Causal Inference Project Report, by Tom Cal (December 10, 2025)

1. Causal Question and Motivation

Causal Inference Project Objectives

This project serves a primary educational and methodological objective:
To apply and evaluate causal inference techniques in a controlled setting.

By utilizing a semi-synthetic dataset where the "ground truth" is known, I was able to benchmark the performance of several causal estimators in their ability to recover the true treatment effect.

Beyond the educational perspective, there is also a **practical motivation grounded in the complexity of real-world clinical data.** While the outcomes are simulated, the patient covariates and treatment assignments are real.

This means our models contend with actual, complex patterns of selection bias found in hospital records.

Cautionary Note: *Because the outcome variable (Y) is simulated, the findings of this report reflect the statistical performance of the causal models and the process used to create the simulated data, not the actual clinical efficacy of HbA1c testing.*
No medical conclusions regarding diabetes management should be drawn from these results.

Motivation

Diabetes is a chronic condition affecting over 400 million people worldwide.

One component of diabetes management is the measurement of HbA1c, which provides a view of a patient's glycemic control over the previous 8 to 12 weeks.

This raises a fundamental causal question:
Does performing an HbA1c test when a person is hospitalized causally improve a patient's subsequent health outcomes?

If testing leads to better care decisions, then hospitals should enforce testing protocols. However, estimating this effect from observational data is challenging due to confounding; patients who receive tests often differ systematically from those who do not.

This project aims to disentangle these associations to isolate the true causal effect of HbA1c measurement on patient health.

Context

For this analysis, I adopt a semi-synthetic experimental design:

- Treatment (T): I use the actual treatment assignment observed in the real-world data. T=1 if the HbA1c result was measured, and T=0 if it was not measured.

- Confounders (X): I utilize actual patient characteristics to adjust for selection bias.

- Outcome (Y): I generate a simulated continuous “health score" that is closely associated with "the likelihood of avoiding hospital readmission", where a higher score is a better outcome. I generate this likelihood using a regression model. (This was generated with OLS regression and is not a true probability.)

By simulating the outcome while maintaining real-world treatment patterns and confounder distributions, I create a "Ground Truth" for the treatment effect. This allows me to evaluate the performance of various causal inference methods against a known benchmark.

2. Exploratory Data Analysis and Preprocessing

Data Source

I use the Diabetes 130-US Hospitals for Years 1999–2008 Data Set (Strack et al., 2014) from the UCI Machine Learning Repository. This dataset includes over 100,000 patient encounters and rich covariate information, including demographics, procedural details, and medication history.
The website for the dataset is https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Variable Description

To ensure a robust causal analysis, I selected confounders that influence both the likelihood of receiving an HbA1c test and the patient's health outcome. To satisfy the temporal precedence required for causality, I include only confounders that were observed prior to the treatment decision or were static demographic features.

Confounders (X):
Demographics: age, gender, and race
Utilization History: number of inpatient, outpatient, and emergency visits in the preceding year
Clinical Context: admission type, admission source, medical specialty of the admitting physician, and payer code

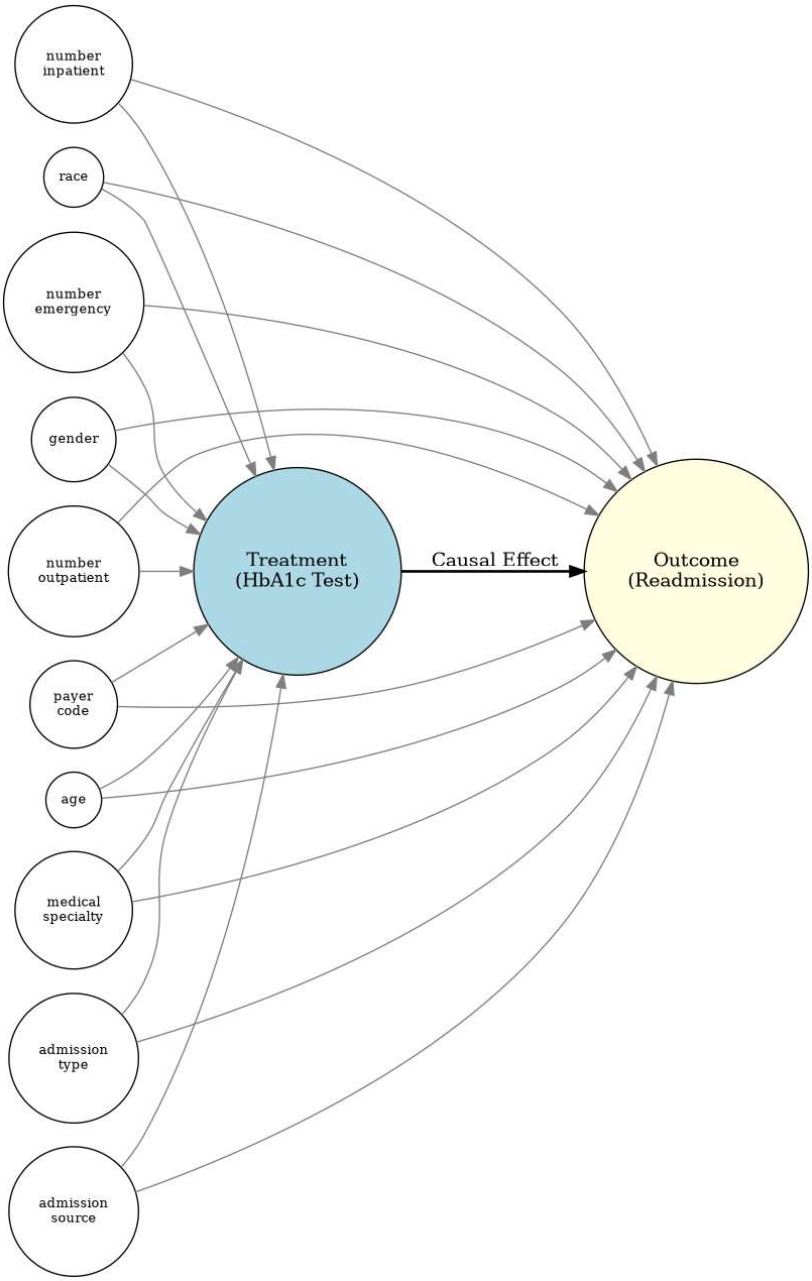
Treatment (T): A binary variable indicating if the HbA1c result was measured (T=1) or not (T=0) during the encounter. 18% of patients received treatment.

Outcome (Y): A simulated continuous score representing the "likelihood of avoiding hospital readmission". Before applying causal adjustment methods, I calculated the unadjusted difference in means (the naive estimator) to establish a baseline association.

The mean simulated outcome for the treated group is 0.611, compared to 0.588 for the control group. The naive difference without controlling for confounders is 0.023.
This raw association likely conflates the true treatment effect with selection bias, necessitating the use of the adjustment methods detailed below.

On the next page I present a **Directed Acyclic Graph (DAG).**

Directed Acyclic Graph (DAG)



Preprocessing and Data Cleaning

Prior to analysis, I assessed the confounders to ensure the positivity (overlap) assumption held and that there was a "meaningful" amount of overlap between control and treated groups for each level of each confounder. The coarsening I used aligned with and was inspired by coarsening in the (Strack et al., 2014) paper.

I "coarsened" high-cardinality confounders such as medical_specialty and admission_type that originally contained many rare categories into broader categories to ensure that every subgroup had a "meaningful" amount of overlap.

Synthetic Data Generation (Ground Truth)
To enable the rigorous evaluation of our causal estimators, I generated a semi-synthetic outcome and known "true" Conditional Average Treatment Effects (CATEs).

Outcome Simulation
I modeled the outcome Y as a non-linear function of the confounders plus the treatment effect. To mimic the noise inherent in medical data, I injected Gaussian noise.

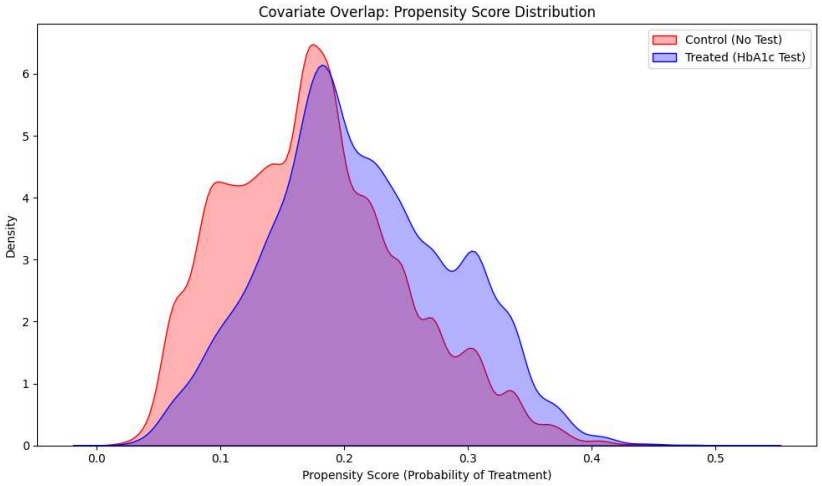
```
outcome ~
  treatment
+ number_emergency + number_outpatient + number_inpatient
+ I(number_emergency**2) + I(number_outpatient**2) + I(number_inpatient**2)
+ race_tc + gender + age_p + admission_type_coarse + admission_source_p
+ payer_code_coarse + medical_specialty_p
+ (treatment * age_p)
+ (treatment * medical_specialty_p)
+ (treatment * number_inpatient)
+ (treatment * number_emergency)'
+ (treatment * number_outpatient)'
+ Gaussian noise
```

Effect Heterogeneity (CATE)
I explicitly designed the treatment effect to vary based on patient characteristics.
Specifically, I included several interaction terms that included treatment in the ground truth model. This created a heterogeneous treatment effect that our models attempt to recover.

Propensity Score distributions.

This figure visualizes the Propensity Score distributions.

The visual overlap between the Treated and Control densities confirms that the overlap assumption is satisfied.



Analysis of Effect Heterogeneity (Ground Truth)

Because I utilize a semi-synthetic dataset with a known ground truth, I can examine if treatment assignment was correlated with the treatment's efficacy (*Selection on Gains*).

The average true causal effect for the treated patients 0.0176, while for the control patients it is 0.0141. This difference indicates that the treatment was not assigned randomly with respect to potential benefit; rather, patients with higher potential gains were more likely to be treated.

Assessment of Balance and Overlap

I assessed the distributional balance between the treated and control groups, to ensure there was overlap for each level of each categorical confounder.

3. Causal Identification and Estimation

Identification Assumptions

To interpret our statistical estimates as causal effects, we rely on three fundamental assumptions. These assumptions allow us to attribute the observed differences in outcomes to the treatment (HbA1c testing) rather than to confounding factors.

a. Stable Unit Treatment Value Assumption (SUTVA):
We assume that the treatment status of one patient (e.g., Patient A) does not affect the outcome of another (Patient B). Given that these are distinct hospital encounters, non-interference is plausible. We also assume "consistency," meaning the treatment is well-defined; there is only one version of "receiving an HbA1c test."

b. Conditional Independence Assumption (CIA / Unconfoundedness):
We assume that conditional on the observed confounders X, the treatment assignment T is independent of the potential outcomes. By including a comprehensive set of confounders, I have tried to block the "backdoor paths" that create spurious associations.

c. Positivity (Common Support):
I assume that for all patient profiles X, there is a non-zero probability of receiving either treatment status. As demonstrated in Section 2, our propensity score analysis confirmed sufficient overlap between the treated and control groups to satisfy this requirement.

Causal Estimation Methods

To estimate the Conditional Average Treatment Effects (CATE), I employed three distinct "Meta-Learner" architectures:

Method 1: Ordinary Least Squares(OLS) Regression
Method 2: S-Learner (Single Learner)
Method 3: T-Learner
Method 4: X-Learner
Method 5: Double Machine Learning (DoubleML)

Confounder Set (X):
For the **OLS model**, I used:

'number_emergency', 'number_outpatient', 'number_inpatient',
'race_tc', 'gender', 'age_p', 'admission_type_coarse',
'admission_source_p', 'payer_code_coarse', 'medical_specialty_p',
'(treatment * age_p)', '(treatment * number_inpatient)'

With OLS, I included two of the treatment interaction terms, to help capture heterogeneity, and because I felt it was “reasonable” to think that an analyst without knowledge of the process used to create the simulated data might through intuition and exploratory data analysis decide to use these two interaction terms.

For the **ML models** I used:
number_emergency, number_outpatient, number_inpatient, race_tc, gender, age_p, admission_type, admission_source, payer_code, and medical_specialty

Base Learner
All ML models used CatBoost (Gradient Boosted Trees) as the base learner, with the same settings.

CatBoost Settings Used

```
'iterations': 100,       # Reduced from 1000 for speed
'depth': 6,              # Standard depth for capturing interactions
'learning_rate': 0.1,    # Slightly higher because iterations are lower
'thread_count': 1,       # CRITICAL: Must be 1 for parallel bootstrap
'verbose': 0,            # Silent mode
'allow_writing_files': False # Prevents creating log folders everywhere
```

Why CatBoost?
I chose CatBoost for its superior handling of categorical variables without extensive preprocessing, and because it is known to work well with tabular datasets. Also, CatBoost natively encodes the predictive relationship between categorical levels and the target variable, thereby avoiding the sparsity and loss of semantic grouping inherent in traditional one-hot encoding.

4. Results, Comparative Analysis
In the table on the next page, I show the following metrics for each model.

Metrics Defined
Estimated ATE: The average treatment effect across the entire population.

ATE 95% CI: The range within which the true ATE likely falls (95% probability), calculated via bootstrapping.

ATE Bias: The absolute difference between the estimated ATE and the ground truth ATE (0.015). Lower is better.

CATE Lift (Top 10%): The average treatment effect within the top 10% of users predicted to benefit most. Higher indicates better targeting capability.

CATE Rank Corr: Spearman rank correlation between predicted treatment effects and actual treatment effects. Measures how well the model sorts users from "high responders" to "low responders" (1.0 is perfect).

CATE PEHE (RMSE): Precision in Estimation of Heterogeneous Effects. The Toot Mean Squared Error(RMSE) between the predicted individual effect and the true individual effect. Lower is better.

CATE R-squared: The proportion of variance in the true treatment effect explained by the model. Higher is better.

Results

	OLS	S-Learner	T-Learner	X-Learner	Double ML
Estimated ATE	**0.016**	0.012	0.013	0.012	**0.014**
ATE 95% CI	[0.012 0.020]	[0.007 0.015]	[0.005 0.018]	[0.005 0.018]	[0.009 0.017]
ATE Bias	**0.001**	0.003	0.002	0.003	**0.001**
CATE Lift (Top 10%)	0.002	**0.033**	0.027	0.030	0.034
CATE Rank Corr	0.229	**0.840**	0.814	0.827	0.773
CATE PEHE (RMSE)	0.022	**0.013**	0.020	0.014	0.016
CATE R-squared	0.040	**0.650**	0.222	0.632	0.473

Note: The true ATE was 0.015

Method Comparison

The analysis reveals a distinct trade-off between global accuracy (ATE) and individual targeting (CATE).

OLS and Double ML provided the most accurate global estimates of ATE, with both achieving a near-perfect ATE Bias of 0.001. Double ML, in particular, proved robust for reporting, with a tight confidence interval [0.009, 0.017] that correctly captures the true effect of 1.5%.

OLS failed completely at targeting (Rank Corr 0.229, R-squared 0.040), confirming that manual interaction terms were insufficient to capture the non-linear nature of the treatment effect.

The S-Learner emerged as the superior model for heterogeneity. It dominated the CATE metrics, achieving the highest Rank Correlation (0.840) and explaining 65% of the variance in treatment effects (R-squared 0.650). This suggests that while it slightly underestimated the *average* benefit (0.012 vs 0.015), it was exceptionally good at identifying *who* benefited.

The T-Learner struggled comparatively, showing the lowest R-squared (0.222) among the ML models. This is likely due to data imbalance and sample splitting, where the division of data into treatment and control arms reduced the effective sample size for learning complex interactions.

Model Triangulation and Sensitivity

The results strongly triangulate around a positive effect. Every single model produced a confidence interval strictly above zero. The fact that OLS, S-Learner, and X-Learner all agree on a positive impact suggests the causal finding is robust and not an artifact of a specific algorithm.

However, the variance in CATE metrics indicates high sensitivity to model architecture when estimating individual effects. This highlights the risk of relying on a single meta-learner without validation; a T-Learner alone would have significantly underrepresented the potential for targeted uplift.

5. Evaluation
Evaluation vs Ground Truth Against the ground truth ATE of 1.5%:

OLS and Double ML were the best at estimating ATE (Bias: 0.001).

OLS was accurate on average but "right for the wrong reasons," as evidenced by its inability to distinguish high responders from low responders.

S-Learner slightly underestimated the total impact (Est: 1.2%) but was highly precise in mapping the ground truth heterogeneity. Its CATE RMSE of 0.013 was the lowest of all models, meaning its individual-level predictions were closest to reality.

6. Conclusion and Future Work

Summary of Findings
The intervention drives a statistically significant positive lift of approximately 1.5%.
While the average effect is modest, the impact is highly concentrated: targeting the top 10% of responders yields a 3.3% lift, more than double the average.

The **S-Learner** is the optimal tool for operationalizing this strategy, offering superior capability in distinguishing these high-value responders.

Limitations
Synthetic Validation: While ground truth validation confirms performance, real-world deployment lacks ground truth labels for ATE and CATE.

Future Work
- Better understand causal inference for binary outcomes
- Better understand best practices when creating semi-synthetic datasets
- Better understand bootstrapping for unit-level CATE CI's
- Implement cross-validation instead of a train test split
- Learn whether automated hyperparameter tuning of ML models is useful

7. Link to the code repository
https://github.com/tc-git-1/causal_inference_project