Sentinel

Comparative evaluation of automated approaches for confounder selection in ultra high-dimensional data

Leveraging unstructured electronic health records for large-scale confounding control in real-world evidence studies

Richard Wyss, PhD, MSc

BWH

DoPE BWH

Disclosures and Funding

- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).

- The authors report no conflicts of interest

Background

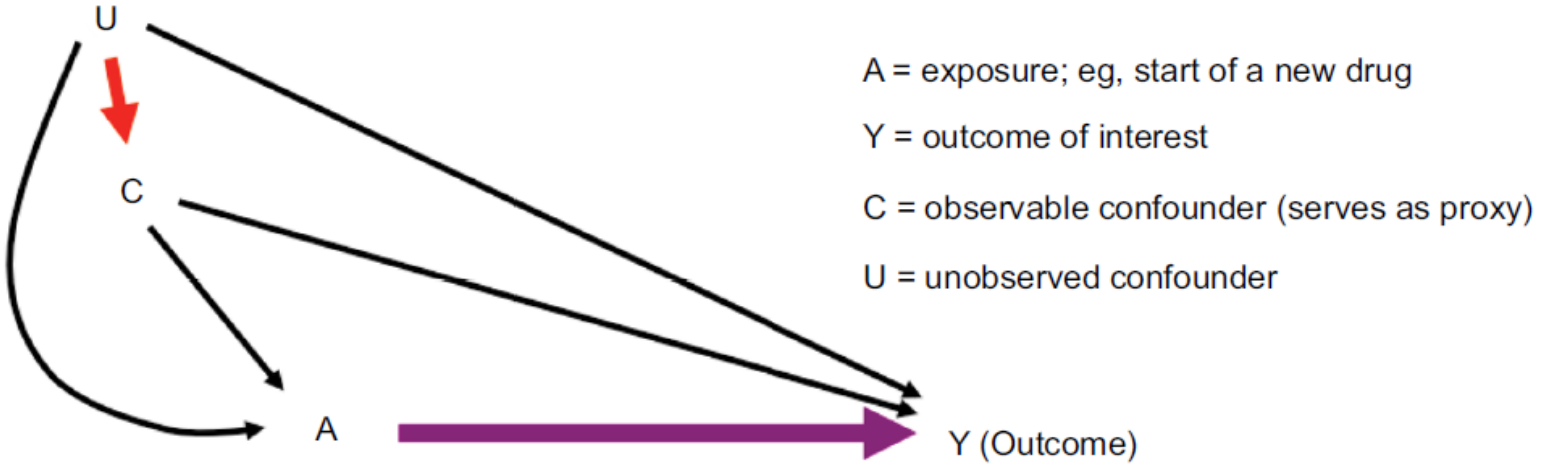
Background: Challenges for Confounding Control in RWE Studies

- Confounding arising from non-randomized treatment choices remains a fundamental challenge for extracting valid evidence to help guide treatment and regulatory decisions.

- Standard tools for confounding adjustment have typically relied on adjusting for a limited number of investigator specified variables.
  - Adjusting for investigator-specified variables alone is often inadequate
    - Some confounders are unknown at the time of drug approval
    - Many confounders are not directly measured in routine-care databases.

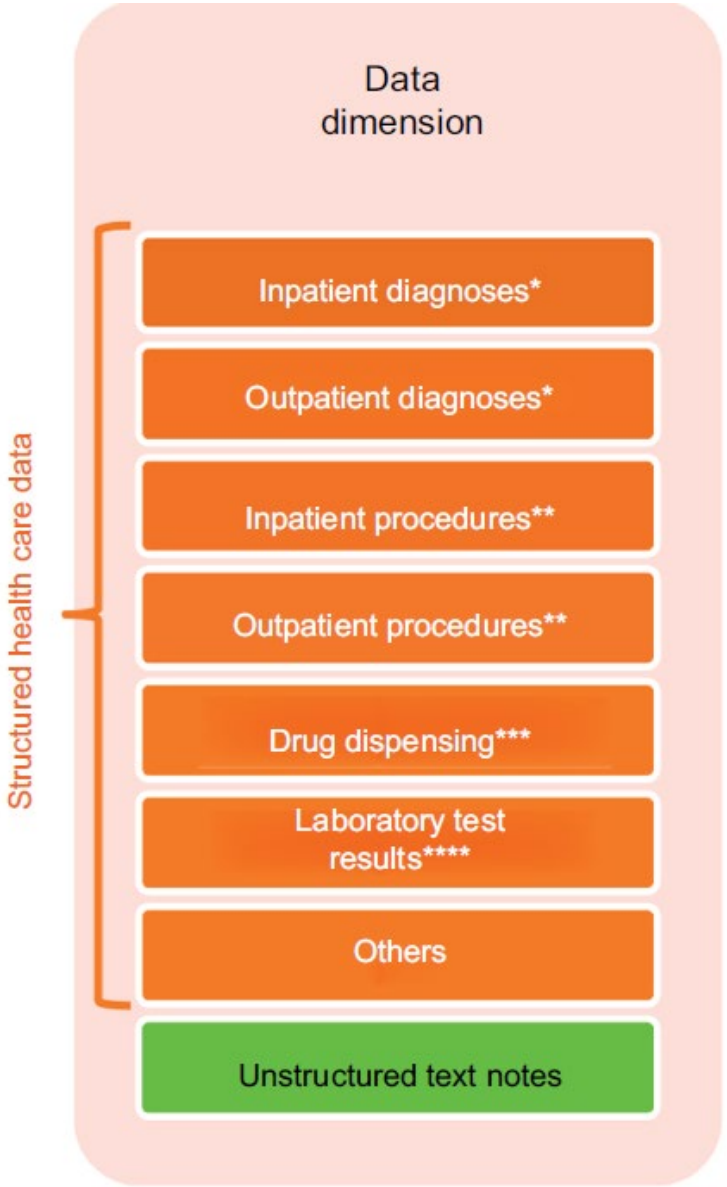
Background: Proxy Confounder Adjustment

- Healthcare databases may be understood and analyzed as a high-dimensional set of "proxy" factors that indirectly describe the health status of patients (Schneeweiss 2009, 2017).



Unobserved confounder	Observable proxy measurement	Coding examples
Very frail health	Use of oxygen canister	CPT-4
Sick but not critical	Code for hypertension during a hospital stay	ICD-9, ICD-10
Health-seeking behavior	Regular check-up visit; regular screening examinations	ICD-9, CPT-4, #PCP visits

Background: High-Dimensional Proxy Confounder Adjustment



- How to identify/generate proxy variables for adjustment?
  - High-dimensional propensity score (Schneeweiss 2009)
    - Does not require data pre-processing
  - OMOP approach:
    - Pre-process data into a common data model then use machine learning algorithms for variable selection (e.g., Lasso)
- **Current approaches for generating proxy variables for confounder adjustment do not leverage information from unstructured EHR text notes.**

Background: Leveraging Unstructured Electronic Health Records for Large-Scale Proxy Adjustment.

- NLP tools turn free-text notes from EHR data into structured features that can supplement confounding adjustment.
  - However, traditional applications are difficult to scale for large-scale proxy adjustment.

- **Project Objective (use of NLP-generated information from unstructured data):** To explore if unsupervised NLP can be used to generate high-dimensional sets of features from free-text notes for improved large-scale proxy confounding control
  - **Aim 1:** To use scalable applications of NLP to generate structured features from high-dimensional data for large-scale proxy adjustment.
    - leverages work from RO1 (Josh Lin, PI; Richie Wyss, Co-investigator; Sebastian Schneeweiss, Co-investigator)
  - **Aim 2: To better understand what machine learning tools for confounder selection perform well for large-scale proxy adjustment in ultra high-dimensional RWE studies.**

Methods

Methods: Data Source for Generating Cohort Studies

- Mass General Brigham (MGB) Research Patient Data Registry (RPDR)
  - The electronic health records (EHR) of all the patients aged 65 and above identified in the Mass General Brigham (MGB) Research Patient Data Registry (RPDR) were linked to Medicare claims data

- Linked RPDR-Medicare claims were used to generate 3 cohort studies comparing different classes of medications (details on later slide).
  - Purpose: case studies for evaluating and testing various methods for NLP feature generation for ultra high-dimensional proxy confounder adjustment.

Methods: Using NLP to Generate Structured Features.

- We used ‘bag-of-words’ to generate features for the top 20,000 most prevalent terms from free-text notes.
  - Very common, simple, and flexible NLP approach
  - Measures the frequency (occurrence) of words within a document
    - Order and structure of words in the document is discarded.
    - The model is only concerned with whether words occur in the document, not where in the document or in relation to other words
- Each word count is then a feature that can be used for modeling

Methods: Study Cohorts

Table 1. Study Cohorts

		Total N			# Baseline Covariates		
No.	Description	Study Population	Treatment (%)	Outcome (%)	Investigator Specified	Claims Codes	EHR features
1.	High vs low intensity statin with an outcome of major cardiac events	3,529	1,244 (35.3)	138 (3.9)	39	18,409	20,017
2.	Oral anti-coagulants vs non-use with an outcome of stroke and major bleeding	9,571	5,991 (62.6)	158 (1.7)	39	19,517	20,051
3.	High vs. low dose PPI with an outcome of peptic ulcer complications	20,862	7,108 (34.1)	234 (1.1)	39	28,041	20,025

Methods: How to best identify confounder information in ultra high-dimensional real-world data?

- Predictive performance did not improve when modeling the outcome, but does this mean that there is no additional confounder information in EHR generated variables?

- Begin by considering various methods for confounder selection
  - Focus on lasso-based approaches
    - Regular Lasso
    - Outcome adaptive lasso
    - Collaborative controlled lasso
    - Outcome highly-adaptive lasso

Methods: How to make objective decisions on which modeling approach is best?

- **Cannot use actual study with estimated effects to make modeling decisions**
- Recent papers have proposed using synthetic control studies to help assess validity of alternative causal inference models and tailor analyses to the given study (Alaa & Van Der Scharr 2019; Schuler et al. 2017; Athey S et al. 2019; Bahamyirou A., et al. 2018; Schuemie MJ, et al. 2018; Petersen et al. 2012)
  - Provides an objective assessment of validity and model selection.
  - A common theme is that they use a variation of 'plasmode simulation' (Franklin et al. 2014).

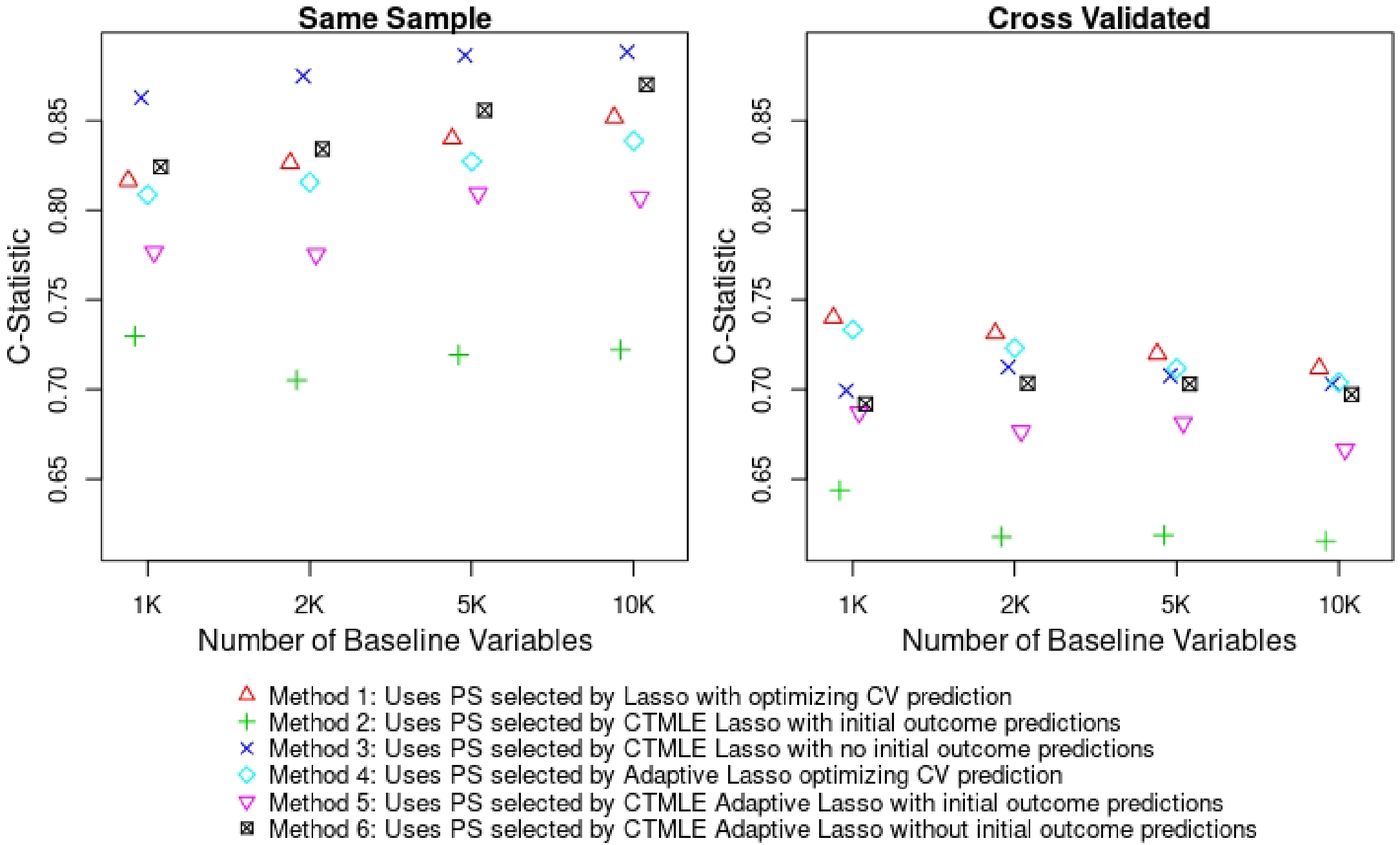
Variation of the parametric bootstrap where we bootstrap from the original study population, but simulate some aspects of the data structure while leaving other features of the data unchanged.
Typically, we set the outcome data aside (outcome blind data), then simulate the outcome while leaving baseline covariates and treatment status unchanged.
Try to generate synthetic control outcomes (and treatment) that mimic as closely as possible the observed confounding structure in the study cohort.
Will be inexact, but close approximations can be useful for testing robustness and validity of causal inference methods for the study at hand.

Confounder Selection & Propensity Score Models	
Lasso PS Models	Description
Standard Lasso	Lasso modeling treatment assignment with penalty factor (lambda) that optimizes CV treatment prediction
CTMLE Lasso w/ predictions	Collaborative controlled lasso—Lasso modeling treatment assignment but uses ctmle to choose penalty factor. We include initial predictions for the counterfactual outcomes using an outcome lasso model.
CTMLE Lasso w/ no predictions	Collaborative controlled lasso—Lasso modeling treatment assignment but uses ctmle to choose penalty factor. We did not include initial predictions for the counterfactual outcomes (only included treatment in the initial outcome model).
Outcome Adaptive Lasso (OAL)	Adaptive lasso modeling treatment assignment with a penalty factor set by user. We assigned a penalty of 0 for all variables selected by the outcome lasso and a penalty of 1 for all other variables (i.e., we forced variables selected by outcome lasso into the lasso model for treatment).
CTMLE OAL w/ predictions	Collaborative controlled outcome adaptive lasso with initial predictions for the counterfactual outcomes
CTMLE OAL w/ no predictions	Collaborative controlled outcome adaptive lasso with no initial predictions for the counterfactual outcomes (initial outcome model includes only treatment)

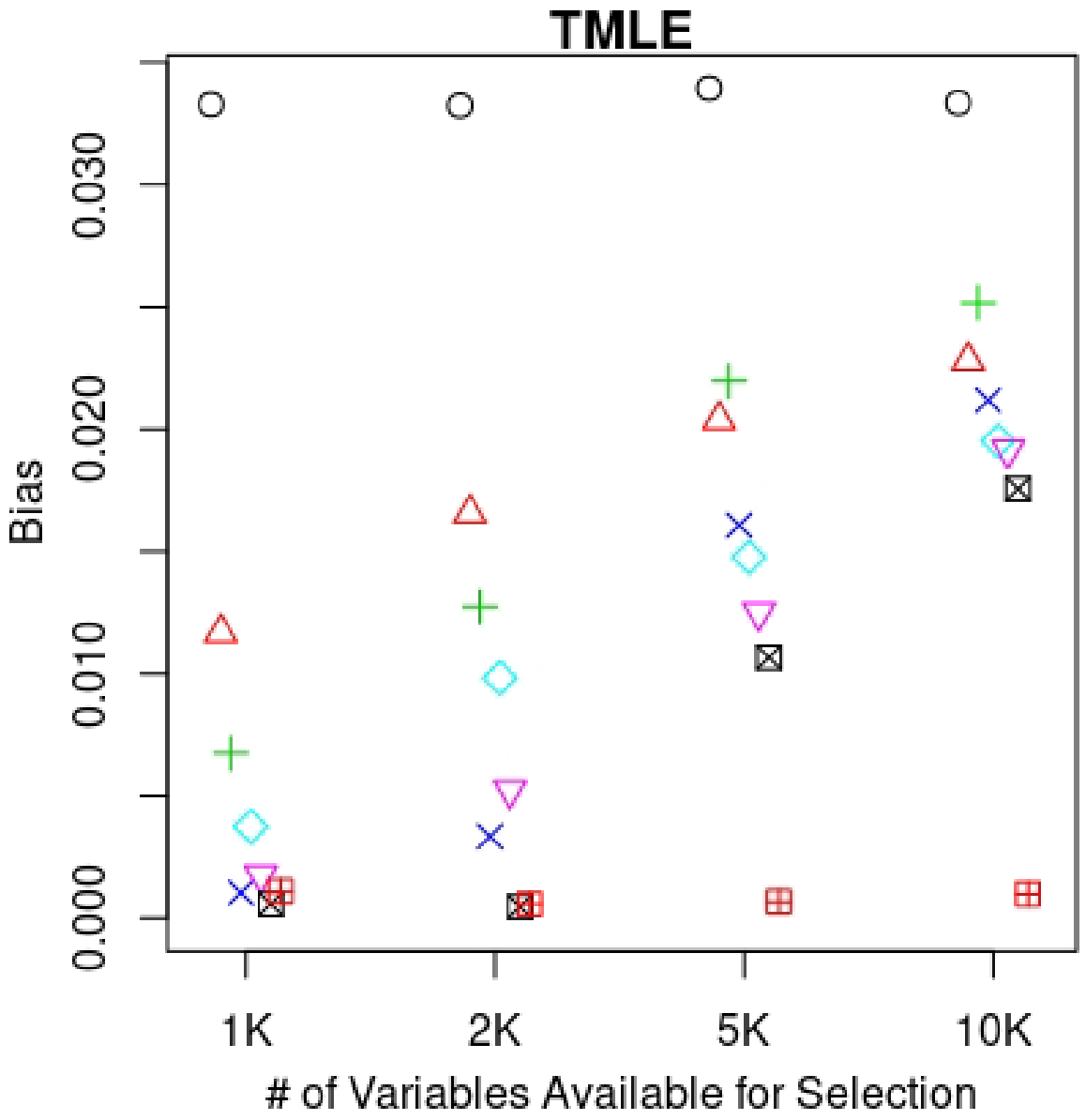
- **For each PS model, we estimated the treatment effect using Targeted Maximum Likelihood Estimation (TMLE) that included initial predictions from an outcome lasso model and PS weighting**

Simulation Results

Selected Simulation Results for Prediction


Same Sample
Cross Validated
C-Statistic
0.65
0.70
0.75
0.80
0.85
1K
2K
5K
10K
Number of Baseline Variables
Method 1: Uses PS selected by Lasso with optimizing CV prediction
Method 2: Uses PS selected by CTMLE Lasso with initial outcome predictions
Method 3: Uses PS selected by CTMLE Lasso with no initial outcome predictions
Method 4: Uses PS selected by Adaptive Lasso optimizing CV prediction
Method 5: Uses PS selected by CTMLE Adaptive Lasso with initial outcome predictions
Method 6: Uses PS selected by CTMLE Adaptive Lasso without initial outcome predictions

Selected Simulation Results for Bias



Lambda Selection for Lasso PS Model

- Unadjusted
- PS Model 1: Traditional Lasso
- PS Model 2: CTMLE Lasso with predictions
- PS Model 3: CTMLE Lasso no predictions
- PS Model 4: Outcome Adaptive Lasso (OAL)
- PS Model 5: CTMLE OAL with predictions
- PS Model 6: CTMLE OAL no predictions
- Oracle: includes all confounders

General points for discussion

- Selecting models based on collaborative learning improved bias reduction even though predictive performance declined.
  - Outcome adaptive lasso with collaborative selection generally performed best.
  - Some degree of overfitting is beneficial for confounding control when using Machine Learning to data-adaptively select (model) high-dimensional sets of variables

- Bias increased as the number of spurious variables available for selection increased.
- Bias can result from two sources
  1. Lasso model not selecting confounding variables
  2. Even when lasso selects confounders there can still be regularization bias (Chernozhukov 2018).

- Use relaxed lasso to reduce regularization bias in sparse high-dimensional data (Meinshausen 2007).

Relaxed lasso

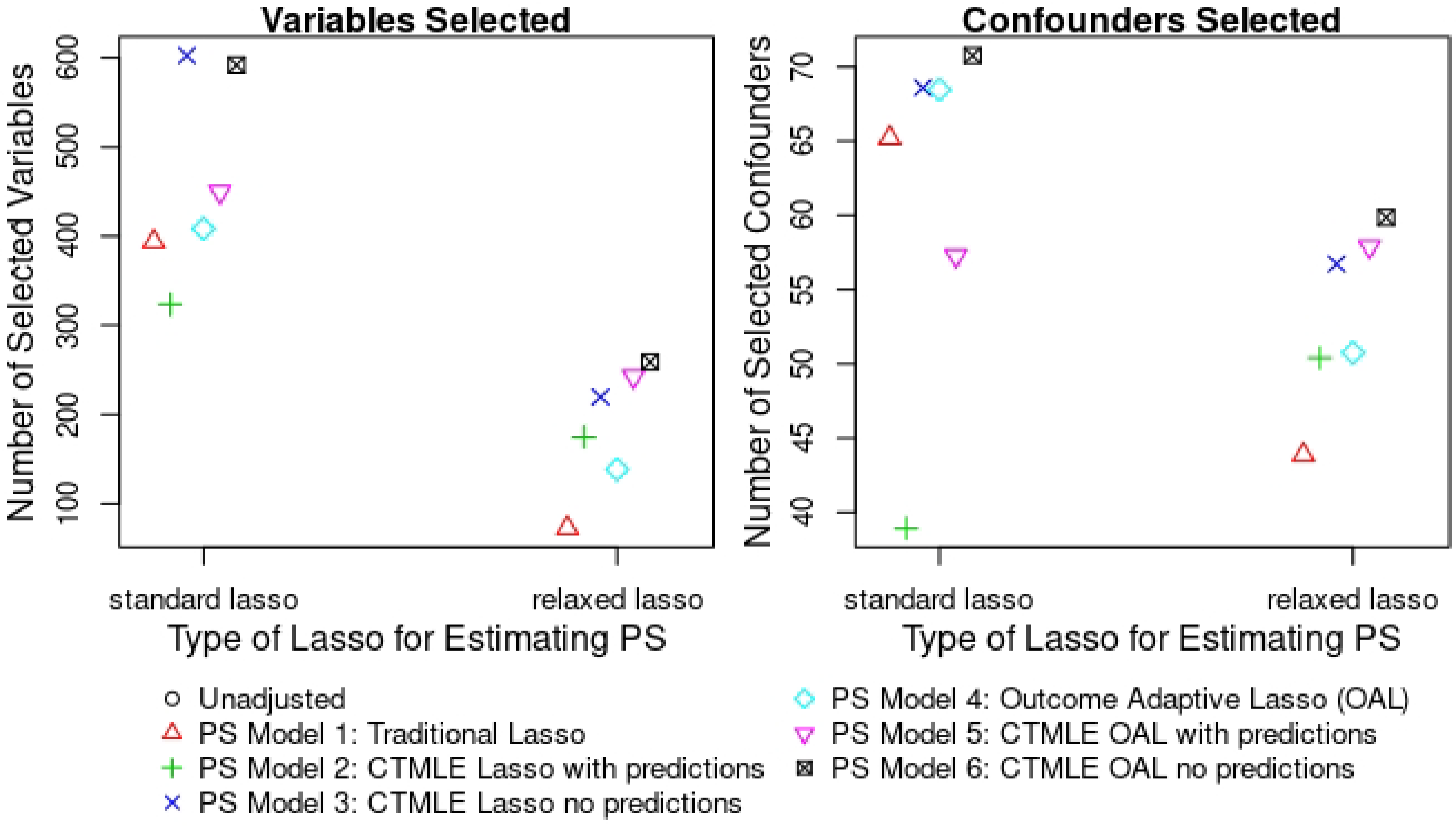
Use relaxed lasso to reduce regularization bias (Meinshausen 2007).

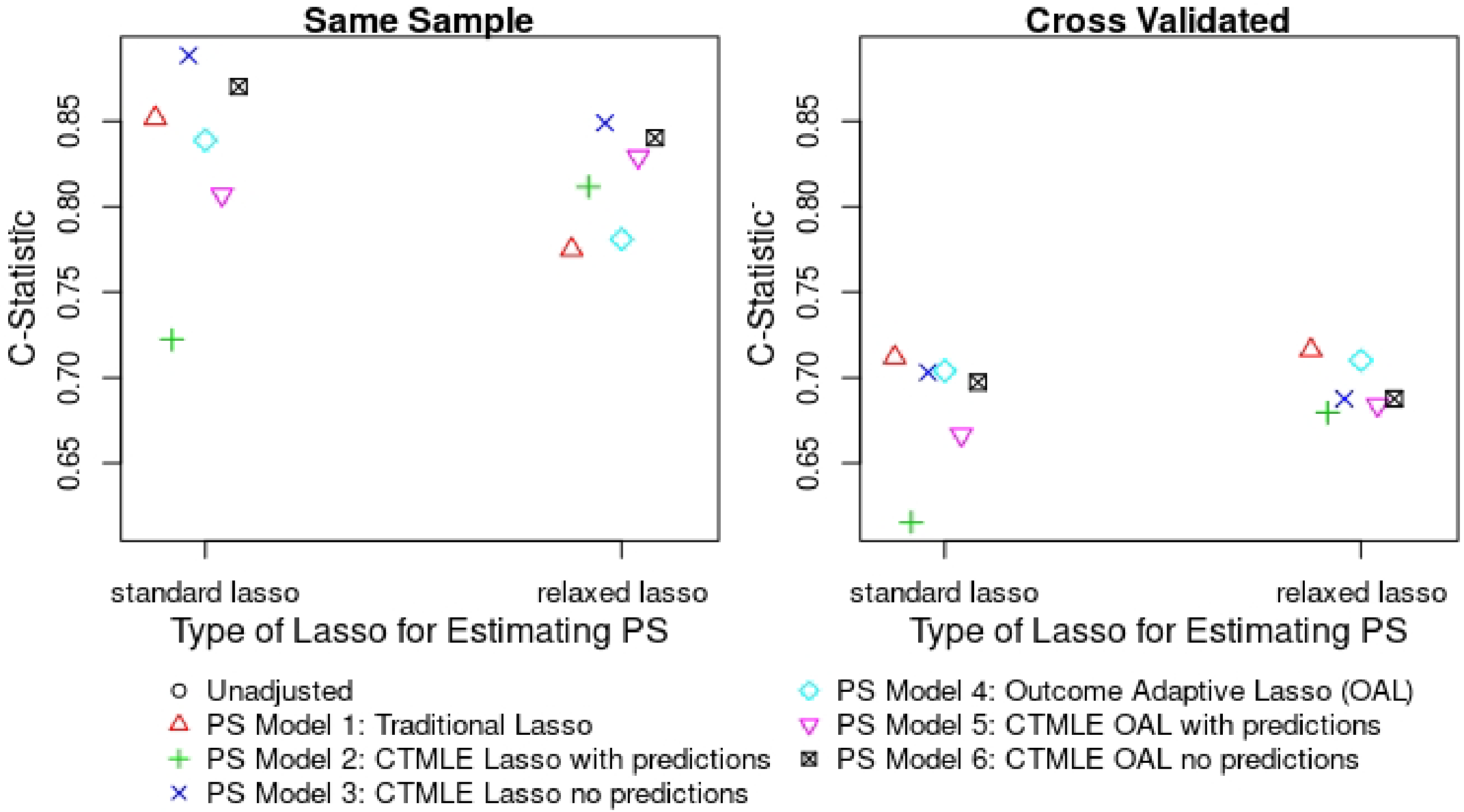
- Runs regularized regression twice:
  1. First runs lasso to select lambdas to control variable selection (which variables are selected for each lambda);
  2. Second step runs regularized regression again for each set of variables selected by each lambda with less penalization to control shrinkage level of coefficients. The shrinkage penalization in the second step can be selected using Cross Validation.

- *'Idea of the relaxed lasso is to take the lasso fitted object and then for each lambda, refit the variables in the active set with either no penalization or less penalization. This gives the "relaxed" fit'*. (Hastie & Tibshirani 2021)

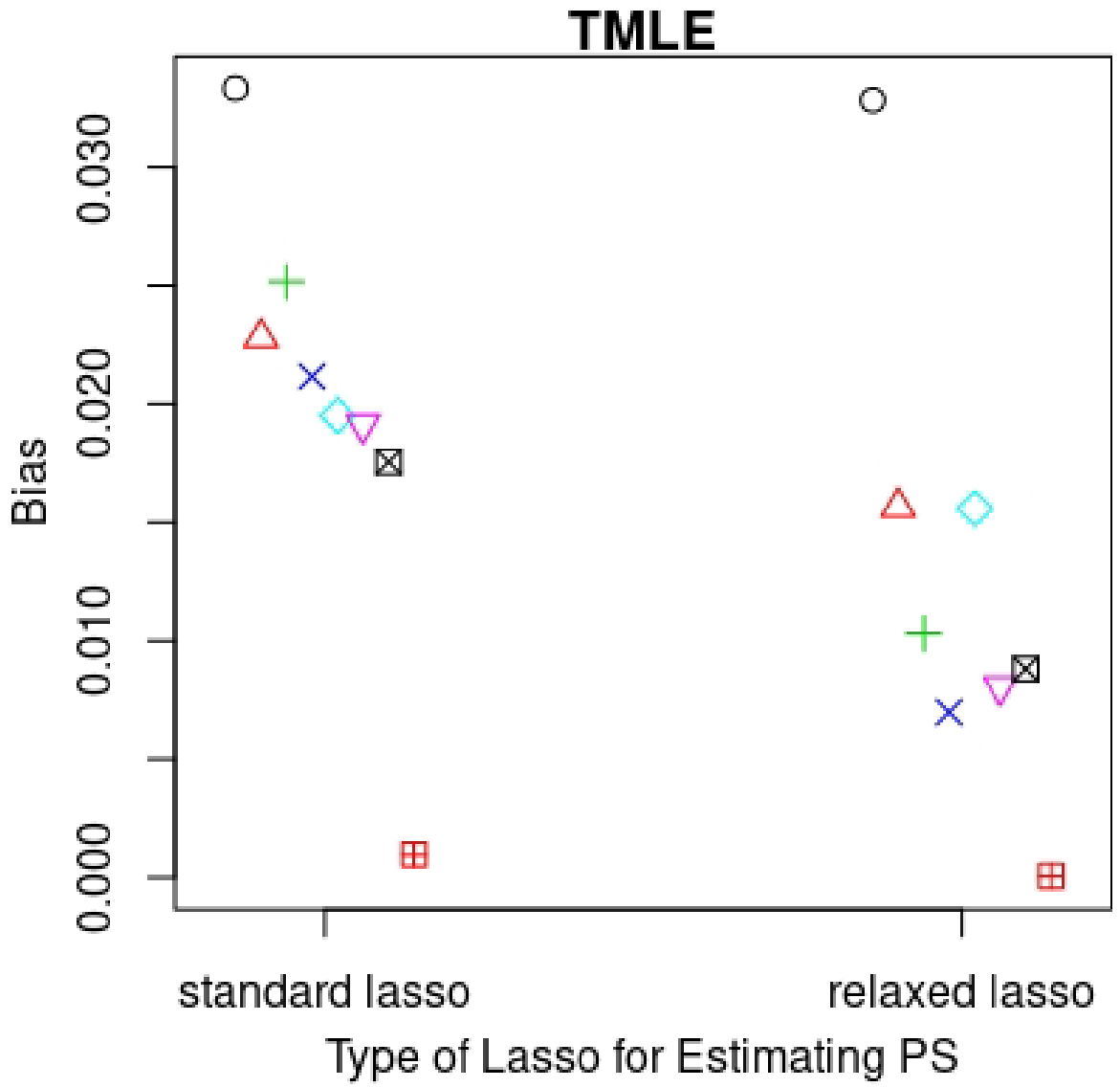
- Relaxed lasso can often improve predictive performance by fitting more parsimonious models with less penalization in sparse high-dimensional data (Meinhausen 2007).

Selected Simulation Results for Variable Selection and Prediction with Relaxed Lasso


Variables Selected
Number of Selected Variables
100
200
300
400
500
600
standard lasso
relaxed lasso
Type of Lasso for Estimating PS
Confounders Selected
Number of Selected Confounders
40
45
50
55
60
65
70
standard lasso
relaxed lasso
Type of Lasso for Estimating PS
Unadjusted
PS Model 1: Traditional Lasso
PS Model 2: CTMLE Lasso with predictions
PS Model 3: CTMLE Lasso no predictions
PS Model 4: Outcome Adaptive Lasso (OAL)
PS Model 5: CTMLE OAL with predictions
PS Model 6: CTMLE OAL no predictions


Same Sample
Cross Validated
C-Statistic
0.65
0.70
0.75
0.80
0.85
standard lasso
relaxed lasso
Type of Lasso for Estimating PS
Unadjusted
PS Model 1: Traditional Lasso
PS Model 2: CTMLE Lasso with predictions
PS Model 3: CTMLE Lasso no predictions
PS Model 4: Outcome Adaptive Lasso (OAL)
PS Model 5: CTMLE OAL with predictions
PS Model 6: CTMLE OAL no predictions

Selected Simulation Results for Bias with Relaxed Lasso



- Unadjusted
- PS Model 1: Traditional Lasso
- PS Model 2: CTMLE Lasso with predictions
- PS Model 3: CTMLE Lasso no predictions
- PS Model 4: Outcome Adaptive Lasso (OAL)
- PS Model 5: CTMLE OAL with predictions
- PS Model 6: CTMLE OAL no predictions
- Oracle: includes all confounders

Discussion

General Points for Discussion after running 'relaxed' lasso

- Relaxed lasso reduced bias in effect estimate compared with standard lasso

- Selecting models based on collaborative learning still improved bias reduction at the expense of predictive performance.
  - Outcome adaptive lasso with collaborative selection generally performed best.
  - Some degree of overfitting is beneficial for confounding control when using Machine Learning to data-adaptively select (model) high-dimensional sets of variables

- Still some bias with large numbers of variables
  - May need large samples to use ML to identify confounders in sparse high-dimensional data.

Future work/next step is to apply top performing models from simulations to empirical studies

Research team

Food and Drug Administration
- Hana Lee, PhD, MS
- Sarah Dutcher, PhD, MS

Brigham and Women's Hospital
- Richard Wyss, PhD, MS
- Rishi Desai, PharmD, PhD
- Josh Lin, MD, PhD
- Shirley Wang, PhD, MS
- Yinzhu Jin, MS, MPH
- Shamika More, MS
- Luke Zabotka, BA

DPM/HPHCI/Operations Center
- Darren Toh, ScD

Kaiser Washington/University of Washington
- Jennifer Nelson, PhD

University of California, Berkeley
- Mark van der Laan, PhD
- Lars van der Laan

University of Michigan
- Xu Shi, PhD

Putnam Data Sciences
- Susan Gruber, PhD, MS, MPH

Thank you