Machine Learning for Healthcare

Subgroup 1: Depression

Scott Brinley
Nina Bryan
Samahriti Mukherjee

Subgroup 2: Diabetes

Margot Langenbach
Thomas Mezgebu
Josue Perez

Subgroup 3: Hypertension

Olivia Jonokuchi
Syon Parashar
Aytijhya Saha
Christian Sanchez

Presentation Outline

Introduction

- Data
- Prediction Basics
- Causal Basics

Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

Final Takeaways

Presentation Outline

- Introduction
  - Data
  - Prediction Basics
  - Causal Basics
- Disease Subgroups
  - 1. Depression
  - 2. Diabetes
  - 3. Hypertension
- Final Takeaways

Subgroup Structure / Disease Justification

Subgroup	Prevalence in World	Prevalence in US	Annual Cost (USD)
Depression	5.0%	4.7%	$1 Trillion
Type 2 Diabetes	8.5%	11.3%	$825 Billion
Hypertension	31.1%	48.1%	$370 Billion

Presentation Outline

Introduction

Disease Subgroups

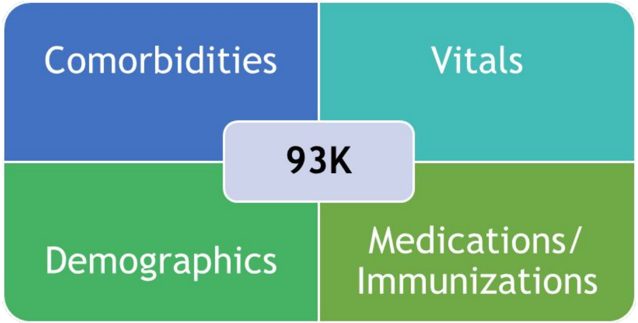
Final
Takeaways

- Data
- Prediction Basics
- Causal Basics

1. Depression
2. Diabetes
3. Hypertension

Data Overview

- Electronic Health Records from the Michigan Genomics Initiative (MGI) Database



- The white population in this data set is 7x larger than all other races combined
- Missingness: Single Imputation using the MICE R package (Multivariate Imputation by Chained Equations)

1 - Quartile in which the average of proportion of households with income greater than $75K, proportion of population age 16+ employed in professional or managerial occupations and proportion of adults with Bachelor's Degree or higher falls under

Presentation Outline

- Introduction
  - Data
  - Prediction Basics
  - Causal Basics
- Disease Subgroups
  1. Depression
  2. Diabetes
  3. Hypertension
- Final Takeaways

Prediction Problem

$$\hat{f}(x) = P\,[Disease_i \; = 1 \mid \hat{X}_i \; = x\,]$$

Models implemented with 70% train and 30% test data:

- Naive Bayes
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Logistic Regression
- Ridge Regression
- Lasso Regression
- Linear Support Vector Machine (LSVM)
- Decision Tree
- Random Forest
- XGBoost
- Neural Network
- Super Learner

Presentation Outline

Introduction

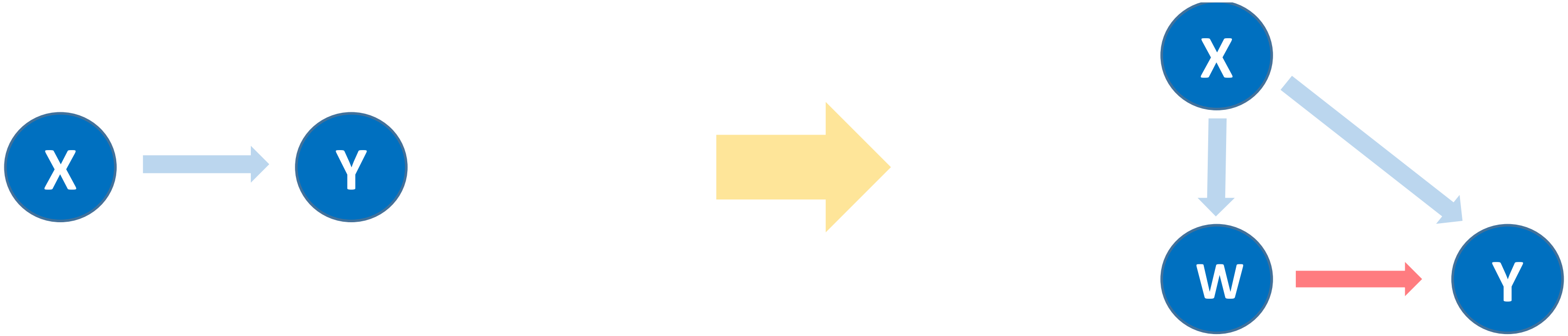
- Data
- Prediction Basics
- Causal Basics

Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

Final Takeaways

Causal Inference



The causal effect of covariates (*X*) on outcome (*Y*) considering treatment assignment (*W*).

$$\tau := E[Y_i(1) - Y_i(0)]$$

X_i : Vector of predictors (Hypertension, Obesity, BMI, etc.)

W_i : Treatment assignment; W_i = 1 shows treatment, W_i = 0 is control

Y_i : Observed outcome; $Y_i(W_i = 1) \rightarrow Y_i(1)$ represents outcome when treated, $Y_i(W_i = 0) \rightarrow Y_i(0)$ represents untreated outcome

Causality Assumptions

Conditional Unconfoundedness:

- The effect of the treatment is independent of the treatment assignment given the covariates

$$Y_i(1), Y_i(0) \perp W_i \mid X_i$$

Overlap:

- Let *Propensity Score* be defined as $e(X_i) := P[W_i = 1 \mid X_i]$
  We assume that

$$0 < e(x) < 1 \qquad for\ all\ x.$$

  This assumption is known as *Overlap*

Consistency:

- The outcome is only function of the treatment for an individual

$$Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i)$$

Estimation Methods

+ Confounders

+ Propensity Score

Difference in means

$$\hat{\tau}^{DIFF} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i - \frac{1}{n_0} \sum_{i:W_i=0} Y_i.$$

Direct estimator

$$\hat{\tau}^{DM} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0) \right).$$

Inverse Probability–Weighted (IPW) estimator

$$\hat{\tau}^{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i \frac{W_i}{\hat{e}(X_i)} - Y_i \frac{1 - W_i}{1 - \hat{e}(X_i)} \right).$$

Augmented Inverse Probability-Weighted (AIPW) estimator

$$\hat{\tau}^{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\underbrace{\hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)}_{\text{Direct estimator}} + \underbrace{\frac{W_i}{\hat{e}(X_i)} (Y_i - \hat{\mu}(X_i, 1)) - \frac{1 - W_i}{1 - \hat{e}(X_i)} (Y_i - \hat{\mu}(X_i, 0))}_{\text{IPW}} \right)$$

What Works for Whom?

Heterogeneous Treatment Effects (HTE)

Different individuals are affected differently by the treatment.

Conditional Average Treatment Effect (CATE)

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

What Works for Whom?

Heterogeneous Treatment Effects (HTE)

Different individuals are affected differently by the treatment.

Conditional Average Treatment Effect (CATE)

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

Causal Tree

Causal Tree / Causal Forest

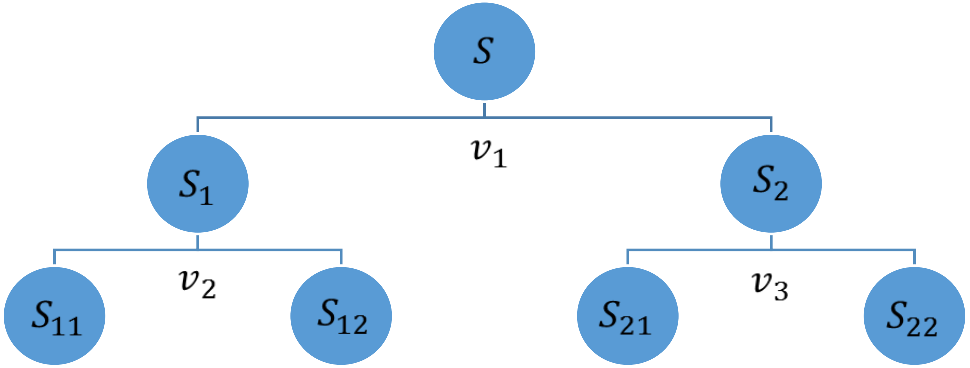
Causal Tree

Classification Tree

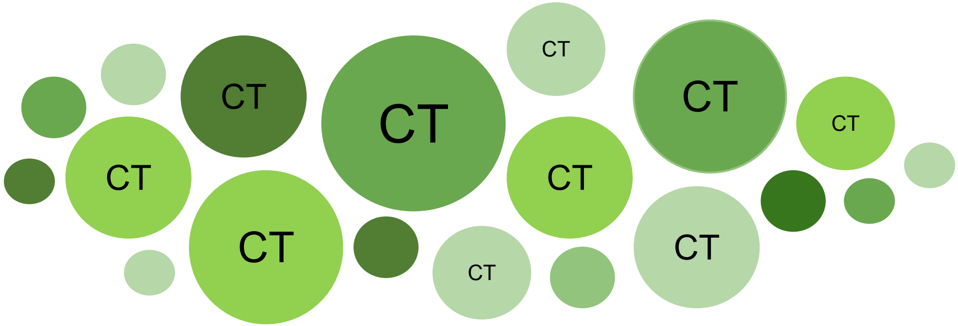
- Improve the predicted power

Causal Tree

- Difference in causal effect

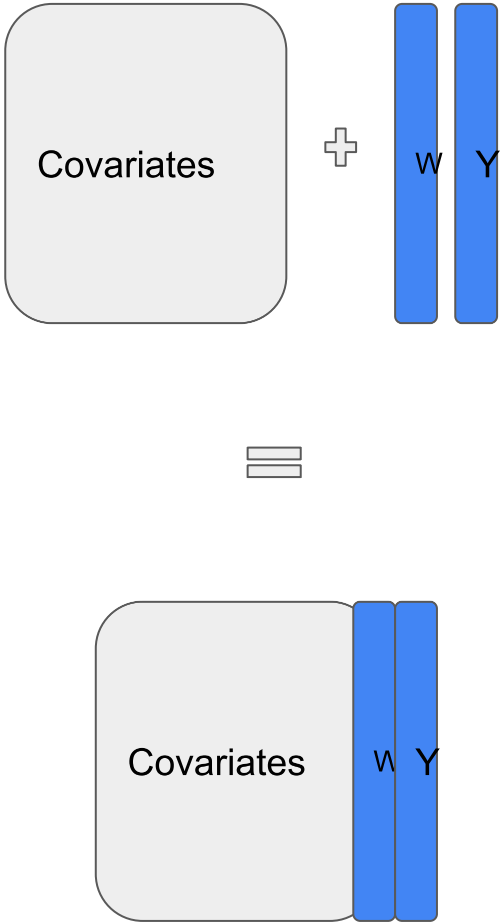


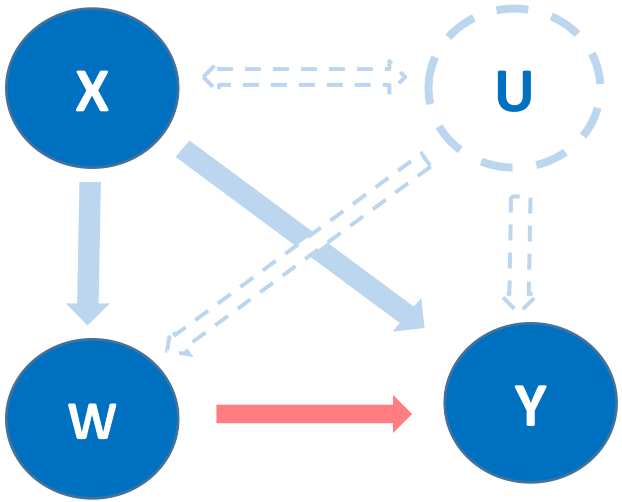
Causal Forest



Overview of Sensitivity Analysis

- Estimation methods work well for cases that lack unobserved confounders
    - However, this assumption does not hold well for observational settings
- The goal of sensitivity analysis is to determine how different strengths of a potential unobserved confounder would affect causal effect estimates


Covariates
W
Y
Covariates
W
Y


X
U
W
Y

Presentation Outline

Introduction

- Data
- Prediction Basics
- Causal Basics

Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

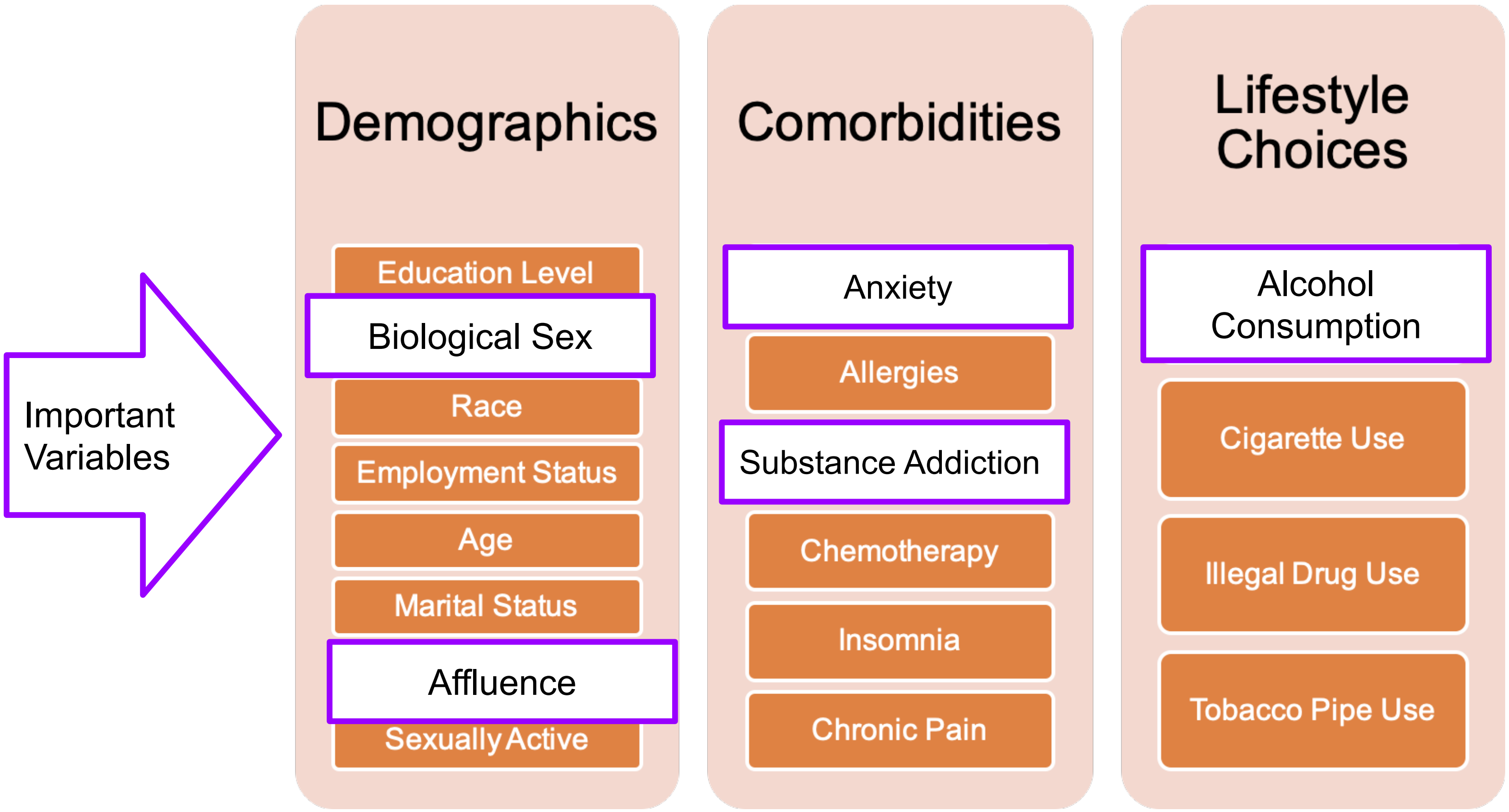
Final Takeaways

Depression Prediction and Causal Considerations

Samahriti Mukherjee[1], Nina Bryan[2], Scott Brinley[2]

Indian Statistical Institute[1], University of Michigan[2]

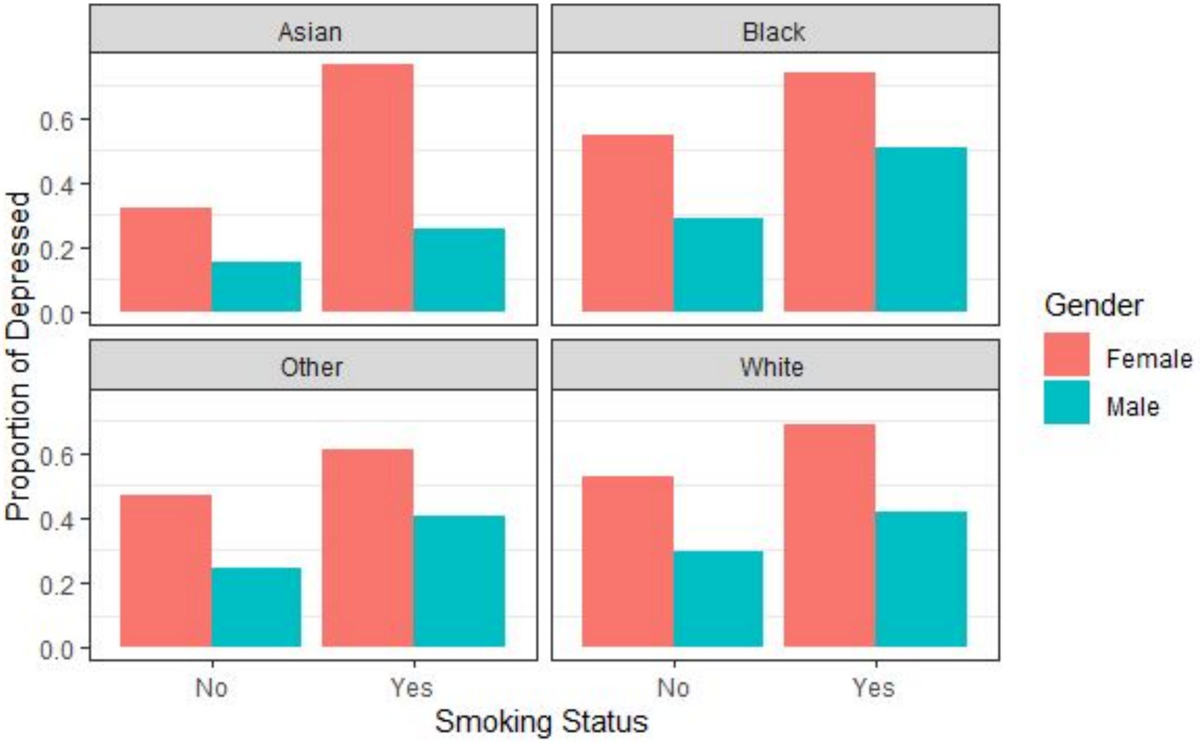
Types of Covariates



Summary of Data Set

Variables	Prevalence of Depression by Groups	
Gender	Male	31%
	Female	53.9%
Smoking Status	No	41.2%
	Yes	55.6%
	Not Asked	29%
Alcohol Usage	Yes	41.6%
	No	49.3%
Marital Status	Married	39.7%
	Unmarried	50.1%
Race	Asian	24.4%
	Other	37.2%
	White	43.9%
	Black	46.4%

Depression w.r.t. Smoking Status, Genders, Races



Prediction Results

Classifiers	Test error	Sensitivity	Specificity	AUC
Naïve Bayes	0.189	0.831	0.788	0.881
LDA	0.101	0.859	0.944	0.936
QDA	0.184	**0.885**	0.736	0.881
Logistic Regression	0.102	0.863	0.938	**0.937**
Ridge Regression	0.105	0.865	0.93	0.897
Lasso Regression	0.101	0.86	0.943	0.902
LSVM	0.105	0.856	**0.954**	0.901
RSVM	0.103	0.858	0.943	0.9
Decision tree	0.101	0.859	0.944	0.902
Random Forest	0.245	0.687	0.834	0.761
XGBoost [1]	0.1	0.865	0.939	0.902
Super Learner	**0.095**	0.876	0.932	0.904

[1] We used XgBoost, LDA and Lasso to run the Super Learner classifier

Implication of Findings

- Simple Linear Relationship
- Different Feature Importance Scores for Subgroups
- Similar Interactions Mentioned in Literature Review

New Outcome, Treatment

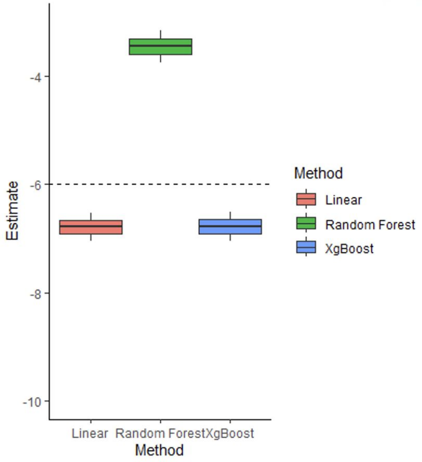
Primary Outcome: Depressive symptom measured by the Hamilton Depression Rating Scale (HDRS-17) which has the maximum score of 52 on a 17-point scale

HDRS Interval	Depression Severity
0-7	Absence of Depression
8-16	Mild
17-23	Moderate
>= 24	Severe

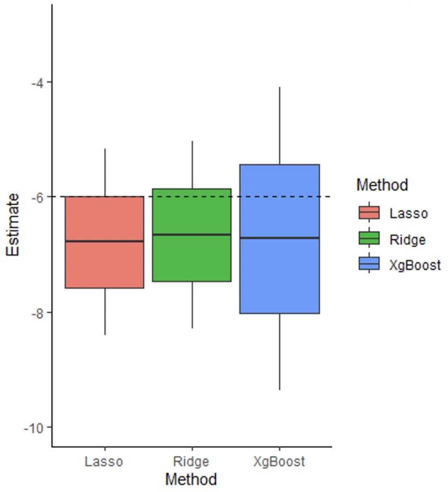
- **Treatment**: Mindfulness-Based Cognitive Therapy (MBCT) vs. Generic Antidepressant

Cognitive Behavioral Therapy + Mindfulness Meditation = Mindfulness-Based Cognitive Therapy (MBCT)

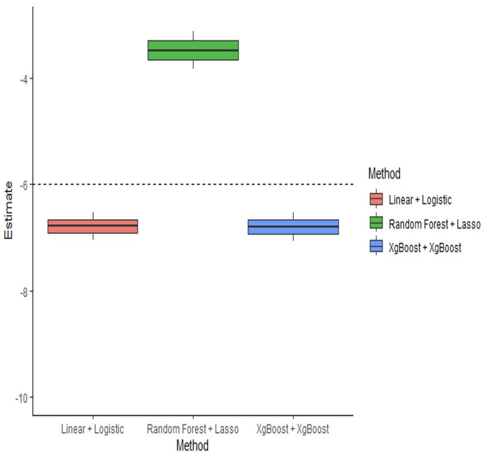
Estimators



Direct Estimate



Inverse
Propensity-Weighted (IPW)



Augmented Inverse
Propensity-Weighted
(AIPW)

Direct Estimate

Model	Relative Bias Percentage	Standard Deviation of ATE	Relative RMSE
Linear	13.1%	0.066	$\boxed{\textbf{0.13}}$
Boosting	13%	0.067	$\boxed{\textbf{0.13}}$
Random Forest	42.6%	0.075	0.43

Relative Bias Percentage

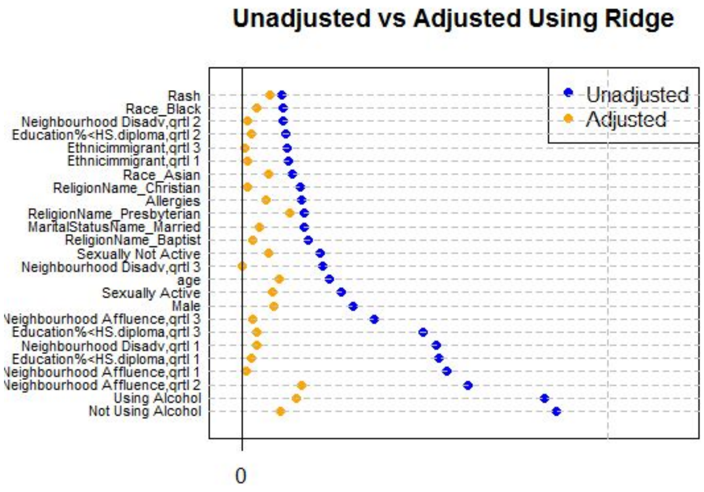
$$\left|\frac{\hat{\theta}-\theta}{\theta}\right| x\ 100$$

Relative Root Mean Square Error

$$\sqrt{\frac{(\hat{\theta}-\theta)^2+SE^2}{\theta^2}}$$

IPW

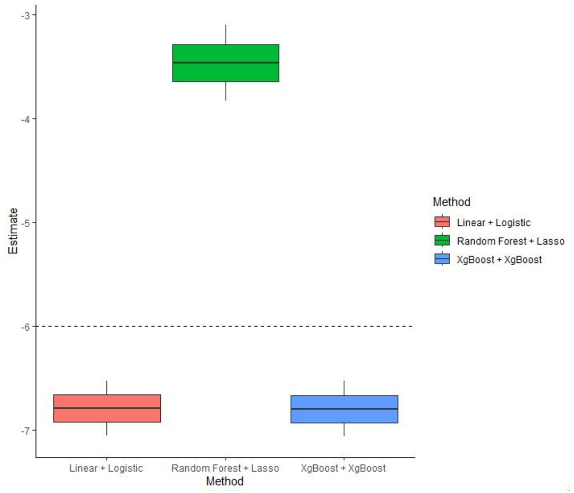
Model	Relative Bias Percentage	Standard Deviation of ATE	Relative RMSE
Logistic	1882.1%	400.477	68.83
Random Forest	276.7%	4.796	2.88
Lasso	13.2%	0.407	0.15
XGBoost	12.2%	0.661	0.16
Ridge	11.1%	0.409	**0.13**



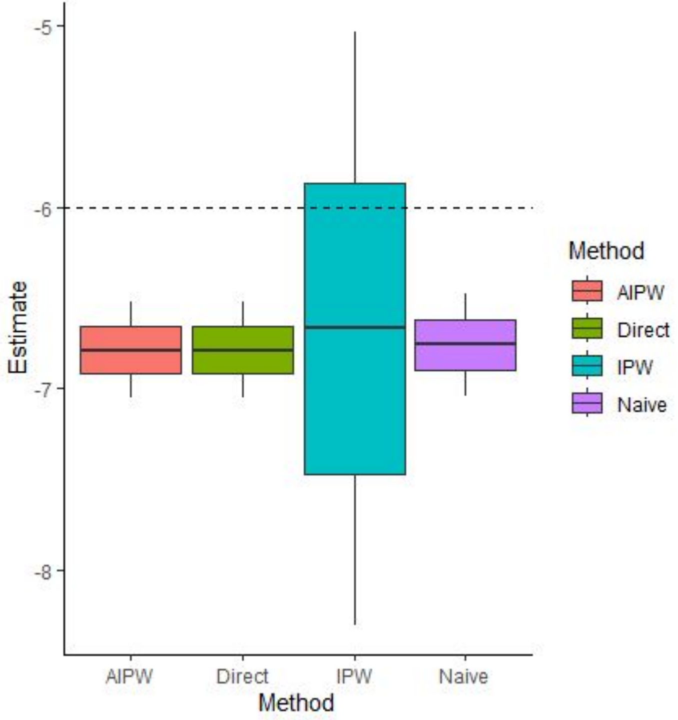
ASMD $\left(\frac{|\bar{Z}_1-\bar{Z}_0|}{\sqrt{s_1^2+s_0^2}}\right)$ for the weighted version $\left(\frac{Z_i W_i}{\hat{e}(X_i)}\ and\ \frac{Z_i(1-W_i)}{(1-\hat{e}(X_i))}\right)$ close to 0.

AIPW

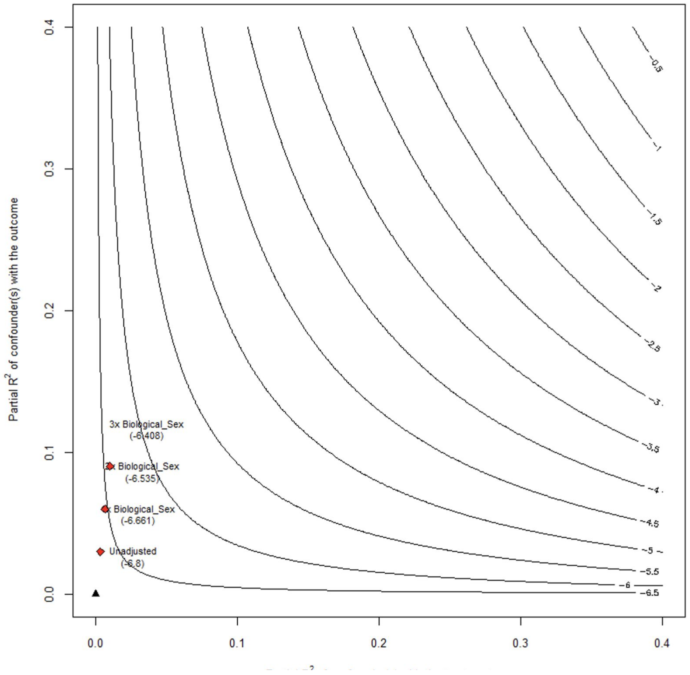
Model	Relative Bias Percentage	Standard Deviation of ATE	Relative RMSE
$\hat{u}$ = Linear $\hat{e}$ = Logistic	13.1%	0.066	**0.13**
$\hat{u}$ = XgBoost $\hat{e}$ = XgBoost	13.2%	0.067	**0.13**
$\hat{u}$ = Random Forest $\hat{e}$ = Lasso	42.2%	0.091	0.42

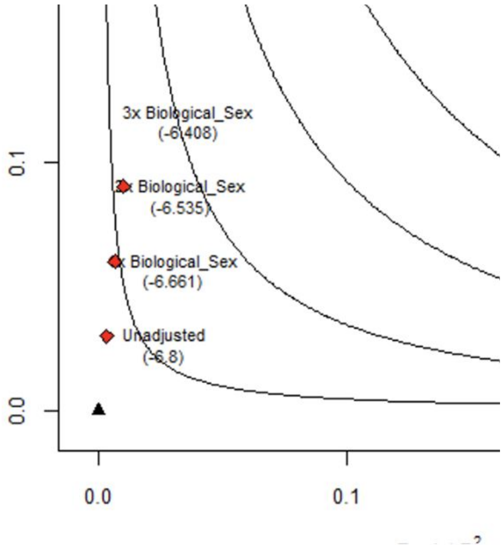


Best Estimator Models



Sensitivity Analysis





Implication of Findings

- Alcohol Usage Highest Mean Difference
- No evidence of difference in treatment effects between subgroups
- Estimators are sensitive to the simulated unobserved confounder

Presentation Outline

Introduction

- Data
- Prediction Basics
- Causal Basics

Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

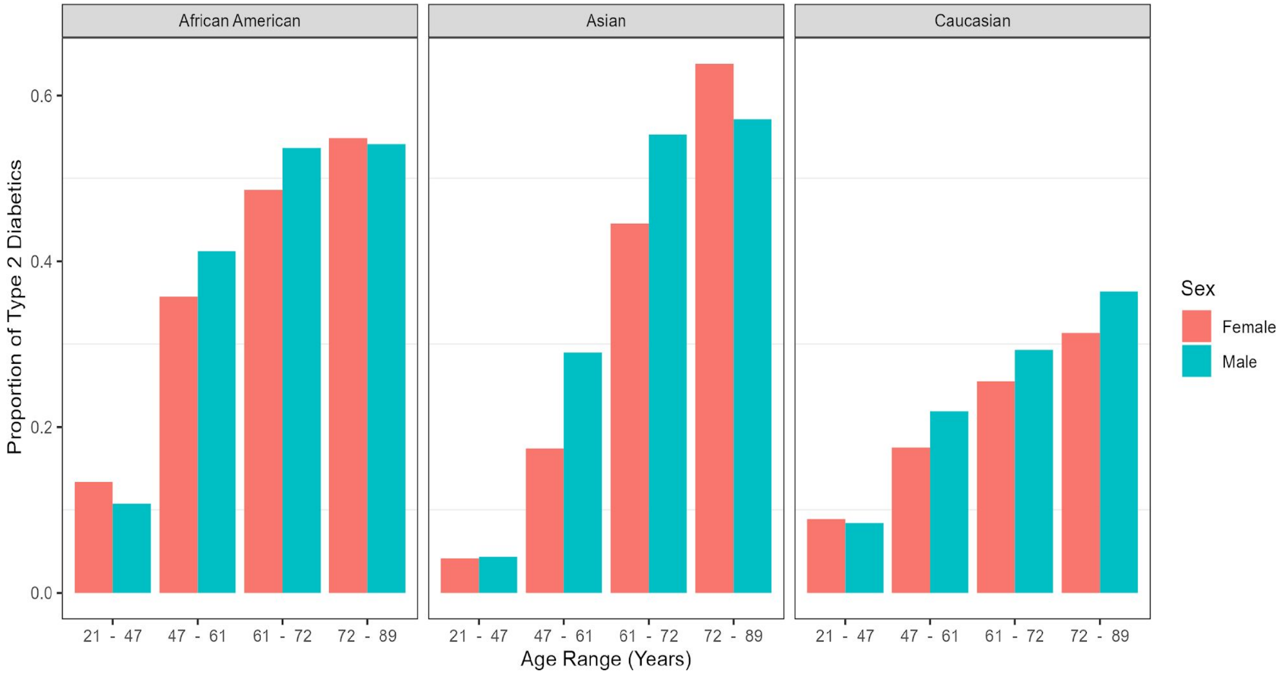
Final Takeaways

Type 2 Diabetes Mellitus Prediction and Causal Considerations

Margot Langenbach, Thomas Mezgebu, Josue Perez

University of North Carolina at Chapel Hill, University of Michigan at Ann Arbor, Universidad de Guanajuato

Exploratory Data Analysis



18 Covariates

Diseases

Epilepsy
Hyperlipidemia
Hypertension
Hypothyroidism
Obesity
Osteoporosis
Pancreatic Cancer
Peripheral Vascular Disease
Retinoschisis
Sleep Apnea
Vitamin D Deficiency

Demographics

Affluence
Age
BMI

Lifestyle Choice

Alcohol Consumption
Cigarette Use
Illegal Drug Use

Important Variables

Prediction Models

Prediction error:

Super Learner (0.222)

* LDA, Neural Net, Random Forest, XGBoost

Sensitivity:

Random Forest (0.775)

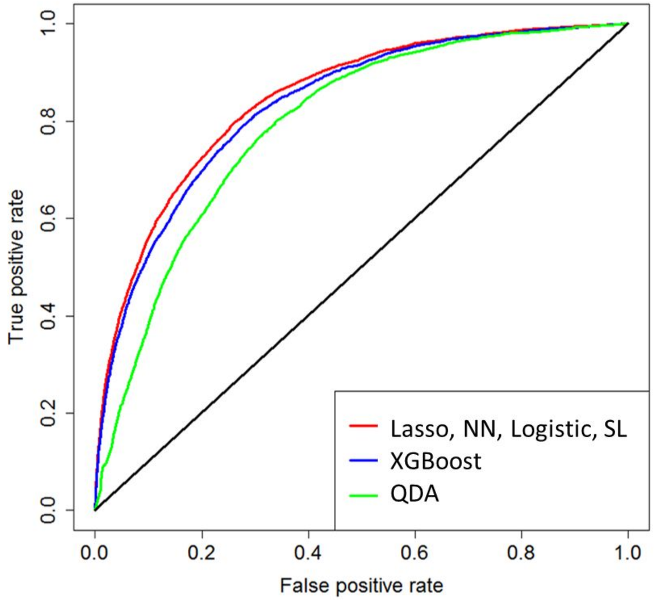
Specificity:

XGBoost, Lasso (0.796)

AUC:

Super Learner (0.852)

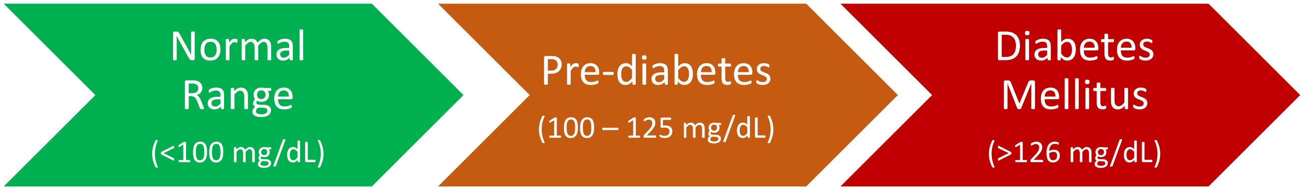
* LDA, Neural Net, Random Forest, XGBoost



Overview of Causal Problem

Outcome of Interest: Expected Fasting Plasma Glucose (*FPG*)

- The FPG is the simplest and quickest way to measure blood glucose in order to diagnose diabetes
- FPG is measured in milligrams per deciliter (*mg/dL*)
- Goal of diabetes management is to achieve FPG levels within normal range



Treatment: Metformin vs. Lifestyle Modifications

- Metformin is an oral medication used to treat high blood sugar levels caused from Type 2 Diabetes Mellitus
- Metformin controls blood sugar levels by decreasing the amount of glucose absorbed from food and made by the liver

Model	Estimated Causal Effect	Standard Error	Relative Bias Efficiency	Relative Mean Squared Error
Lasso	-48.39	0.121	7.008	0.070
Linear	-48.43	0.117	7.080	0.071
XGBoost	-47.98	0.115	6.217	0.062

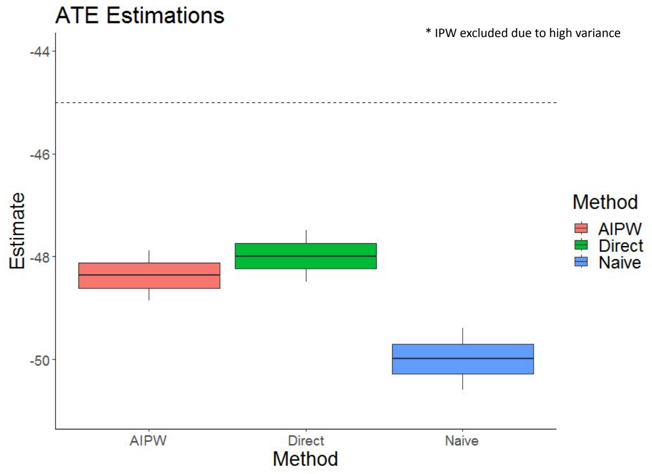
Direct Estimator

Model	Estimated Causal Effect	Standard Error	Relative Bias Efficiency	Relative Mean Squared Error
Linear-Logit	-48.43	0.123	7.082	0.071
Lasso-NNet	-48.359	0.124	6.946	0.070
Linear-Lasso	-48.412	0.123	7.048	0.071
Lasso-Logit	-48.401	0.119	7.027	0.070

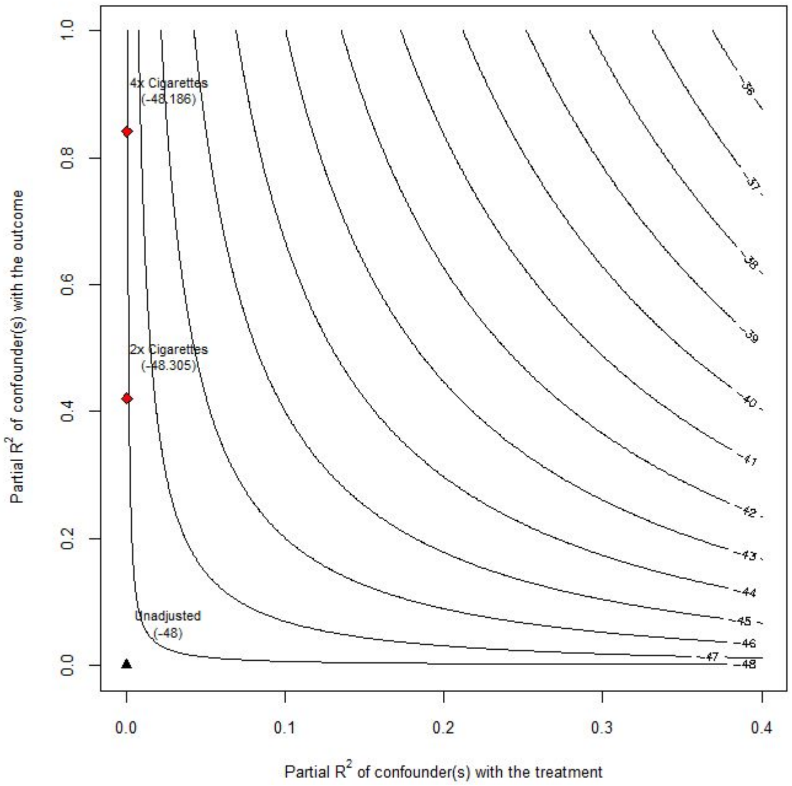
AIPW Estimator

IPW Estimator

Model	Estimated Causal Effect	Standard Error	Relative Bias Efficiency	Relative Mean Squared Error
Lasso	-49.27	5.41	8.047	0.148
Logistic	-47.26	6.40	6.180	0.146
Neural Net	-48.95	7.52	10.018	0.209



Sensitivity Analysis



Presentation Outline

Introduction

- Data
- Prediction Basics
- Causal Basics

Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

Final Takeaways

Hypertension:
Prediction and Causal Considerations

Olivia Jonokuchi, Syon Parashar, Aytijhya Saha, Christian Sanchez

University of California, Santa Barbara, Cardiff University, Indian Statistical Institute, Universidad de Guanajuato

Hypertension Data Summary

Total Number of Patients with a Yes/No answer for Hypertension: 68,720

Hypertension		No	Yes
Count (%)		34990 (50.9%)	33730 (49.1%)
Biological Sex (Female)		41.3%	52.6%
Age (Mean)		52	67
BMI (Mean)		28.2	31.8
Race	Caucasian	85.2%	87.8%
	Others	14.8%	12.3%
Affluence	1st Quartile	16.0%	19.0%
	2nd Quartile	21.0%	23.6%
	3rd Quartile	24.3%	25.2%
	4th Quartile	38.8%	32.2%
Diabetes (Yes)		9.0%	41.6%
Obesity (Yes)		22.9%	49.3%
Renal Failure (Yes)		5.3%	31.5%

18 Predictors Used for the Prediction Problem

Demographic Predictors: Biological Sex, Race, Age, Marital Status

Social Predictors: Affluence[1], Disadvantage[2], Alcohol Use Status, Illegal Drug User Status, Sexually Active Status, Cigarette Use Status

Clinical Predictors: Obesity, Diabetes, Renal Failure, Sleep Apnea, Coronary artery disease, Hyperlipidemia, Atherosclerosis, Body Mass Index (BMI)

[1] Quartile in which the average of proportion of households with income greater than $75K, proportion of population age 16+ employed in professional or managerial occupations and proportion of adults with Bachelor's Degree or higher falls under

[2] Quartile in which the average of proportion non-Hispanic Black, proportion of female headed families with children, proportion of households with public assistance income or food stamps, proportion of families with income below the federal poverty level and proportion of population age 16+ unemployed falls under

18 Predictors Used for the Prediction Problem

Demographic Predictors
Biological Sex, Race, **Age**, Marital Status

Social Predictors
Affluence[1], Disadvantage[2], **Alcohol Use Status**, Illegal Drug User Status, Sexually Active Status, Cigarette Use Status

Clinical Predictors
Obesity, Diabetes, Renal Failure, Sleep Apnea, Coronary artery disease, Hyperlipidemia, Atherosclerosis, Body Mass Index (BMI)

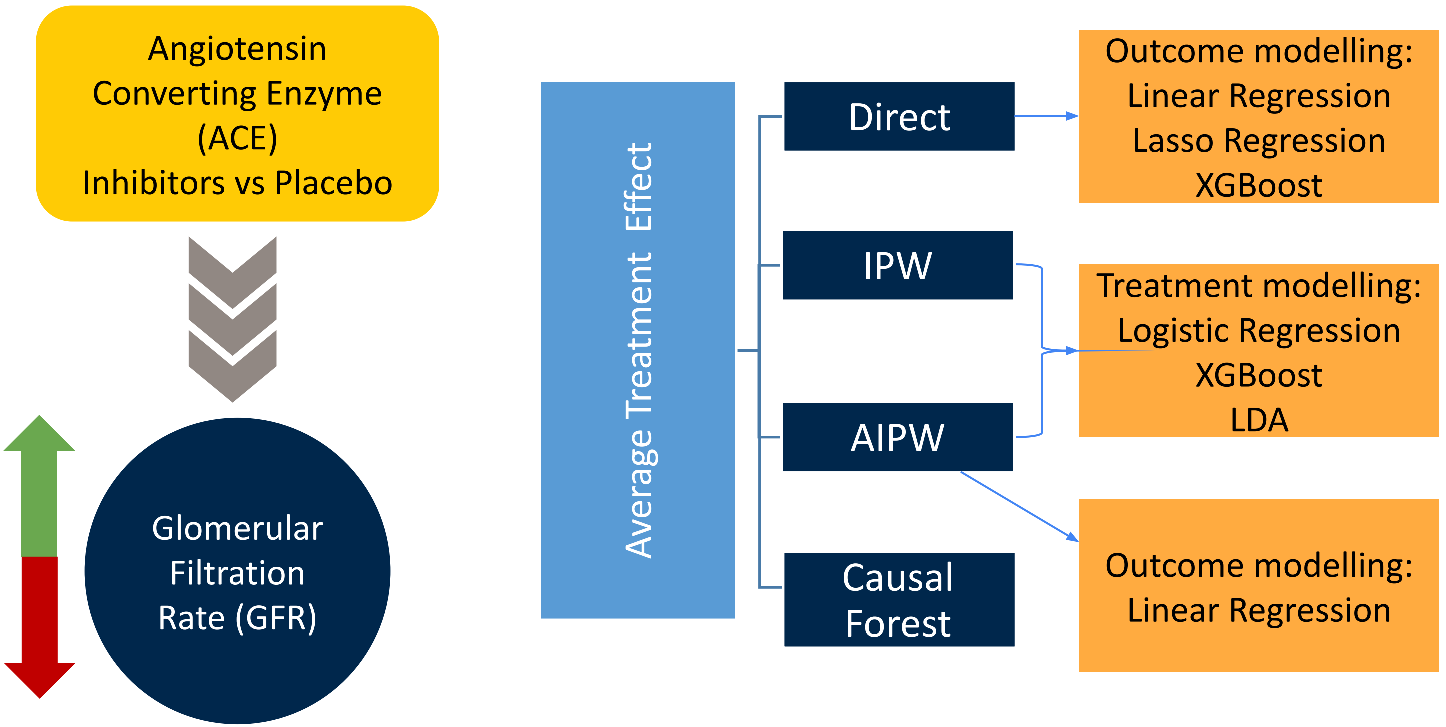
[1] Quartile in which the average of proportion of households with income greater than $75K, proportion of population age 16+ employed in professional or managerial occupations and proportion of adults with Bachelor's Degree or higher falls under
[2] Quartile in which the average of proportion non-Hispanic Black, proportion of female headed families with children, proportion of households with public assistance income or food stamps, proportion of families with income below the federal poverty level and proportion of population age 16+ unemployed falls under

Prediction Results

Classifiers	Test error	Sensitivity	Specificity	AUC
Naïve Bayes	0.2468	0.7503	0.7558	0.8212
LDA	0.2222	0.7860	0.7695	0.8612
QDA	0.2506	0.7474	0.7512	0.8197
Logistic Regression	0.2216	0.7852	0.7715	0.8634
Ridge Regression	0.2208	0.7788	0.7793	0.8632
Lasso Regression	0.2211	0.7873	0.7704	0.8634
Group Lasso	0.2207	0.7726	0.7857	0.8605
Elastic Net (α = 0.6)	0.2207	0.7870	0.7714	0.8634
Decision tree	0.2603	0.6896	0.7884	0.7864
Random Forest	0.2449	0.7396	0.7701	0.8343
XGBoost	0.2264	0.7768	0.7703	0.8538
LSVM	0.2203	0.7786	0.7807	0.8627
Super Learner[1]	0.2191	0.7784	0.7833	0.8643
Neural Net	0.2195	0.7829	0.7780	0.8640

[1] We used XGBoost, Random Forest, and GLMNet to run the Super Learner classifier

The Causal Problem

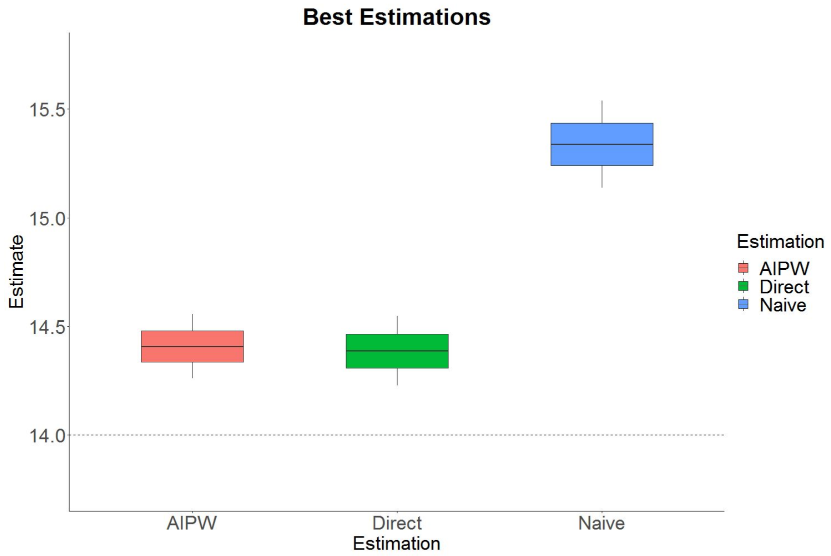


Same Covariates Used for the Causal Problem

Demographic Predictors	Biological Sex, Race, Age, Marital Status
Social Predictors	Affluence , Disadvantage[1], Alcohol Use Status, Illegal Drug User Status, Sexually Active Status, Cigarette Use Status
Clinical Predictors	Obesity, Diabetes, Renal Failure, Sleep Apnea, Coronary artery disease, Hyperlipidemia, Atherosclerosis, Body Mass Index (BMI)

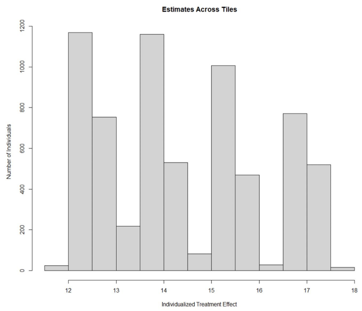
[1] Quartile in which the average of proportion non-Hispanic Black, proportion of female headed families with children, proportion of households with public assistance income or food stamps, proportion of families with income below the federal poverty level and proportion of population age 16+ unemployed falls under

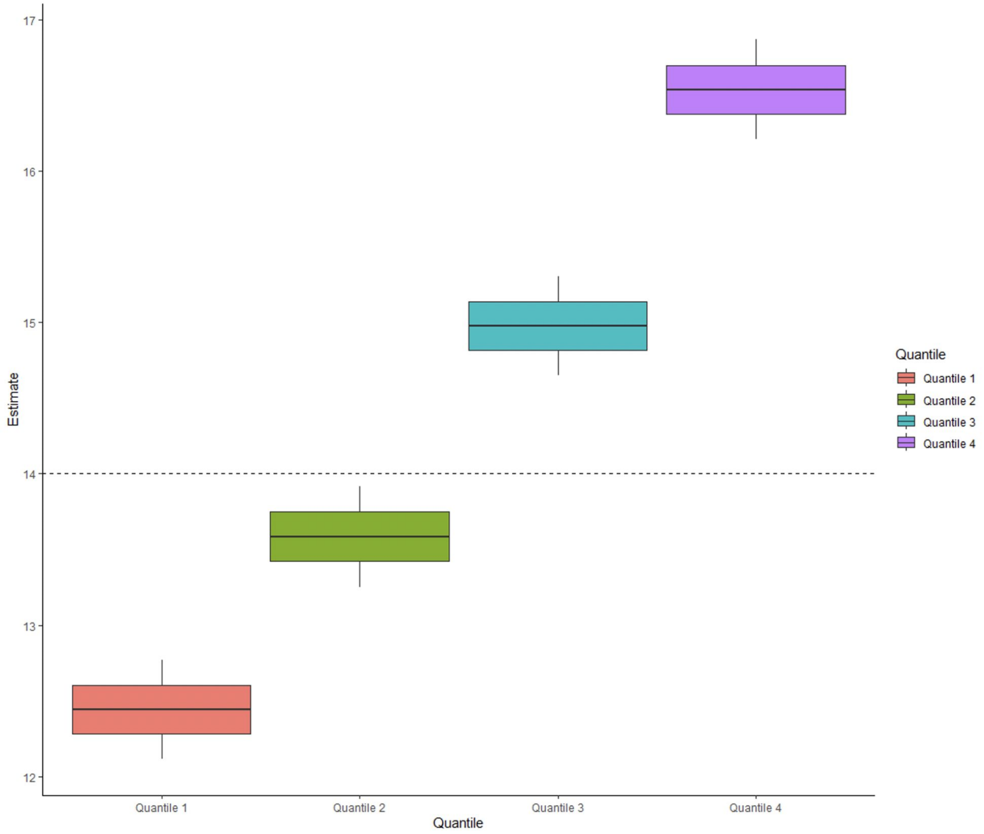
Estimator Results



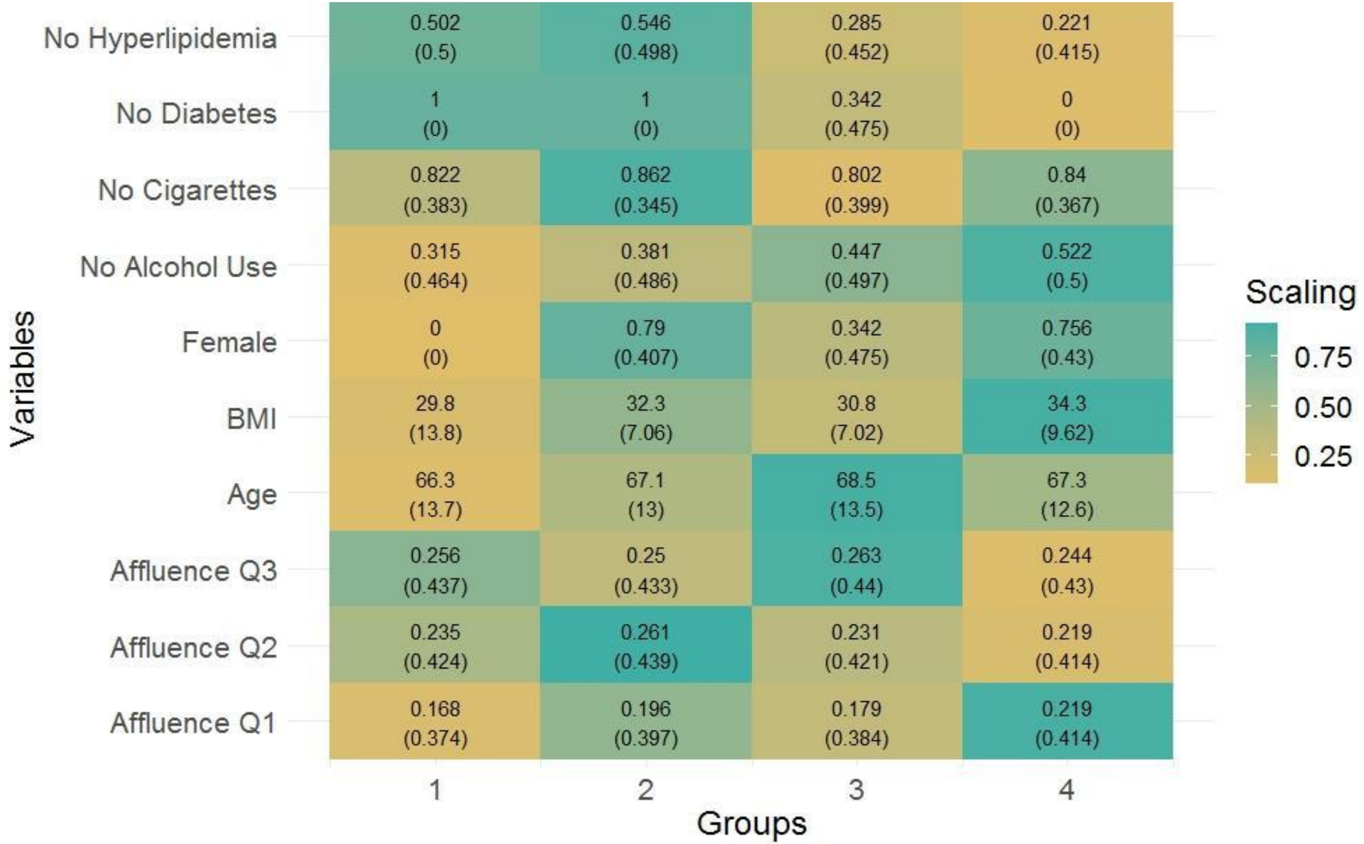
Estimator	Best Model	Relative Bias (%)	Relative Root Mean Square Error
Direct Estimation	Lasso Regression	2.76	0.0277
AIPW	Linear Regression - LDA	2.94	0.0295
Naive	–	9.56	0.0956

Average Treatment Effect Estimates

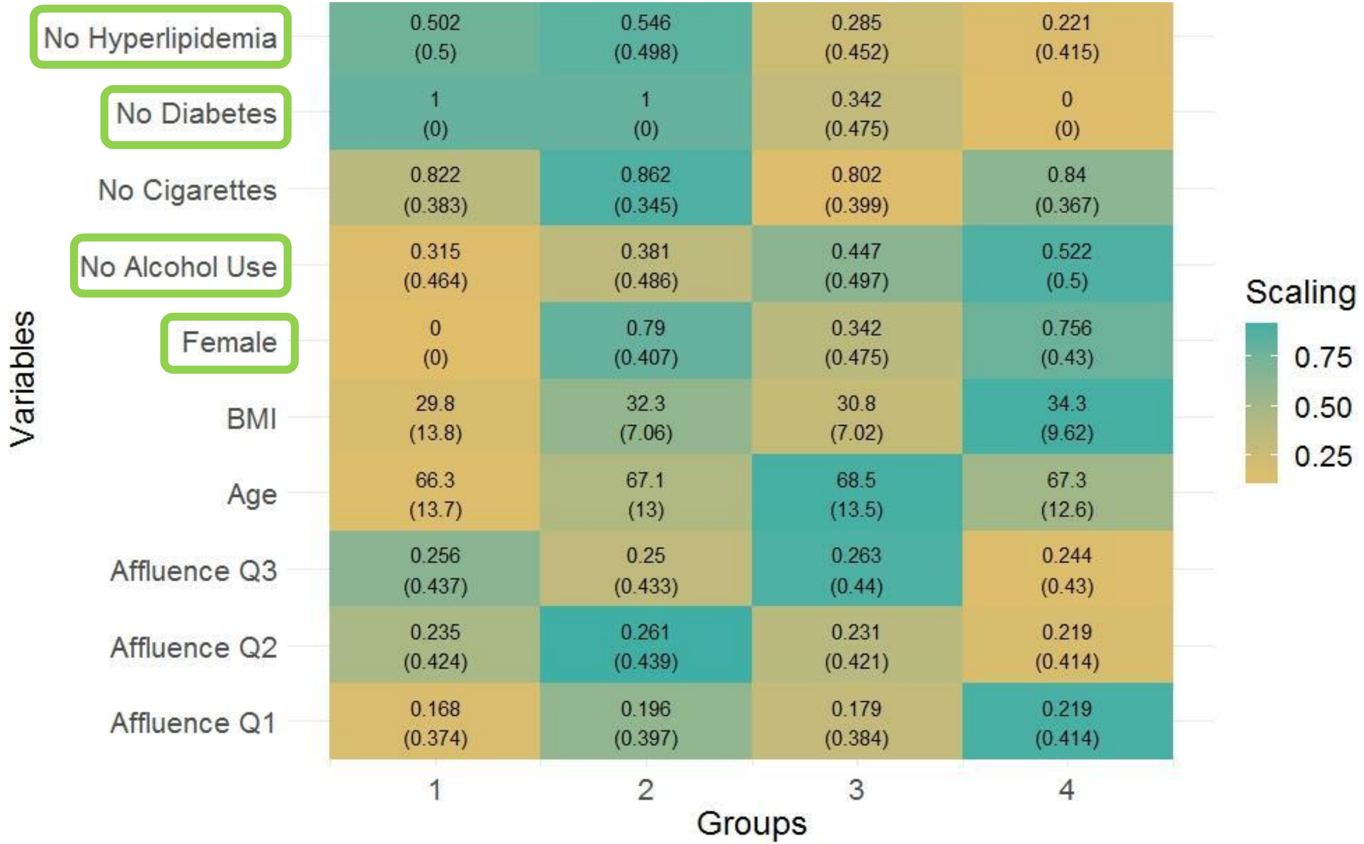




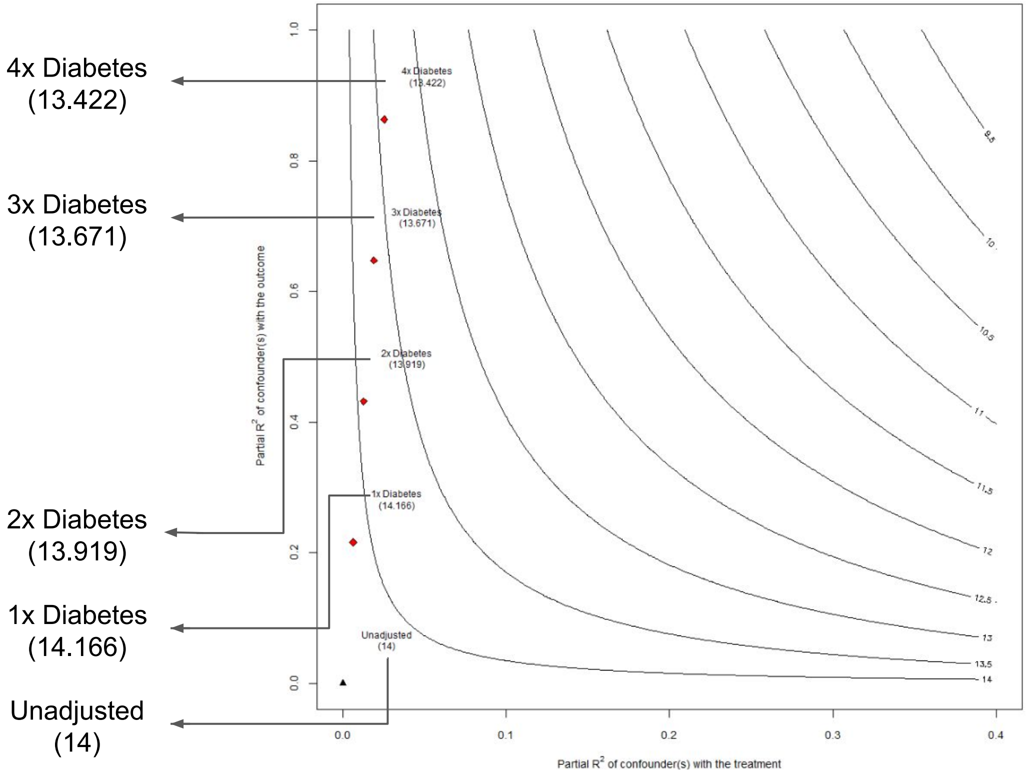
HTE: Average Covariate Values



HTE: Average Covariate Values



Sensitivity Analysis



Presentation Outline

Introduction

- Data
- Prediction Basics
- Causal Basics

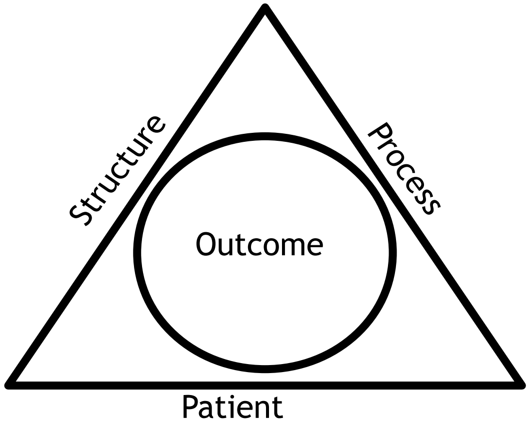
Disease Subgroups

1. Depression
2. Diabetes
3. Hypertension

Final Takeaways

Final Group Takeaways

1. **Generalizability**: Demographics in the MGI dataset are not representative of the US population or even the Michigan state population.



Final Group Takeaways

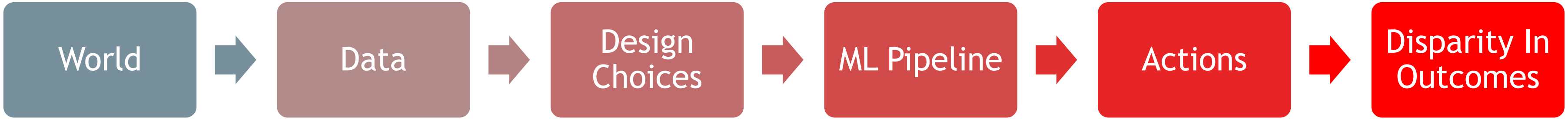
2. **Beware of Bias:** introducing and enforcing biases in data and models generates results that don’t accurately represent the population. ML models with more interpretability provide more insights of biases going into and out of the model.

Garbage in, Garbage Out

<u>Most Relevant Biases</u>

- Selection
- Omitted Variable
- Measurement
- Confounding
- Observational
- Funding
- Non-response
- Omitted Variable
- Assignment

Sources of Bias and Disparity



Final Group Takeaways

3. **Variable Selection:** methods included literature review, backward elimination, forward selection with varying degrees of preventing overfitting

4. **Multicollinearity**: A strong correspondence (linear combination) between two or more explanatory variables. Regression coefficients are indeterminate and their standard errors are not defined

Tip: Identifying high correlation does **NOT** always identify the source of MC

Problems

- Wider confidence intervals
- Decreased statistical power
- Exclusion of significant predictors
- Skewed or misleading results (inaccurate parameter estimates)

Solutions

- Ridge
- Lasso

Final Group Takeaways

5. **Strengths and Weakness of ML models:**

	Parametric	Non-Parametric
Benefits	- Easier to understand, increased interpretability - Usually very fast, less data is required	- Flexibility with no assumptions of the underlying function - Can result in higher performance models for prediction
Drawbacks	- Limited model complexity could result in poor fit	-Requires a lot more data and slower -higher risk of overfitting

Acknowledgements

We would like to give a big thanks Rahul Ladhania (PI) and Ritoban Kundu (GSI) for all of their guidance and support throughout this research project.

We also want to thank Bhramar Mukherjee, Sabrina Olsson, Hanna Venera, Youqi Yang, and all other BDSI faculty and staff that helped made this summer institute possible.

References

Brown, S. (2021, April 21). *Machine Learning, Explained*. MIT Sloan. mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Centers for Disease Control and Prevention. (2021, May 18). *High blood pressure symptoms and causes.* Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/about.htm

Centers for Disease Control and Prevention. (2023, March 17). *Know your risk for high blood pressure.* Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/risk_factors.htm

Cleveland Clinic Medical (n.d.). *High blood pressure: What you need to know.* Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/4314-hypertension-high-blood-pressure

Mayo Foundation for Medical Education and Research. (n.d.). *High blood pressure (hypertension).* Mayo Clinic. https://mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

Depression statistics. Depression and Bipolar Support Alliance. (2019, July 12). https://www.dbsalliance.org/education/depression/statistics/

Goodwin, R. D., Dierker, L.C., Wu, M., Galea, S., Hoven, C.W., & Weinberger, A.H. (2022). Trends in US Depression Prevalence From 2015-2020: The Widening Treatment Gap. American journal of preventive medicine, 63(5):726-733. https://doi.org/10.1016/j.ampere.2022.05.014

Questions?