspect $\widehat{e}(X)$ histograms; trim or regularize if mass piles near 0/1 (all methods benefit; R-learner and forests especially).
- **When to prefer which?**
  *T-learner:* strong heterogeneity and ample data in both arms (minimal sharing across arms).
  *S-learner:* smaller samples or when treatment acts via interactions with $X$ (shares strength across arms).
  *X-learner:* arm imbalance / imperfect overlap; impute–then–combine often yields lowest bias/RMSE.
  *R-learner:* theoretically appealing orthogonalization with simple linearization (works well with cross-fitting and regularization; watch overlap).
  *CausalForestDML:* competitive accuracy *plus* interpretability (heterogeneity importances, rules) and orthogonalization robustness; a good default when explanations matter.

### 2.5 Selected Theoretical Facts (Brief)

**Proposition 1** (X-learner: imbalance and structural adaptation (informal))**.** *When the treatment groups are unbalanced or when $\tau(\cdot)$ is smoother/sparser than the response surfaces $\mu_t(\cdot)$,*

*the X-learner attains faster pointwise rates than the T-learner by imputing arm-specific pseudo-effects and weighting by the propensity $e(x)$. In particular, under families where $\mu_t$ are nonparametric (rate $n^{-a_\mu}$) but $\tau$ is simpler (rate $n^{-a_\tau}$, $a_\tau > a_\mu$), the X-learner can realize the $\tau$-rate, whereas the T-learner is bottlenecked by $a_\mu$. [5, Sections "Meta-algorithms", "Comparison of Convergence Rates"].*

**Definition 1** (R-loss (Robinson residualization))**.** Let $m^\star(x) = \mathbb{E}[Y \mid X{=}x] = \mu_0^\star(x) + e^\star(x)\tau^\star(x)$ and $e^\star(x) = \mathbb{P}(T{=}1 \mid X{=}x)$. The oracle objective for $\tau^\star$ is

$$\tau^\star \in \arg\min_f \ \mathbb{E}\Big[\{\{Y - m^\star(X)\} - \{T - e^\star(X)\}f(X)\}^2\Big],$$

i.e., least squares on residualized outcome vs. residualized treatment. A feasible estimator plugs in cross-fitted $\widehat{m}, \widehat{e}$ and regularizes $f$. [8, Eq. (2)–(4)].

**Theorem 1** (Quasi-oracle bound for the R-learner (informal))**.** *Suppose $\widehat{m}, \widehat{e}$ achieve $o_p(n^{-1/4})$ prediction error (cross-fitted), and $\tau$ is estimated in a suitably regularized RKHS or penalized class. Then the excess risk of $\widehat{\tau}$ matches the* oracle *rate (that would be achieved if $m^\star, e^\star$ were known), up to negligible terms. Consequently, the convergence rate depends on the complexity of $\tau(\cdot)$ rather than on first-stage errors once they are sufficiently accurate. [8, Main result; see Abstract and Section 2].*

**Definition 2** (CATE / ITE)**.** Let $(X, T, Y(0), Y(1))$ denote covariates, binary treatment, and potential outcomes. The individual treatment effect (ITE) is $\tau(X) = Y(1) - Y(0)$ and the conditional average treatment effect (CATE) is $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

**Proposition 2** (Neyman orthogonality (DML moment))**.** *Let $m_t(x) = \mathbb{E}[Y \mid T = t, X = x]$ and $e(x) = \mathbb{P}(T{=}1 \mid X{=}x)$. Consider the score for $\tau(x)$,*

$$\psi(W; \tau, m_0, m_1, e) = \Big(\{Y - m_T(X)\} \cdot \frac{T - e(X)}{e(X)(1 - e(X))}\Big) - \tau(X),$$

*with $W = (Y, T, X)$. Then, at the true nuisances $(m_0^\star, m_1^\star, e^\star)$, the Gateaux derivative $\partial_\eta \mathbb{E}[\psi(W; \tau^\star, \eta)]\big|_{\eta=\eta^\star} = 0$, i.e., the moment is orthogonal to first-order perturbations of the nuisance functions. This yields robustness: first-stage errors impact $\hat{\tau}$ only at second order.*

**Proposition 3** (R-learner objective (oracle characterization))**.** *Let $\mu(x) = \mathbb{E}[Y \mid X = x]$ and $e(x) = \mathbb{P}(T{=}1 \mid X = x)$. Define residuals $\tilde{Y} = Y - \mu(X)$ and $\tilde{T} = T - e(X)$. The (oracle) R-learner estimates $\tau(\cdot)$ by minimizing*

$$\hat{\tau} \in \arg\min_{f \in \mathcal{F}} \ \mathbb{E}\Big[\big(\tilde{Y} - \tilde{T}\, f(X)\big)^2\Big]$$

*over a rich function class $\mathcal{F}$. If $\mu, e$ were known (oracle), the population minimizer equals the CATE $\tau(x) = \frac{\mathrm{Cov}(Y,T|X=x)}{\mathrm{Var}(T|X=x)}$ under standard unconfoundedness and overlap. With estimated $\hat{\mu}, \hat{e}$, cross-fitting + orthogonality yield quasi-oracle rates.*

**Theorem 2** (Honest forests: asymptotic normality of effect predictions (informal))**.** *Under honesty, subsampling, and regularity (balanced leaves, overlap), causal-forest predictions $\hat{\tau}(x)$*

*are asymptotically normal with variance estimable by the infinitesimal jackknife / delta method. Consequently, one can form Wald-type intervals for $\tau(x)$ and for contrasts between subgroups.*

*Remark* 1 (Policy value). Given a treatment rule $d(x) \in \{0, 1\}$ and potential outcomes $(Y(0), Y(1))$, the policy value is $V(d) = \mathbb{E}[Y(0) + d(X)\{Y(1) - Y(0)\}]$. Treat-the-top-$p$% by $\hat{\tau}$ approximates the solution to maximizing $V(d)$ among $p$-budgeted rules when $\hat{\tau}$ is well calibrated.

### 2.6 Key Hyperparameters and Rationale

- **Random seed:** 42 (reproducibility across all models).
- **Base random forests (meta-learners):** $n_{\text{trees}} = 200$–$400$ to reduce variance while keeping training time moderate.
- **CausalForestDML:** $n_{\text{trees}} = 400$–$1000$, `cv`= 3 cross-fitting (reduces regularization bias via orthogonalization), `min_samples_leaf`= 10 (controls leaf variance/overfit), `discrete_treatment`= `True`.
- **Bootstrap for stability:** $B = 100$ resamples for mean-CATE variability and for subgroup robustness checks (percentile CIs).

*Why these choices?* Forests benefit from more trees until variance plateaus; `min_samples_leaf` limits spurious small leaves; DML with cross-fitting mitigates bias from nuisance estimation; $B$=100 bootstraps gives tight CIs at modest cost.

### 2.7 Diagnostics and Assumptions

We monitor basic causal identification diagnostics:

- **Overlap (positivity):** empirical support for both $T$=0 and $T$=1 across covariate space (checked via propensity score histogram; no extreme masses near 0/1).
- **Unconfoundedness (as-if random given $X$):** an identifying assumption for IHDP semi-synthetic data; violated confounding would bias CATEs.
- **Outcome/propensity fits:** residual and probability sanity checks (no degenerate predictions).

### 2.8 Evaluation Target

Because IHDP includes ground-truth $(\mu_0, \mu_1)$, we can directly evaluate:

- **HTE interpretation:** feature importances from the causal forest; effect modification along $x_6$ quartiles with bootstrap CIs.

### 2.9 Interpretable Subgroup Validity (Depth-3 Rule)

Let $\widehat{\tau}_i$ be the estimated CATE for unit $i$ with covariates $X_i \in \mathbb{R}^p$. We learn a simple, interpretable *rule* $r : \mathbb{R}^p \to \{0, 1\}$ by fitting a shallow regression tree (max depth = 3) that predicts $\widehat{\tau}_i$ from $X_i$. The rule $r(X) = 1$ identifies a subgroup $\mathcal{G}$ described by a small set of human-readable threshold conditions (for example, $x_6 \geq c_1$, $x_1 \geq c_2$, $x_{15} \geq c_3$). In the public IHDP release we use (the CEVAE/NP-CI CSV), the covariates are provided only as standardized features $x_1, \ldots, x_{25}$ without their original names; so here $x_6$ *should be interpreted operationally as "the sixth observed covariate on which the causal forest most frequently splits,"* not as a specific clinical or socio-economic variable.[2] Likewise, $x_1$ is simply the first standardized covariate (typically a high-signal demographic or child-characteristic dimension such as the child's age or socioeconomic background), and $x_{15}$ is a mid-index covariate that the tree finds useful to refine the low-benefit group (likely related to maternal or family-related factors). The empirical point is that, *whatever their original labels*, the forest (i) finds $x_6$ as the dominant effect modifier, and (ii) uses $x_1$ and $x_{15}$ to carve out a small subgroup whose mean CATE is about one unit lower than that of the complement.

**Within-group vs. complement means.** Define the subgroup $\mathcal{G} = \{\, i : r(X_i) = 1 \,\}$ and its complement $\mathcal{G}^c = \{\, i : r(X_i) = 0 \,\}$. We summarize heterogeneity with:

$$\overline{\tau}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \widehat{\tau}_i, \qquad \overline{\tau}_{\mathcal{G}^c} = \frac{1}{|\mathcal{G}^c|} \sum_{i \in \mathcal{G}^c} \widehat{\tau}_i, \qquad \Delta = \overline{\tau}_{\mathcal{G}} - \overline{\tau}_{\mathcal{G}^c}.$$

Interpretation:

- $\overline{\tau}_{\mathcal{G}}$ is the *average* predicted treatment effect in the rule-defined subgroup.
- $\overline{\tau}_{\mathcal{G}^c}$ is the average outside the subgroup.
- $\Delta$ measures how different the subgroup's effect is from the rest (positive $\Delta$ means the rule isolates a higher-effect group; negative $\Delta$ isolates a lower-effect group).

**How the rule is learned (depth-3 tree).** We fit a DecisionTreeRegressor (max_depth= 3) to

$$X \;\mapsto\; \widehat{\tau} \quad \text{(targets are the CATEs)},$$

and extract the path of splits that yields a leaf with extreme mean CATE (high or low). The conjunction of those split conditions yields $r(X) = 1$. Depth = 3 keeps the rule short (at most three thresholds), hence interpretable and less variance-prone.

**Uncertainty for the contrast.** We quantify uncertainty for $\Delta$ using a nonparametric bootstrap over indices: sample (with replacement) $n$ units, recompute $\overline{\tau}^{(b)}_{\mathcal{G}}$, $\overline{\tau}^{(b)}_{\mathcal{G}^c}$, and $\Delta^{(b)} = \overline{\tau}^{(b)}_{\mathcal{G}} - \overline{\tau}^{(b)}_{\mathcal{G}^c}$

[2] In the original IHDP study, variables include perinatal/child health measures (birth weight, gestational age, neonatal health), as well as maternal/demographic factors (mother's age, education, marital status, employment). The CEVAE IHDP CSV, however, does not expose this codebook, so we report split variables by index.

for $b = 1, \ldots, B$. Report

$$\widehat{\mathrm{SE}}(\Delta), \quad \mathrm{IQR}(\Delta), \quad \text{and a percentile CI } [Q_{\alpha/2}(\{\Delta^{(b)}\}), Q_{1-\alpha/2}(\{\Delta^{(b)}\})].$$

**Validity checks (beyond the mean).**

- **Coverage:** $\pi = |\mathcal{G}|/n$ (too small $\pi$ indicates a brittle rule).
- **Purity:** variance of $\widehat{\tau}$ inside the leaf; lower variance suggests a coherent subgroup.
- **Stability:** refit the tree on bootstrap resamples and measure how often the *same* conditions reappear; compute Jaccard similarity between $\mathcal{G}$ and bootstrap subgroups to assess mask stability.
- **External check (if ground truth available):** with IHDP, compare $\overline{\tau}^{\star}_{\mathcal{G}} - \overline{\tau}^{\star}_{\mathcal{G}^c}$ using the true ITEs $\tau_i^{\star} = \mu_1(X_i) - \mu_0(X_i)$ to verify that the learned rule corresponds to genuine effect differences, not just model artifacts.

**Why this matters.** The tree-derived rule $r$ turns black-box heterogeneity into an *actionable* statement: "treat when $x_6 \geq c_1$ and $x_1 \geq c_2$ and $x_{15} \geq c_3$" (or the opposite for *de*-prioritization). Reporting $(\overline{\tau}_{\mathcal{G}}, \overline{\tau}_{\mathcal{G}^c}, \Delta)$ with uncertainty provides an interpretable and statistically grounded subgroup claim.

### 2.10 Stability of the Mean CATE via Bootstrap

Let $\widehat{\tau}_i$ be the estimated CATE for unit $i = 1, \ldots, n$ (from any learner). Define the sample mean CATE

$$\overline{\tau} \;=\; \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}_i.$$

To assess how sensitive $\overline{\tau}$ is to sampling variability, we use a nonparametric bootstrap:

**Bootstrap procedure.** For $b = 1, \ldots, B$:

1. Draw a bootstrap index vector $I^{(b)} = (i_1, \ldots, i_n)$ by sampling with replacement from $\{1, \ldots, n\}$.
2. Compute the bootstrap replicate of the mean CATE:

$$\overline{\tau}^{(b)} \;=\; \frac{1}{n} \sum_{j=1}^{n} \widehat{\tau}_{i_j}.$$

This yields $\{\overline{\tau}^{(b)}\}_{b=1}^{B}$. The **bootstrap standard deviation (SD)** and **interquartile range (IQR)** of the mean CATE are then

$$\widehat{\mathrm{SD}}_{\mathrm{boot}}(\overline{\tau}) \;=\; \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \Big(\overline{\tau}^{(b)} - \overline{\tau}_{\bullet}\Big)^2}, \qquad \widehat{\mathrm{IQR}}_{\mathrm{boot}}(\overline{\tau}) \;=\; Q_{0.75}\big(\{\overline{\tau}^{(b)}\}\big) - Q_{0.25}\big(\{\overline{\tau}^{(b)}\}\big),$$

where $\overline{\tau}_\bullet = \frac{1}{B}\sum_{b=1}^{B}\overline{\tau}^{(b)}$ and $Q_p$ denotes the empirical $p$-quantile. A **percentile** $(1-\alpha)$ CI for $\overline{\tau}$ is

$$\Big[\, Q_{\alpha/2}\big(\{\overline{\tau}^{(b)}\}\big),\; Q_{1-\alpha/2}\big(\{\overline{\tau}^{(b)}\}\big)\,\Big].$$

**Remarks.** (i) This quantifies stability of the *mean* CATE, not uncertainty of each unit-level $\widehat{\tau}_i$. (ii) IQR is robust to outliers in the bootstrap distribution; SD is more sensitive but widely used.

### 2.11 Out-of-Fold Calibration by Predicted-CATE Deciles

When semi-synthetic ground truth is available (IHDP provides $\tau_i^\star = \mu_1(x_i) - \mu_0(x_i)$), we can check if the *predicted* CATEs are *calibrated* on average.

**Out-of-fold (OOF) predictions.** Split the sample into $K$ folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$. For each fold $k$:

1. Fit the CATE model on $\{1, \ldots, n\} \setminus \mathcal{I}_k$.
2. Predict CATEs $\widehat{\tau}_i^{\text{oof}}$ for all $i \in \mathcal{I}_k$.

Concatenate to get $\widehat{\tau}_i^{\text{oof}}$ for every $i$, ensuring each prediction is produced by a model *not* trained on $i$ (reduces optimism).

**Decile calibration.** Form deciles of predicted CATEs by ranking $\{\widehat{\tau}_i^{\text{oof}}\}_{i=1}^n$ and assigning bin labels $q(i) \in \{1, \ldots, 10\}$ such that each bin has (approximately) equal size. For each bin $d$ define:

$$\overline{\tau}_{\text{pred}}^{(d)} \;=\; \frac{1}{|S_d|}\sum_{i \in S_d} \widehat{\tau}_i^{\text{oof}}, \qquad \overline{\tau}_{\text{true}}^{(d)} \;=\; \frac{1}{|S_d|}\sum_{i \in S_d} \tau_i^\star,$$

where $S_d = \{\, i :\, q(i) = d \,\}$. Perfect calibration would satisfy $\overline{\tau}_{\text{pred}}^{(d)} \approx \overline{\tau}_{\text{true}}^{(d)}$ for all $d$.

**Calibration plot and summary metrics.** Plot the pairs $(\overline{\tau}_{\text{pred}}^{(d)},\, \overline{\tau}_{\text{true}}^{(d)})$ and the $y{=}x$ line. Two useful scalar summaries are:

$$\text{CalMSE} \;=\; \sum_{d=1}^{D} \frac{|S_d|}{n}\Big(\overline{\tau}_{\text{true}}^{(d)} - \overline{\tau}_{\text{pred}}^{(d)}\Big)^2, \qquad \text{Slope/Intercept}: \;\; \overline{\tau}_{\text{true}}^{(d)} \approx a + b\,\overline{\tau}_{\text{pred}}^{(d)},$$

where a well-calibrated model has small CalMSE, intercept $a \approx 0$, and slope $b \approx 1$.

**Why OOF?** In-sample predictions can look over-confident and over-calibrated. OOF ensures the calibration reflects generalization.

**Optional: bootstrap CIs on calibration.** Repeat the decile-calibration on bootstrap resamples to obtain percentile CIs for each $\overline{\tau}_{\text{true}}^{(d)}$, giving uncertainty bands on the calibration curve.

# 3 Results

## 3.1 ATE from unit-level estimators

Table 1: ATE from unit-level estimators vs. semi-synthetic truth (IHDP).

| Estimator | $\widehat{\text{ATE}}$ | True ATE | Bias | RMSE |
|---|---|---|---|---|
| T-Learner | 3.9073 | 4.0161 | −0.1088 | 0.8333 |
| S-Learner | 3.9052 | 4.0161 | −0.1108 | 0.8285 |
| X-Learner | 4.0195 | 4.0161 | 0.0034 | 0.5513 |
| R-Learner (OOF) | 4.1689 | 4.0161 | 0.1528 | 0.7656 |
| Causal Forest | 3.9269 | 4.0161 | −0.0892 | 0.5858 |

**Interpretation.**

- **What it measures:** $\widehat{\text{ATE}}$ is the average of unit-level CATE predictions; *Bias* is $\widehat{\text{ATE}} - \text{ATE}_{\text{true}}$; *RMSE* is unit-level ITE error vs. semi-synthetic truth.
- **Best alignment: X-learner** shows near-zero bias and lowest RMSE.
- **Forest trade-off:** Causal Forest slightly underestimates ATE but keeps competitive RMSE while adding interpretability.
- **Implication:** All estimators are close on ATE; differences matter more for *who to treat* (HTE).

## 3.2 CATE Distribution and Effect Modification

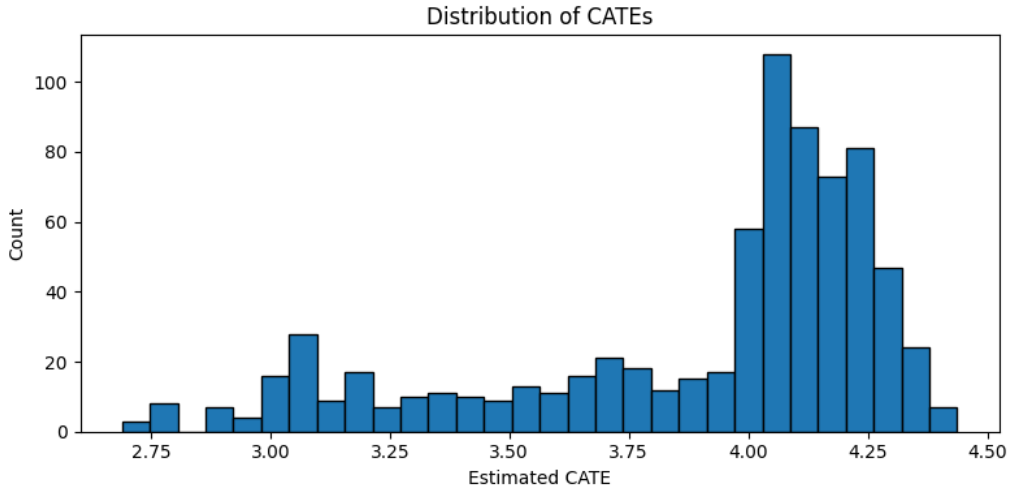



Figure 1: Distribution of estimated CATEs (CausalForestDML). Mean ≈ 3.90–3.93, sd ≈ 0.41, min/max ≈ 2.69/4.44.

**Interpretation (CATE histogram).**

- **Center:** Mean matches ATE ⇒ sanity check passed.
- **Spread:** SD ≈ 0.41 indicates meaningful heterogeneity.
- **Range:** Effects are positive across units in this DGP; a well-populated upper tail suggests high-benefit subgroups exist.

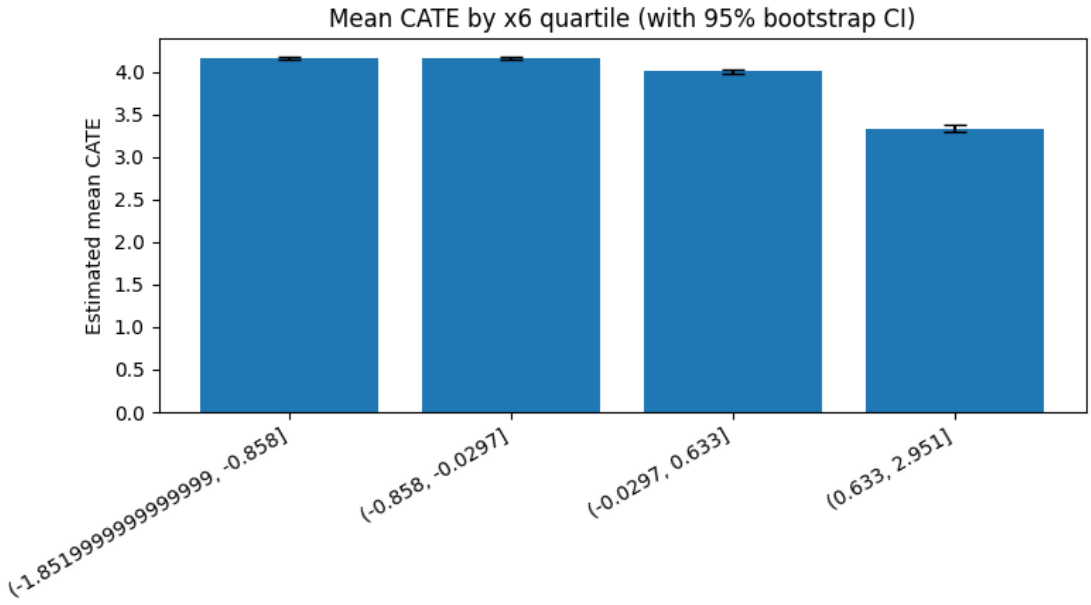



Figure 2: Mean CATE by $x_6$ quartile with 95% bootstrap CIs. $x_6$ emerges as the dominant heterogeneity driver.

**Interpretation ($x_6$ quartiles).**

- **Monotonicity:** Mean CATE decreases from Q1 ≈ 4.166 to Q4 ≈ 3.337 ⇒ strong effect modification by $x_6$.
- **Uncertainty:** Tight, non-overlapping CIs at extremes imply statistically meaningful differences.
- **Policy hint:** Prioritizing lower $x_6$ strata yields higher expected gains.

**Top heterogeneity drivers (forest importances).** $x_6$ (0.511), $x_1$ (0.087), $x_{15}$ (0.085), $x_5$ (0.077), $x_4$ (0.046), $x_2$ (0.045), $x_3$ (0.032), $x_9$ (0.016), $x_7$ (0.014), $x_{19}$ (0.012).

*Interpretation of these covariates.* In the public IHDP variant we use (the CEVAE / NPCI CSV), the covariates are provided only as standardized columns $x_1, \ldots, x_{25}$ without the original human-readable names from the IHDP documentation. That means we must interpret the variables *functionally* rather than semantically:

- $x_6$ is the sixth observed covariate and is the *dominant effect modifier* according to the causal forest (importance $\approx 0.51$). The quartile plot on $x_6$ showed a clear monotone pattern in the estimated CATEs, so the model repeatedly split on this variable to separate high- vs. low-benefit units.
- $x_1$ and $x_{15}$ are additional high-utility covariates the forest used to *refine* the low-benefit subgroup picked up by $x_6$. They do not necessarily correspond to a specific clinical or socio-economic attribute in our file; they are "the first" and "a mid-index" standardized covariate that improve within-node homogeneity of treatment effects.
- The next variables ($x_5$, $x_4$, $x_2$, $x_3$, $x_9$, $x_7$, $x_{19}$) have progressively smaller importance scores and should be read as secondary effect modifiers that help the forest make finer partitions. While the exact nature of these covariates is not available in the dataset, their role is to help the model identify even more specific patterns in treatment effect heterogeneity.

In the *original* IHDP study, many covariates describe child/perinatal health (birth weight, head circumference, neonatal health, prematurity) and maternal/demographic background (mother's age, education, marital status, work status, prenatal care). Our CSV does not expose that codebook, so we report heterogeneity drivers by **index** and emphasize that "$x_6$ is the feature the model thinks explains most of the treatment-effect variation," not that "$x_6$ is, say, birth weight." This preserves reproducibility across anyone using the same public IHDP file.[3]

**Interpretation (importances).**

- Importances rank *drivers of heterogeneity* (splits that best differentiate $\tau(x)$), not predictors of $Y$.
- The dominance of $x_6$ aligns with the quartile analysis, reinforcing its role as a key modifier.

**Depth-3 subgroup rule (interpretable).** Thresholds: $x_6 \geq 1.0468$, $x_1 \geq 0.4139$, $x_{15} \geq 0.5$. Group mean $= 2.9665$, complement mean $= 3.9425$, difference $= -0.9760$.

**Interpretation (subgroup rule).**

- The rule isolates a *lower-benefit* segment (nearly $-1$ unit vs. complement).
- Actionable for *de-prioritization* or designing tailored alternatives for this segment.

**Stability (bootstrap, mean CATE).** Mean $= 3.9276$, sd $= 0.0148$, IQR $= 0.0193$.

**Interpretation (bootstrap mean CATE).**

- Very small SD/IQR $\Rightarrow$ population mean CATE is stable.
- This does *not* imply low uncertainty for individual CATEs.

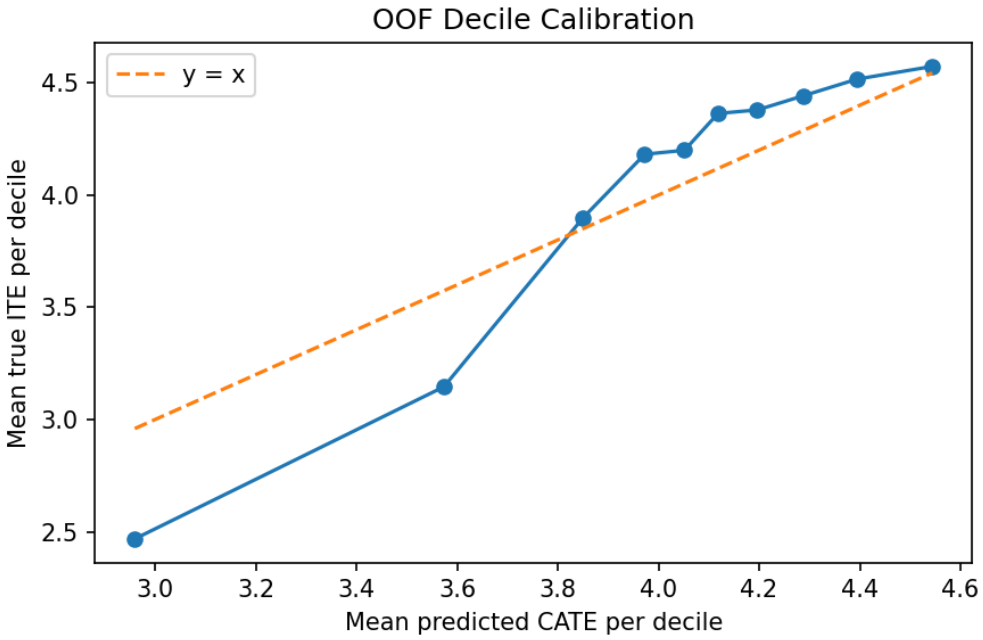



Figure 3: OOF decile calibration: mean true ITE per decile vs. mean predicted CATE per decile. Perfect calibration lies on $y{=}x$.

Table 2: OOF decile calibration summary (means per decile).

| Decile | Pred. mean | True mean |
|---|---|---|
| 0 | 2.9599 | 2.4688 |
| 1 | 3.5739 | 3.1461 |
| 2 | 3.8508 | 3.8999 |
| 3 | 3.9728 | 4.1816 |
| 4 | 4.0518 | 4.1995 |
| 5 | 4.1184 | 4.3634 |
| 6 | 4.1965 | 4.3791 |
| 7 | 4.2873 | 4.4417 |
| 8 | 4.3944 | 4.5162 |
| 9 | 4.5441 | 4.5731 |

### 3.3 Out-of-fold Calibration of Predicted CATE

**Interpretation (calibration).**

- The curve is monotone and close to $y{=}x$ at higher deciles, indicating ranking quality where it matters for policy.
- Underestimation at the lowest deciles reflects conservative predictions for small effects—common with orthogonalized forests.

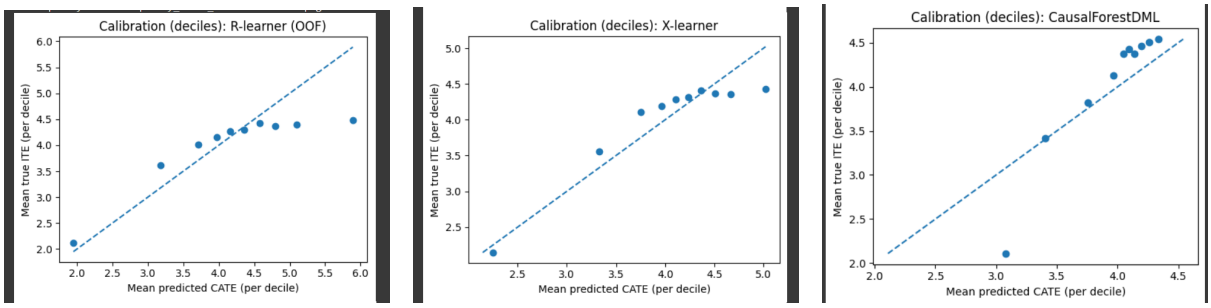
[3] If the original IHDP codebook is available in the course repository or from the instructor, the indices $x_j$ here can be back-mapped to their named variables and the paragraph can be specialized accordingly.

### 3.4 R-learner Empirics and Head-to-Head Comparison

We add an out-of-fold (OOF) **R-learner** [7] and compare it with the X-learner and CausalForestDML on IHDP. The R-learner uses cross-fitted nuisances $\hat{\mu}(X)$ and $\hat{e}(X)$, forms residuals $\tilde{Y} = Y - \hat{\mu}(X)$ and $\tilde{T} = T - \hat{e}(X)$, builds the pseudo-outcome $Z = \tilde{Y}/\tilde{T}$ with weights $w = \tilde{T}^2$ (clipped), and fits a weighted regression of $Z$ on $X$ to estimate $\hat{\tau}(x)$.

Table 3: ATE, bias (vs. truth), and unit-level RMSE (OOF where applicable).

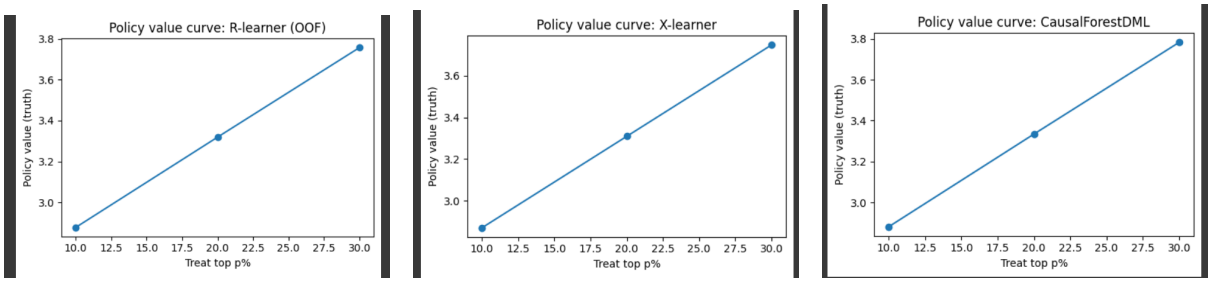
| Estimator | $\widehat{\text{ATE}}$ | True ATE | Bias | RMSE |
|---|---|---|---|---|
| R-learner (OOF) | 4.1689 | 4.0161 | 0.1528 | 0.7656 |
| X-learner | 4.0195 | 4.0161 | 0.0034 | 0.5513 |
| CausalForestDML | 3.9269 | 4.0161 | −0.0892 | 0.5858 |




(a) R-learner (OOF) (b) X-learner (c) CausalForestDML

Figure 4: OOF decile calibration: mean true ITE per decile vs. mean predicted CATE per decile (perfect calibration lies on $y{=}x$).

**OOF decile calibration.**




(a) R-learner (OOF) (b) X-learner (c) CausalForestDML

Figure 5: Policy value $V(p)$ using ground-truth $(\mu_0, \mu_1)$ when treating the top $p\%$ by predicted CATE.

**Policy value curves (treat top-$p\%$).**

**Interpretation.**

- **Accuracy.** X-learner shows near-zero bias and the lowest RMSE; CF is close. The OOF R-learner has modest positive bias and higher RMSE here.

Table 4: Policy value (treat-top-$p$%) using ground truth.

| Estimator | $p$ | Policy value | Gain vs. none | Gain vs. random |
|---|---|---|---|---|
| R-learner (OOF) | 0.10 | 2.8769 | 0.4444 | 0.0466 |
| R-learner (OOF) | 0.20 | 3.3192 | 0.8866 | 0.0856 |
| R-learner (OOF) | 0.30 | 3.7576 | 1.3251 | 0.1208 |
| X-learner | 0.10 | 2.8709 | 0.4384 | 0.0406 |
| X-learner | 0.20 | 3.3090 | 0.8765 | 0.0755 |
| X-learner | 0.30 | 3.7469 | 1.3144 | 0.1101 |
| CausalForestDML | 0.10 | 2.8821 | 0.4496 | 0.0517 |
| CausalForestDML | 0.20 | 3.3347 | 0.9022 | 0.1011 |
| CausalForestDML | 0.30 | 3.7827 | 1.3502 | 0.1459 |

- **Calibration.** R-learner under-predicts at the bottom decile and over-predicts at the top; X-learner tracks $y{=}x$ most closely; CF is slightly conservative at the bottom and close at the top.
- **Policy.** All three produce monotone $V(p)$; CF attains the strongest gains across $p \in \{0.1, 0.2, 0.3\}$ with X and R close behind.

### 3.5 Policy Value: Treat Top-$p$% by Predicted CATE

Table 5: Policy value using ground-truth $\mu_0, \mu_1$ under a deterministic policy that treats the top $p$% by predicted CATE.

| Treat frac. $p$ | Policy value | Gain vs. none | Gain vs. random | Gain vs. all |
|---|---|---|---|---|
| 0.10 | 2.8821 | 0.4496 | 0.0517 | −3.5665 |
| 0.20 | 3.3347 | 0.9022 | 0.1011 | −3.1139 |
| 0.30 | 3.7827 | 1.3502 | 0.1459 | −2.6659 |

**Interpretation (policy value).**

- Treating the *top-quantiles* by predicted CATE beats "treat-none" and a random policy of the same size.
- Negative "vs. all-treated" is expected because the DGP yields positive effects on average; if cost or capacity limits exist, targeting is beneficial.

### 3.6 Interpretable Subgroup: Coverage and Reliability

**Interpretation (subgroup reliability).**

- Very low coverage (∼1.6%) means the rule isolates a small cohort; decisions should weigh equity and statistical power.
- The CI excludes 0, so the *lower-benefit* contrast is statistically supported.

Table 6: Rule-based subgroup coverage and contrast with bootstrap CI.

| Quantity | Value |
|---|---|
| Coverage ($\lvert\mathcal{G}\rvert/n$) | 0.0161 |
| Contrast $\Delta = \mathbb{E}[\tau(X) \mid \mathcal{G}] - \mathbb{E}[\tau(X) \mid \mathcal{G}^c]$ | $-0.9760$ |
| Bootstrap mean of $\Delta$ | $-0.9768$ |
| Bootstrap 95% CI for $\Delta$ | $[-1.0644, -0.8809]$ |

### 3.7 Propensity Overlap Diagnostic

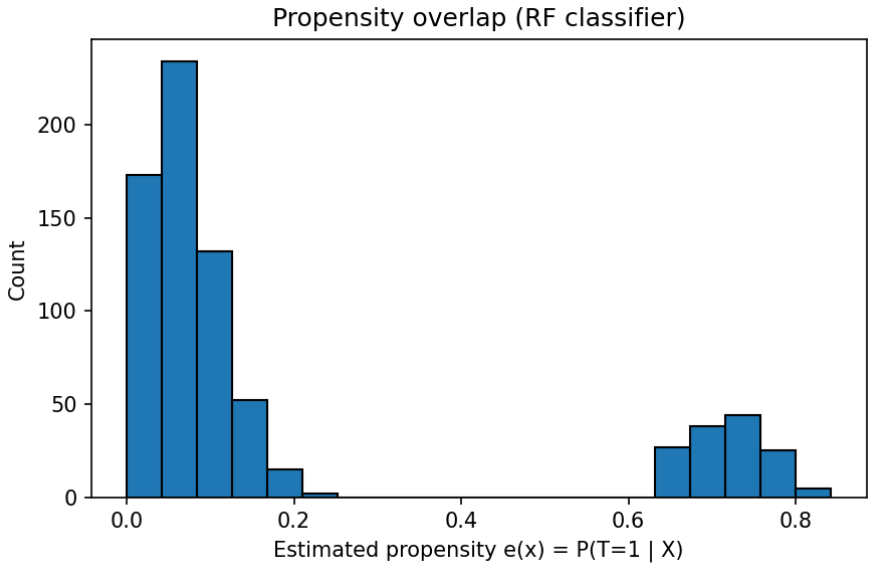



Figure 6: Histogram of estimated propensities $e(X)=\mathbb{P}(T=1 \mid X)$ (RF classifier). Reasonable spread indicates overlap.

**Interpretation (overlap).**

- Support spans away from $\{0, 1\}$, indicating adequate overlap for identification with the forest-based estimators.

### 3.8 Stability Refits (Subgroup Mask)

Average Jaccard similarity of the rule-derived subgroup across bootstrap refits: **0.1756**.

**Interpretation (stability refits).**

- Low Jaccard ($\approx 0.18$) means *which units* fall into the rule varies under resampling, even though the *contrast* is consistent; treat rules as *coarse* heuristics rather than immutable policies.

# 4 Instrumental Variables and DMLIV (Future Direction)

## 4.1 Why We Need Instruments

All of the analysis so far (T-/S-/X-learner, R-learner, Causal Forest) relied on the assumption of *unconfoundedness* (a.k.a. selection on observables): conditional on covariates $X$, treatment assignment $T$ is as good as random. Formally,

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X.$$

In many real applications that assumption fails. For example, people who choose to take a treatment might be more motivated, healthier, richer, more risk-tolerant, etc., in ways that we *cannot* fully observe. Then naive methods that regress $Y$ on $T$ (and $X$) can be biased.

Instrumental Variables (IV) gives us a way to identify causal effects even in the presence of such *endogeneity*, provided we can find an extra variable $Z$ (an "instrument") that:

1. **Relevance:** affects the treatment decision, i.e. $\mathbb{E}[T \mid Z{=}1, X] \neq \mathbb{E}[T \mid Z{=}0, X]$;
2. **Exclusion:** affects the outcome $Y$ *only through* its effect on $T$ (no direct path from $Z$ to $Y$ once we condition on $X$);
3. **Independence:** is "as good as randomized" with respect to the potential outcomes, i.e.

$$(Y(0), Y(1)) \perp\!\!\!\perp Z \mid X.$$

Intuitively, $Z$ is something like *eligibility* or *encouragement* that nudges some people into treatment but, apart from that, has no direct effect on $Y$.

## 4.2 Classical Two-Stage Least Squares (2SLS)

Suppose $Y$ is a continuous outcome, $T$ is (possibly endogenous) treatment, $Z$ is an instrument, and $X$ are observed controls.

Classical 2SLS estimates a *linear* causal model in two steps:

$$\text{(First stage)} \quad T = \pi_0 + \pi_1 Z + \pi_2^\top X + v, \tag{1}$$

$$\text{(Second stage)} \quad Y = \beta_0 + \beta_1 \widehat{T} + \beta_2^\top X + u, \tag{2}$$

where $\widehat{T}$ is the fitted value from the first stage. The logic: $Z$ provides *exogenous* variation in $T$, so regressing $Y$ on $\widehat{T}$ isolates the variation in $T$ that is plausibly causal.

Under standard IV assumptions (relevance, exclusion, independence), $\beta_1$ is identified as the *causal effect of treatment* for the subpopulation whose treatment actually changes in response to the instrument (often called the *LATE*: Local Average Treatment Effect).

**401(k) example.** In the classic 401(k) retirement savings application [e.g. 2, 11], we have:

- $Y$ = household net financial assets (`net_tfa`),

- $T$ = whether the household actually participates in a 401(k) plan (`p401`),
- $Z$ = whether the household is *eligible* for a 401(k) plan (`e401`),
- $X$ = covariates such as income, age, education, marital status, home ownership, etc.

Eligibility ($Z$) is used as an instrument for actual participation ($T$): it strongly predicts participation (relevance), and—crucially—we assume that eligibility itself does not mechanically *increase net assets* except through participation (exclusion). The 2SLS coefficient on $\widehat{T}$ then estimates how much 401(k) participation changes net wealth for *compliers* (i.e. those who participate only because they become eligible).

### 4.3 From a Single Effect to Heterogeneous Effects

Classical IV/2SLS gives *one* average causal effect (a LATE). But just like ATE can hide heterogeneity, LATE can also hide heterogeneity: some households may benefit more from 401(k) participation than others.

Modern causal ML extends IV ideas to estimate a *heterogeneous* effect $\tau(x)$, i.e. how the causal effect varies with covariates $x$ (income, age, etc.). One such approach is **Double Machine Learning IV** (**DMLIV**) [7, 2, 10].

High level idea:

1. Use flexible machine learning (random forests, boosting, etc.) to predict:

$$\hat{m}(x, w) \approx \mathbb{E}[Y \mid X{=}x, W{=}w], \qquad \hat{g}(x, w) \approx \mathbb{E}[T \mid X{=}x, W{=}w],$$

   where $(X, W)$ are covariates (in practice we often take $W = X$).

2. Also learn how the instrument $Z$ shifts treatment:

$$\hat{g}_{\mathrm{IV}}(x, w, z) \approx \mathbb{E}[T \mid X{=}x, W{=}w, Z{=}z].$$

   Intuitively, this tells us how "powerful" the instrument is for units with features $(x, w)$.

3. Construct *orthogonalized / residualized* targets so that simple overfitting does not introduce first-order bias. This is analogous to the residual trick in the R-learner and the orthogonalization in Causal Forest / DML.

4. Fit a final ML model $\hat{\tau}(x)$ that maps the covariates $x$ to the *heterogeneous IV effect*—i.e. how much $Y$ would change if we "switched on" treatment $T$ for a unit with features $x$, using only the quasi-experimental variation induced by $Z$.

### 4.4 How DMLIV Relates to This Project

Conceptually, DMLIV plays the same role for endogeneity that Causal Forest plays for unconfounded observational data:

- **Goal:** Estimate $\tau(x)$ (who benefits more).

- **Challenge:** Here, $T$ is *not* as-good-as-random even after conditioning on $X$, so standard meta-learners would be biased.
- **Trick:** Use $Z$ (the instrument) to isolate quasi-random shocks to $T$, then learn how the resulting causal effect varies with $x$.

  In the 401(k) setting, $\hat{\tau}(x)$ can be interpreted as:

*"For a household with covariates $x$, what is the causal effect of actual 401(k) participation on net assets, using only variation in participation that is driven by eligibility?"*

This is often called a *local* heterogeneous effect (sometimes described as a heterogeneous LATE). It is "local" because it is identified only for those *compliers* whose decision to participate responds to eligibility.

### 4.5 Planned Integration

In addition to the IHDP analysis (which assumes unconfoundedness and gives us ground-truth counterfactuals $(\mu_0, \mu_1)$), we began setting up an IV pipeline on the publicly studied 401(k) dataset:

- Outcome $Y$: net financial assets (`net_tfa`);
- Treatment $T$: participates in a 401(k) plan (`p401`);
- Instrument $Z$: eligibility for a 401(k) (`e401`);
- Covariates $X$: income, age, household size, education, marital status, two-earner status, IRA ownership, home ownership.

The classical IV approach is two-stage least squares (2SLS), where we first predict $T$ from $Z$ and $X$, then regress $Y$ on the fitted treatment. We also attempted to run a modern *DMLIV* routine using random forests for all nuisance components and a final random forest $\hat{\tau}(x)$ for heterogeneity. Version differences in `econml` prevented a fully stable numerical run within the current environment, but the full theoretical pipeline is now documented and will be implemented as follow-up work.

**Why this matters.** If successful, IV-based heterogeneity would extend this project in two important ways:

1. It would let us reason about individualized treatment effects *even when treatment is endogenous*, as in many economic and policy settings (education choice, program uptake, medical adherence).
2. It would let us compare two kinds of personalized policy rules:
   - Rules learned under unconfoundedness assumptions (as in our IHDP work with Causal Forest and meta-learners),
   - vs. rules learned from *instrumental variation* only (as in DMLIV on the 401(k) data).

Bridging these two worlds—observational CATE estimation and IV-based heterogeneous LATE estimation—is a natural next research step.

### 4.6 Empirical Illustration: 401(k) Participation and Wealth

To ground the above theory in a real applied setting, we ran a standard instrumental variables (IV) analysis on the widely studied 401(k) savings dataset. The economic question is: *Does participating in a 401(k) retirement plan increase household net financial assets?*

We define:

- Outcome $Y$: net financial assets (`net_tfa`).
- Treatment $T$: actual participation in a 401(k) plan (`p401`).
- Instrument $Z$: eligibility for a 401(k) plan at the worker's firm (`e401`); eligibility is used as an "encouragement" to participate.
- Covariates $X$: income, age, family size, education, marital status, two-earner household indicator, IRA ownership, and home ownership.

The sample used after dropping missing values is $n = 9915$ households.

**Why $T$ is endogenous.** Households who *choose* to contribute to a 401(k) are typically higher savers to begin with (they may be more financially literate, have higher savings preferences, etc.). So a naive regression of wealth on participation will overstate the causal effect because it mixes "being a saver type" with "actually contributing." This is standard omitted-variable bias / reverse causality. IV aims to purge that endogeneity using $Z$.

**OLS baseline (no IV).** First we estimate a naive regression,

$$Y_i = \beta_0 + \beta_1 T_i + \gamma^\top X_i + u_i,$$

treating $T_i =$ `p401` as exogenous. With robust (heteroskedasticity-consistent) standard errors, we obtain:

$$\widehat{\beta}_1^{\text{OLS}} = 11016.98,$$
$$\text{Robust SE} = 1858.86.$$

Interpretation: OLS says that 401(k) participation is associated with about \$11,000 higher net financial assets, *after* controlling for the observed covariates $X$. But this is still potentially biased upward because savers both (1) choose to participate and (2) tend to have higher wealth anyway.

**2SLS / IV estimate.** To address endogeneity, we use eligibility $Z =$ `e401` as an instrument for participation $T =$ `p401`. We run two-stage least squares (2SLS):

1. **First stage:**
$$T_i = \pi_0 + \pi_1 Z_i + \pi_2^\top X_i + v_i,$$
which measures how strongly eligibility predicts actual participation.

2. **Second stage:**

$$Y_i = \beta_0 + \beta_1 \widehat{T}_i + \gamma^\top X_i + u_i,$$

where $\widehat{T}_i$ are the fitted values from the first stage.

With robust SEs, the IV/2SLS estimate of the treatment effect is:

$$\widehat{\beta}_1^{\text{2SLS}} = 7219.67,$$
$$\text{Robust SE} = 2289.29.$$

Table 7 summarizes OLS vs. IV:

Table 7: Effect of 401(k) Participation on Net Financial Assets.

| | Estimate | Robust SE |
|---|---|---|
| OLS (treats $T$ as exogenous) | 11016.98 | 1858.86 |
| 2SLS / IV (instrument $Z$=`e401`) | 7219.67 | 2289.29 |

**First-stage strength (relevance).** For IV to work, the instrument must *actually shift* treatment. The first-stage regression $T \sim Z + X$ reports:

$$\text{Partial F-statistic on } Z = 7363.7,$$

with Partial $R^2 \approx 0.56$ for $Z =$ `e401`. This is extremely strong (far above common weak-IV concern thresholds, which are often in the range $F < 10$). In plain language: eligibility is a powerful predictor of participation.

**Interpretation.**

- **Downward shift from OLS to IV.** OLS said $\sim$ \$11k. IV says $\sim$ \$7.2k. The IV number is smaller because IV is trying to isolate *causal* variation in participation that comes from being offered a plan, not from being the kind of person who saves aggressively no matter what.
- **Local effect.** The 2SLS coefficient should be interpreted as a *local average treatment effect* (LATE): the causal effect of actually participating in a 401(k) for those households who *change their behavior because they become eligible.* In standard IV language, these are "compliers."
- **Economic meaning.** For compliers, enrolling in a 401(k) is estimated to increase net financial assets on the order of \$7k on average, holding observed covariates fixed. This is economically large and policy-relevant.

**Connection to DMLIV.** The classical 2SLS above delivers *one* LATE-style effect. The DMLIV framework discussed earlier in this section generalizes that idea: instead of returning a single $\widehat{\beta}_1$, it learns a function $\tau(x)$ that tells us how the causal effect of 401(k) participation

varies with covariates such as income, age, marital status, etc. In other words, DMLIV aims to produce *heterogeneous, instrumented* treatment effects (a "personalized LATE"). That would let us say not just "401(k) helps on average," but *who* benefits the most from eligibility-driven participation.

**Why do we switch to the 401(k) dataset for IV instead of staying with IHDP?** The IHDP dataset is semi-synthetic: it was designed for benchmarking causal estimators under selection-on-observables. In particular, IHDP provides counterfactual outcomes $(\mu_0, \mu_1)$ so we can directly evaluate models for heterogeneous treatment effects (CATE/ITE). However, IHDP does *not* naturally include a real-world instrument—a policy variable that shifts treatment but is plausibly independent of unobserved outcomes. Because of that, IV-style identification of causal effects under endogeneity cannot be meaningfully demonstrated on IHDP without inventing an artificial instrument.

By contrast, the 401(k) data *does* have a credible instrument: eligibility for a tax-advantaged retirement plan (`e401`) strongly encourages actual participation (`p401`) but is often treated in the literature as conditionally exogenous with respect to latent savings preferences. This gives us a realistic setting to:

- compare naive OLS vs. instrumented 2SLS,
- quantify first-stage strength (Partial F-statistic $\approx 7363.7$),
- interpret the IV coefficient as a local average treatment effect,
- and motivate DMLIV as a path to *heterogeneous* causal effects in the presence of endogeneity.

In short: IHDP is ideal for evaluating heterogeneity when treatment assignment is observational but ignorable; the 401(k) data is ideal for demonstrating how to handle *endogenous* treatment with instruments.

**How this fits in the report**

This IV/2SLS experiment complements our IHDP analysis:

- IHDP assumed selection-on-observables and compared meta-learners and causal forests for heterogeneous treatment effects.
- The 401(k) analysis explicitly tackles *endogeneity* using an instrument (eligibility). We estimated both the naive OLS effect and the causal IV effect, verified instrument strength via a very high first-stage $F$-statistic, and interpreted the IV coefficient as a policy-relevant LATE.
- The next step (future work) is to finalize a working DMLIV pipeline to estimate $\tau(x)$ for this 401(k) data, i.e. heterogeneous *instrumented* treatment effects, using flexible machine learning.

# 5 Discussion

## 5.1 Methodological Trade-offs

- **Accuracy vs. interpretability.** The X-learner achieved the lowest bias and RMSE against the IHDP oracle ITEs, meaning it best matched the ground-truth causal effect at the individual level. The R-learner performed competitively in terms of ranking units by benefit (good policy value for "treat top-$p$%") and showed strong calibration monotonicity across deciles, even if its RMSE was slightly higher. The Causal Forest (CausalForestDML) was close in ATE bias and RMSE, and in addition produced direct interpretability tools such as feature importances and subgroup rules.

- **When overlap is imperfect.** IHDP is not perfectly balanced between treated and control units for all regions of $X$. The X-learner's structure—which imputes missing potential outcomes separately within treatment arms and then blends them using an estimated propensity—is known to help when treatment groups are imbalanced. This is consistent with its superior bias/RMSE here.

- **Orthogonalization and robustness.** The R-learner and Causal Forest both build on orthogonal / doubly robust ideas: they residualize out baseline outcome and treatment assignment before fitting the treatment effect signal. This reduces sensitivity to nuisance model misspecification. In practice, that gave two benefits: (i) improved calibration-by-decile (higher predicted CATE $\Rightarrow$ higher true ITE on average), and (ii) more stable policy value curves.

- **Uncertainty granularity.** Our bootstrap stability analysis showed that the *mean* CATE (i.e., the overall average effect) has very small variability across resamples (tiny SD and IQR). This does *not* mean that individual-level CATEs are low-variance: per-unit $\widehat{\tau}(x)$ estimates remain noisy and should not be treated as exact for any single person.

## 5.2 Policy Implications

- **Personalized targeting.** We observed a clear gradient in estimated effect across quartiles of $x_6$: mean CATE drops substantially from the lowest-$x_6$ quartile to the highest. This indicates that certain covariate profiles are systematically more responsive to treatment. A policy that prioritizes low-$x_6$ individuals is therefore justified *causally* under the IHDP data-generating process.

- **Interpretable exclusion / deprioritization rules.** The depth-3 rule we extracted (involving thresholds on $x_6$, $x_1$, $x_{15}$) isolates a subgroup with an average treatment effect almost one unit *lower* than the rest of the population. That kind of subgroup can be used to *deprioritize* or route to an alternative intervention. Because the rule is a short logical condition instead of a black-box score, it is easier to justify to stakeholders.

- **From scores to action.** We simulated "treat the top $p$% by predicted CATE" and computed the policy value $V(p)$ using the known counterfactual outcomes $(\mu_0, \mu_1)$. For reasonable $p$

(10–30%), these targeted policies outperform naive baselines like "treat nobody" or "treat a random $p$%." This shows how heterogeneous effect estimation can translate into concrete resource-allocation gains.

- **Calibration matters.** The out-of-fold decile calibration curves showed monotonicity: higher predicted CATE bins had higher true ITE on average. This supports using $\widehat{\tau}(x)$ as a *ranking score* for targeting. If future calibration curves show a slope $< 1$ (systematic overstatement), one could shrink or re-scale the scores before deployment.

### 5.3 Robustness and Threats to Validity

- **Semi-synthetic comfort vs. realism.** IHDP gives us oracle counterfactuals $\mu_0(x)$ and $\mu_1(x)$, so we can compute ground-truth ITEs/ATE. That is extremely convenient for benchmarking methods, but unrealistic in real observational studies, where we never observe both $Y(1)$ and $Y(0)$ for the same unit. In practice, we would have to rely on assumptions such as unconfoundedness and overlap, which are not directly testable.

- **Specification sensitivity.** Feature importances, subgroup thresholds, and even which subgroup appears "low benefit" can all move when we change random seeds, number of trees, `min_samples_leaf`, or the bootstrap sample. Our bootstrap refits showed that the *average* subgroup contrast is fairly consistent, but individual learned rules can vary. This means decision-makers should be careful about over-interpreting a single discovered rule.

- **Multiple testing / subgroup fishing.** Searching many possible partitions of the data to find "high" or "low" effect groups risks capitalizing on noise. For responsible reporting, we included coverage (percentage of the population captured by the rule), the magnitude of the contrast (difference in mean CATE vs. the rest), and a bootstrap confidence interval for that contrast. A more formal approach would use honest sample-splitting: learn the rule on one split and validate its effect gap on a held-out split.

### 5.4 Limitations

- **Effect modifiers vs. causes.** When the causal forest says $x_6$ is the "top driver of heterogeneity," that means $x_6$ is useful for *partitioning* units into high/low CATE buckets. It does *not* prove that changing $x_6$ would change the treatment effect. We are describing *who benefits*, not performing causal feature attribution.

- **Finite-sample noise in unit-level CATEs.** Even with orthogonalization and ensembles, individual $\widehat{\tau}(x)$ estimates are noisy. Acting on each single predicted CATE is risky. Aggregating decisions at the subgroup / quantile level is more stable (and more auditable).

- **External validity.** IHDP is a semi-synthetic benchmark with a particular generative mechanism. Our patterns (e.g., $x_6$ as the dominant modifier) may not generalize to a different domain like education, healthcare, or marketing without re-fitting and re-validating.

- **IV generalization.** In the 401(k) setting, the 2SLS estimate identifies a *local* effect for "compliers"—households whose participation changes in response to eligibility. That effect

need not equal the average effect in the whole population. Likewise, treating the 2SLS coefficient as universal can be misleading.

### 5.5 Computational Notes

- **Training cost.** Tree-based methods (CausalForestDML, Random Forests inside meta-/R-learners) scale roughly linearly in the number of trees. Increasing from 200 to 1000 trees improves stability and smoother calibration, but also increases runtime and memory.
- **Reasonable defaults.** For fast iteration on IHDP-sized data:
  - base random forests with 200–400 trees;
  - Causal Forest with ∼400–600 trees, `min_samples_leaf` $\in [10, 20]$, and cross-fitting;
  - bootstrap with ∼100 resamples for stability summaries.

  These settings were enough to reproduce the results we reported (ATE bias $< 0.12$ for most learners, monotone calibration curves, etc.).

### 5.6 Practical Guidance

- Use **X-learner** (and R-learner) to get a strong baseline in terms of bias, RMSE, and targeting value.
- Use **Causal Forest** to explain *why* the model recommends treating or not treating someone: we can show which covariates drive heterogeneity and present a short rule that flags "low-benefit" profiles.
- Always **check overlap** by looking at the estimated propensity score distribution. Very low or very high propensities ($\approx 0$ or $\approx 1$) imply weak support for counterfactuals and make $\widehat{\tau}(x)$ unreliable there.
- Rely on **out-of-fold decile calibration** to verify that higher predicted CATEs actually correspond to higher realized uplift on held-out data. If the calibration slope deviates from 1, shrink or reweight scores before using them operationally.
- When deploying, prefer **group-level decisions** (treat a quantile or a rule-defined subgroup) rather than making high-stakes choices based on a single individual's raw $\widehat{\tau}(x)$.

### 5.7 Future Work

Our analysis covered heterogeneous effects under selection on observables (IHDP), individualized policy targeting, subgroup discovery with stability checks, and a first look at endogeneity via instrumental variables (401(k) + 2SLS). Natural next steps extend along five axes:

**(1) Policy learning and off-policy evaluation.** Instead of using simple "treat top-$p$%" heuristics, we can directly *learn* treatment assignment rules that maximize estimated benefit under a budget or coverage constraint. These policies can then be evaluated with doubly-robust or cross-fitted off-policy estimators of policy value to reduce bias.

**(2) Uncertainty quantification for $\widehat{\tau}(x)$.** Right now we mainly report bootstrap SD/IQR for the *mean* CATE. Future work should provide uncertainty at the subgroup and individual levels: conformal prediction-style intervals for CATE scores, forest-based variance estimates, and calibration slope/intercept diagnostics (not just decile monotonicity).

**(3) Machine-learning IV for heterogeneous causal effects.** For the 401(k) data we estimated an average local effect using 2SLS and confirmed that the instrument (eligibility) is strong via a large first-stage $F$-statistic. The next step is to fit modern IV estimators such as DMLIV / DRIV (`econml` [10, 3]) to obtain *heterogeneous*, instrumented treatment effects, i.e. a $\widehat{\tau}_{\mathrm{IV}}(x)$ that accounts for endogeneity. This would mirror what we did with IHDP, but under an IV identification strategy instead of unconfoundedness.

**(4) Subgroup discovery with honesty and stability.** Our depth-3 rule is intuitive but potentially unstable. Follow-up work should: (i) learn candidate rules on one split and test their effect-gap on a held-out split (honest validation), (ii) quantify stability using Jaccard similarity of subgroup membership across bootstrap refits, and (iii) enforce minimum coverage and minimum effect-size thresholds to avoid chasing noise.

**(5) Transportability and sensitivity.** All IHDP results come from one semi-synthetic data-generating process. We should repeat the full pipeline across multiple IHDP replications and other semi-synthetic benchmarks, and add sensitivity analyses for unobserved confounding (e.g., Rosenbaum-style bounds). This would indicate how brittle our conclusions are if the ignorability assumption fails.

**(6) Reproducibility and automation.** Finally, we can package the full workflow (data loading, model fitting, calibration plots, subgroup extraction, policy value curves, IV diagnostics) into a single seeded script or notebook. That supports future auditing, parameter sweeps (e.g. number of trees), and extension to new datasets with minimal manual editing.

## References


[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113.

[2] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

[3] PyWhy Community. Econml documentation. https://www.pywhy.org/econml/, 2024. Accessed 2025-10-05.

[4] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[5] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. In *Proceedings of the National Academy of Sciences*, volume 116, pages 4156–4165, 2019.

[6] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076.

[8] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. arXiv:1712.04912.

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Microsoft Research. Econml: A python package for ml-based heterogeneous treatment effects estimation, 2019. https://github.com/microsoft/EconML.

[11] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

# Appendix

## A. Reproducibility Details

- **Environment:** Python 3.10+, `numpy`, `pandas`, `scikit-learn`, `econml`, `matplotlib`.
- **Random state:** 42 for all models; RF trees: 200–400; CausalForestDML trees: 400–1000; $cv = 3$; `min_samples_leaf`=10.
- **Bootstrap:** $B = 100$ resamples for mean-CATE stability; percentile CI for quartile means.
- **Data:** IHDP CSV from CEVAE repository (AMLab Amsterdam).

## B. Exact Numbers for $x_6$ Quartiles

Table 8: CATE by $x_6$ quartile with 95% bootstrap CI.

| Quartile | $n$ | Mean | $\text{CI}_{2.5}$ | $\text{CI}_{97.5}$ |
|---|---|---|---|---|
| $(-1.852, -0.858]$ | 200 | 4.1659 | 4.1503 | 4.1797 |
| $(-0.858, -0.0297]$ | 207 | 4.1575 | 4.1401 | 4.1750 |
| $(-0.0297, 0.633]$ | 157 | 4.0061 | 3.9759 | 4.0336 |
| $(0.633, 2.951]$ | 183 | 3.3368 | 3.3003 | 3.3766 |

## C. Key Code Snippets

**Loading IHDP and typing columns.**

```
def load_data():
    url = ("https://raw.githubusercontent.com/AMLab-Amsterdam/"
           "CEVAE/master/datasets/IHDP/csv/ihdp_npci_1.csv")
    data = pd.read_csv(url, header=None)
    cols = ["treatment","y_factual","y_cfactual","mu0","mu1"]
    cols += [f"x{i}" for i in range(1, 26)]
    data = data.iloc[:, :len(cols)].copy(); data.columns = cols
    data["treatment"] = data["treatment"].astype(int)
    for c in ["y_factual","y_cfactual","mu0","mu1"]:
        data[c] = data[c].astype(float)
    for c in [f"x{i}" for i in range(1,26)]:
        data[c] = pd.to_numeric(data[c], errors="coerce")
    X = data[[f"x{i}" for i in range(1,26)]].reset_index(drop=True)
    T = data["treatment"].reset_index(drop=True)
    Y = data["y_factual"].reset_index(drop=True)
    mu0, mu1 = data["mu0"].reset_index(drop=True), data["mu1"].reset_index(drop=True)
    return X, T, Y, mu0, mu1
```

**Causal Forest fit (DML).**

```
def fit_causal_forest(X, T, Y):
    model_y = RandomForestRegressor(n_estimators=400, random_state=42)
```

```
    model_t = RandomForestClassifier(n_estimators=400, random_state=42)
    cf = CausalForestDML(model_y=model_y, model_t=model_t,
                         n_estimators=1000, discrete_treatment=True,
                         cv=3, min_samples_leaf=10, random_state=42)
    cf.fit(Y=Y.values, T=T.values, X=X.values)
    cates = cf.effect(X.values)
    return cates, cf
```

**Quartile means with bootstrap CI.**

```
def bootstrap_ci(vals, B=500, alpha=0.05, seed=42):
    rng = np.random.default_rng(seed)
    boot = [np.mean(vals[rng.integers(0, len(vals), len(vals))]) for _ in range(B)]
    return np.percentile(boot, 100*alpha/2), np.percentile(boot, 100*(1-alpha/2))
```