Methodological Medical Research while Integrating ML in Causal Inference Problems

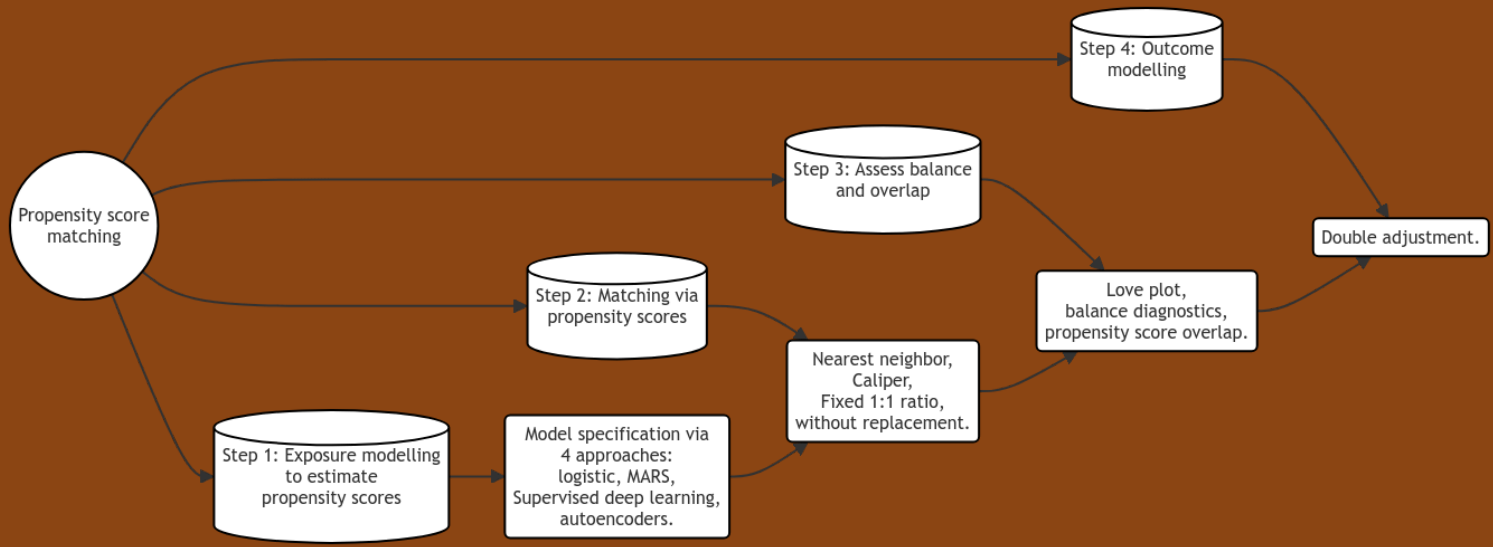
Ehsan Karim; ehsan.karim@ubc.ca

SPPH, UBC

Example Research Topics

1. Deep Learning
    - Supervised Deep Learning Architecture vs. Autoencoders
2. Double Crossfitting (DCF) Guideline
    - How many folds/splits
    - How many repetitions?
3. Residual Confounding Bias Reduction
    - High Dimensional Propensity Score (hdPS) vs ML extensions
4. Enhancing hdPS
    - Original hdPS
    - Pure ML
    - TMLE with no ML
    - TMLE (and choice of Super Learner)
    - TMLE with DCF (and choice of Super Learner)

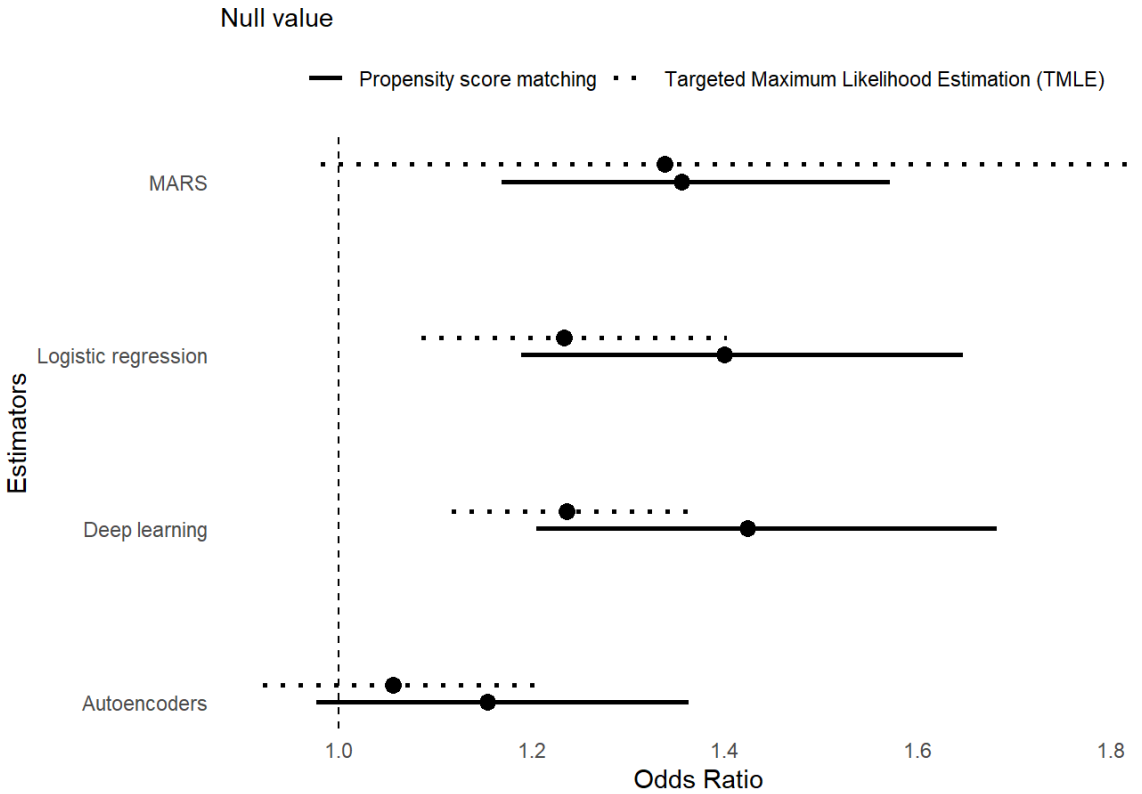
(1) Supervised Deep Learning Architecture vs. Autoencoders



PS is estimated with 4 different methods.

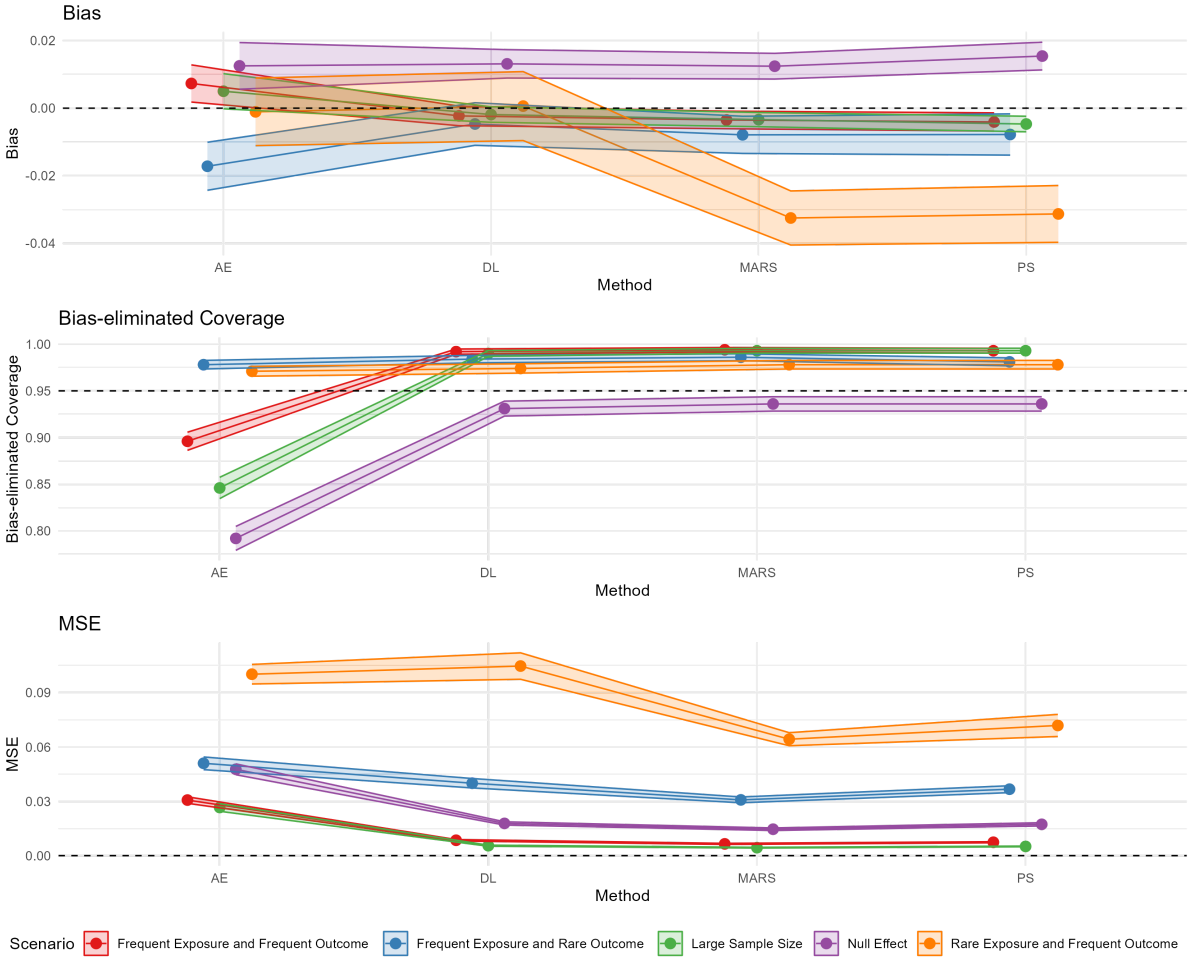
(Weberpals, Becker, Davies, Schmich, Rüttinger, Theis, and Bauer-Mehren, 2021)

(1) Empirical Analysis with RHC data



RHC analysis with PS and TMLE analyses, when PS is estimated with 4 different methods.

(1) Simulation studies



Simulation results, when PS is estimated with 4 different methods.

(2) Double Crossfitting Guideline

RESEARCH-ARTICLE

Finding the Optimal Number of Splits and Repetitions in Double Cross-Fitting Targeted Maximum Likelihood Estimators

Mohammad Ehsanul Karim[*1,2] | **Momenul Haque Mondol**[1]

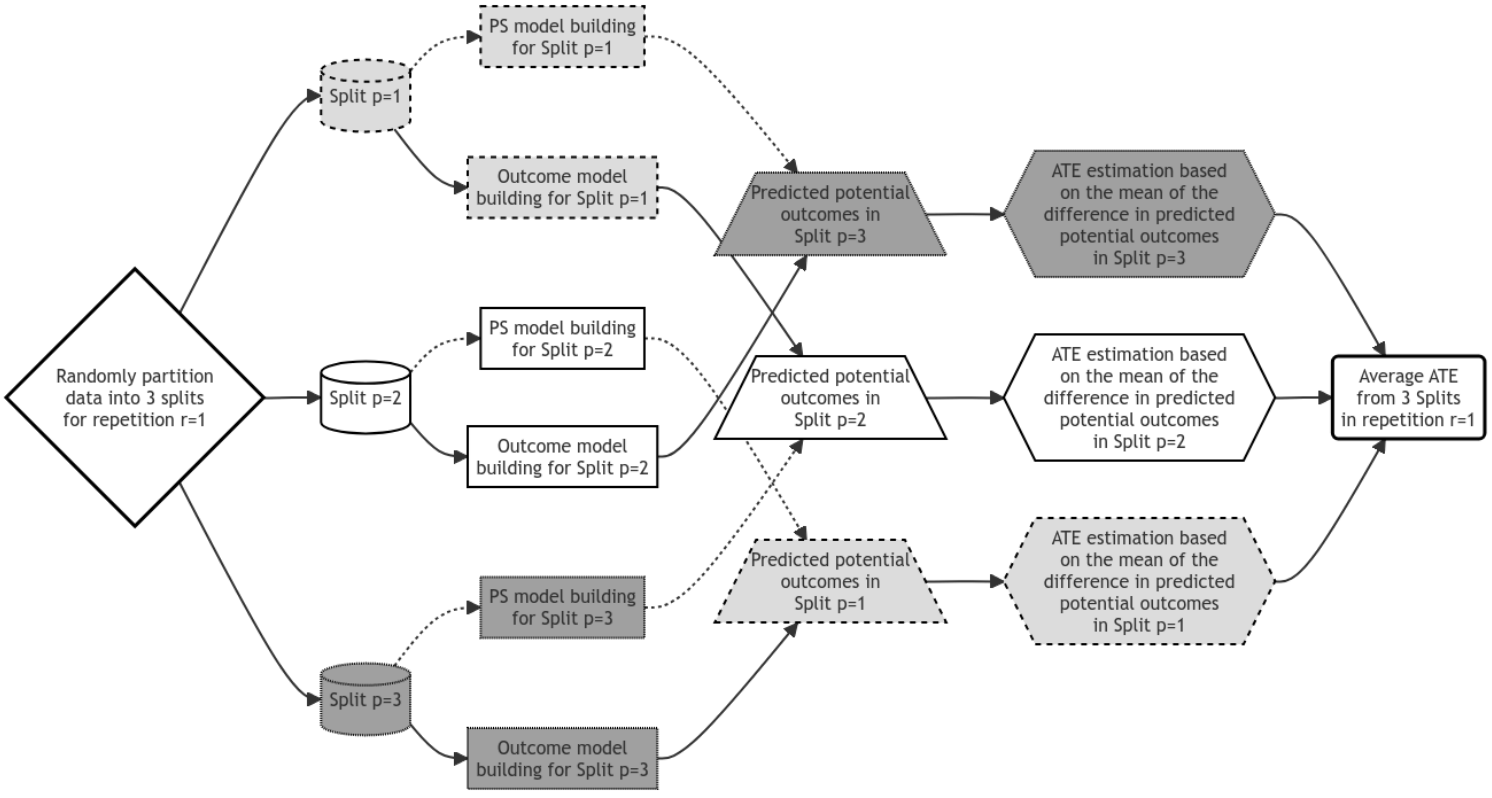
[1]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

[2]Centre for Advancing Health Outcomes, University of British Columbia, Vancouver, BC, Canada

Purpose: Flexible machine learning algorithms are increasingly utilized in real-data analyses. When integrated within double robust methods, such as the Targeted Maximum Likelihood Estimator (TMLE), complex estimators can result in significant undercoverage. The Double Cross-Fitting (DCF) procedure complements these methods by enabling the use of diverse machine learning estimators, yet optimal guidelines

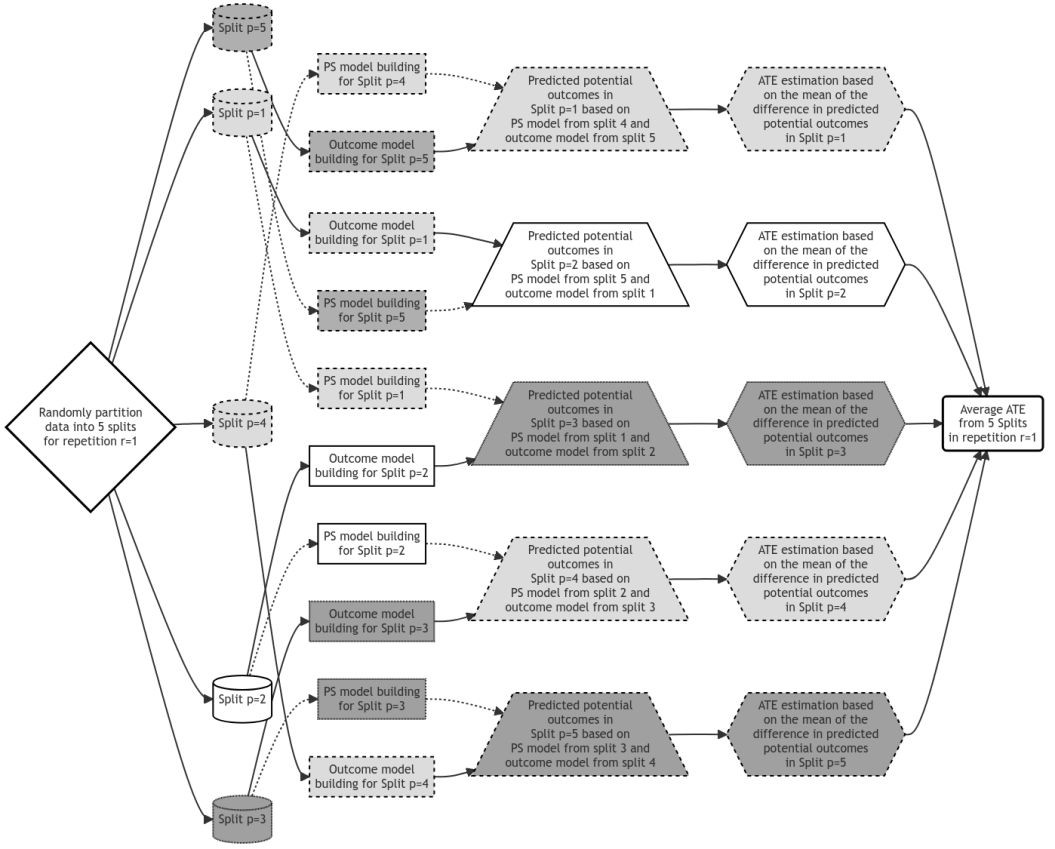
Manuscript under revision

(2) Double Crossfitting Guideline



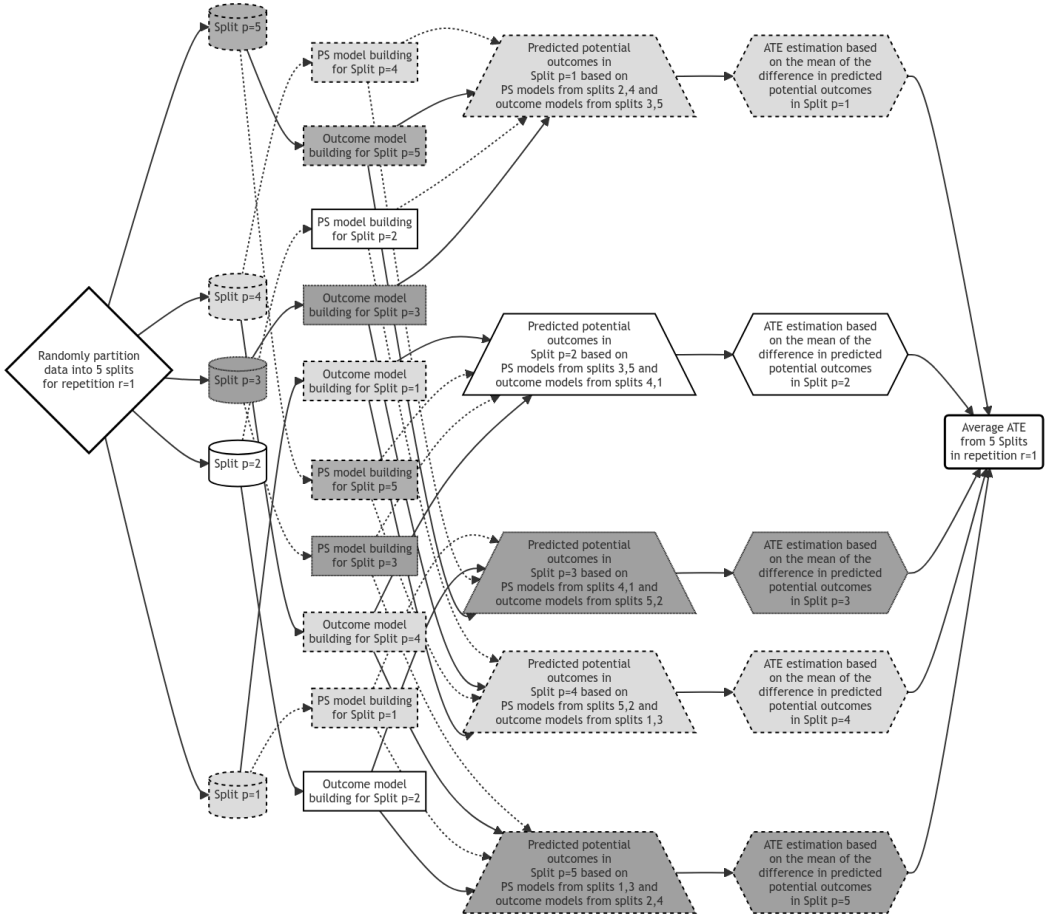
3 splits (folds)

(2) Double Crossfitting Guideline



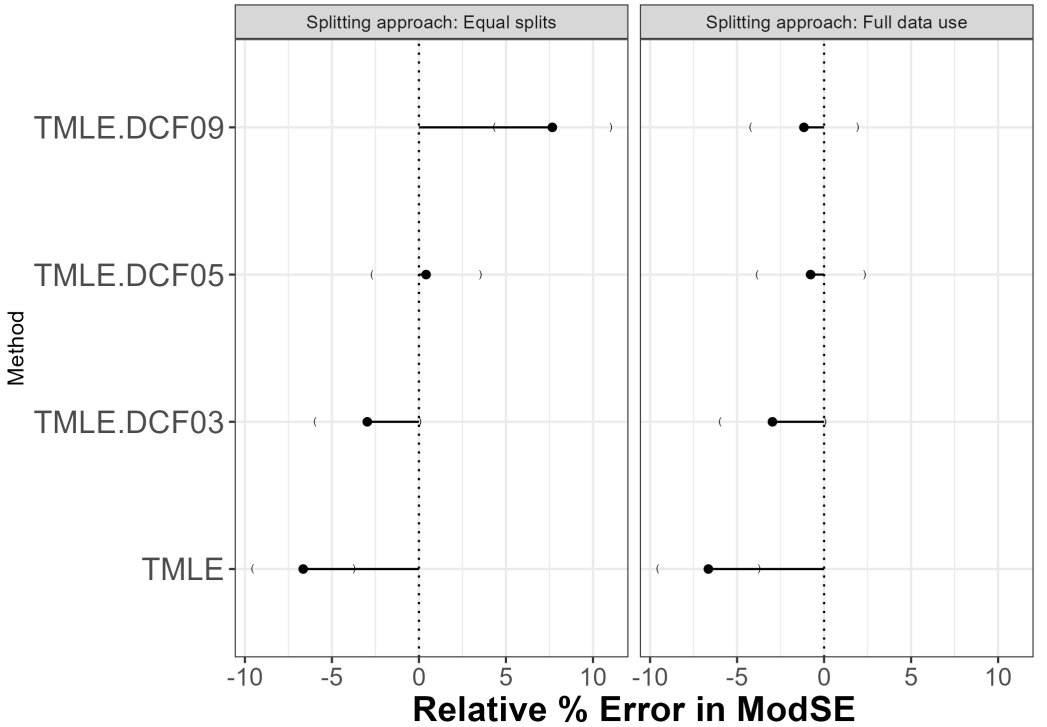
5 splits (folds), with some loss in data

(2) Double Crossfitting Guideline



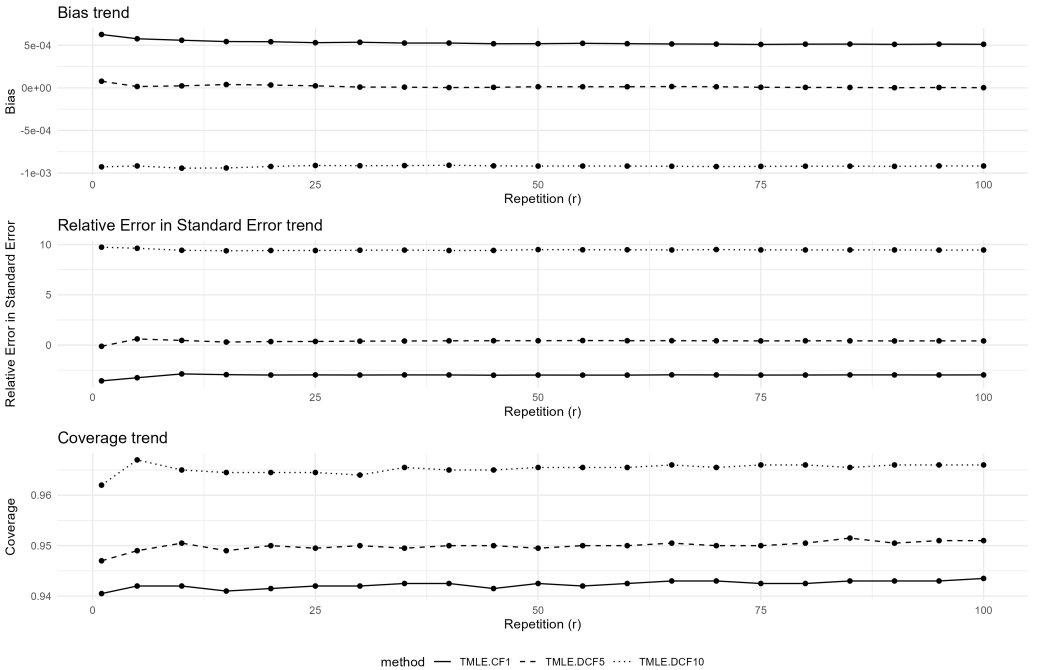
5 splits (folds), but full use of data

(2) Double Crossfitting Guideline: How many splits?



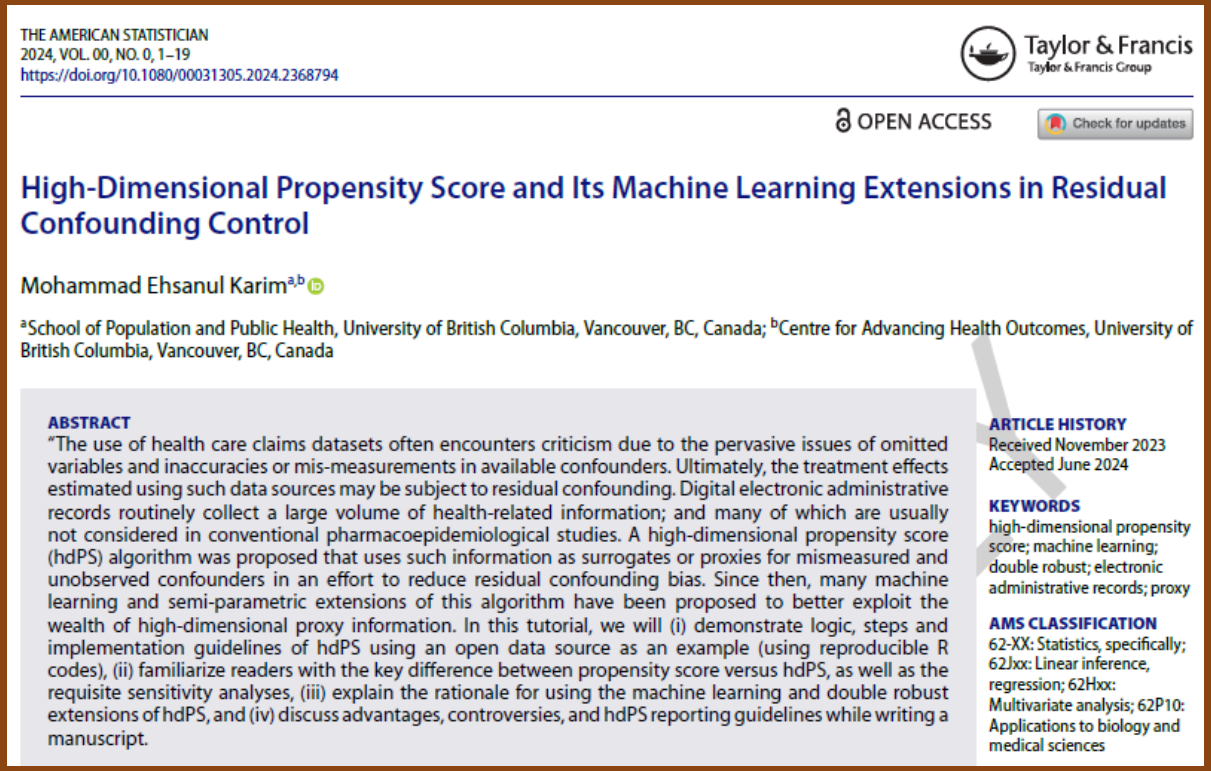
Identifying optimal number of splits

(2) Double Crossfitting Guideline: How many repetitions?



Identifying optimal number of repetitions

(3) Residual Confounding Bias Reduction: High Dimensional Propensity Score vs ML extensions



THE AMERICAN STATISTICIAN
2024, VOL. 00, NO. 0, 1–19
https://doi.org/10.1080/00031305.2024.2368794

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

High-Dimensional Propensity Score and Its Machine Learning Extensions in Residual Confounding Control

Mohammad Ehsanul Karim[a,b]

[a]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada; [b]Centre for Advancing Health Outcomes, University of British Columbia, Vancouver, BC, Canada

ABSTRACT

"The use of health care claims datasets often encounters criticism due to the pervasive issues of omitted variables and inaccuracies or mis-measurements in available confounders. Ultimately, the treatment effects estimated using such data sources may be subject to residual confounding. Digital electronic administrative records routinely collect a large volume of health-related information; and many of which are usually not considered in conventional pharmacoepidemiological studies. A high-dimensional propensity score (hdPS) algorithm was proposed that uses such information as surrogates or proxies for mismeasured and unobserved confounders in an effort to reduce residual confounding bias. Since then, many machine learning and semi-parametric extensions of this algorithm have been proposed to better exploit the wealth of high-dimensional proxy information. In this tutorial, we will (i) demonstrate logic, steps and implementation guidelines of hdPS using an open data source as an example (using reproducible R codes), (ii) familiarize readers with the key difference between propensity score versus hdPS, as well as the requisite sensitivity analyses, (iii) explain the rationale for using the machine learning and double robust extensions of hdPS, and (iv) discuss advantages, controversies, and hdPS reporting guidelines while writing a manuscript.

ARTICLE HISTORY
Received November 2023
Accepted June 2024

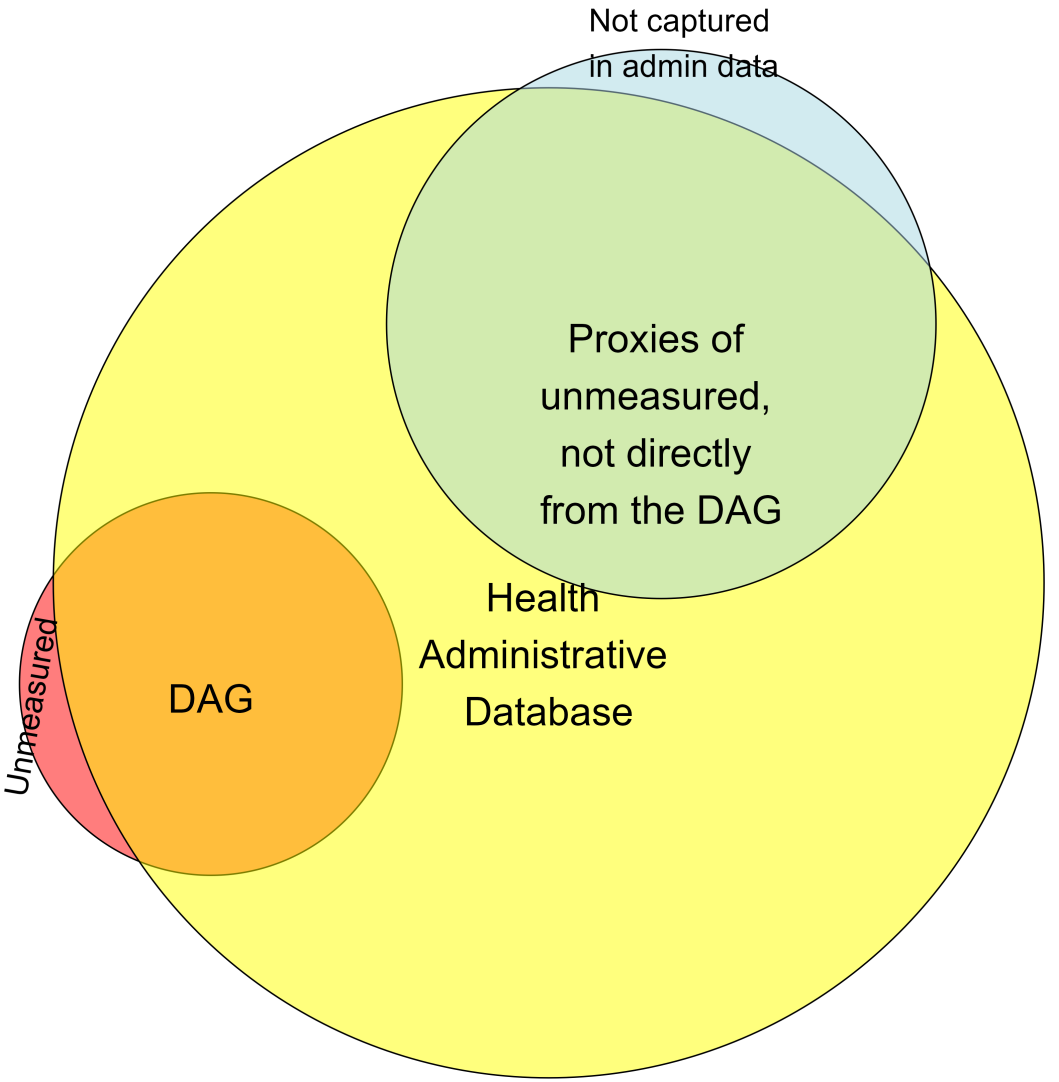
KEYWORDS
high-dimensional propensity score; machine learning; double robust; electronic administrative records; proxy

AMS CLASSIFICATION
62-XX: Statistics, specifically; 62Jxx: Linear inference, regression; 62Hxx: Multivariate analysis; 62P10: Applications to biology and medical sciences

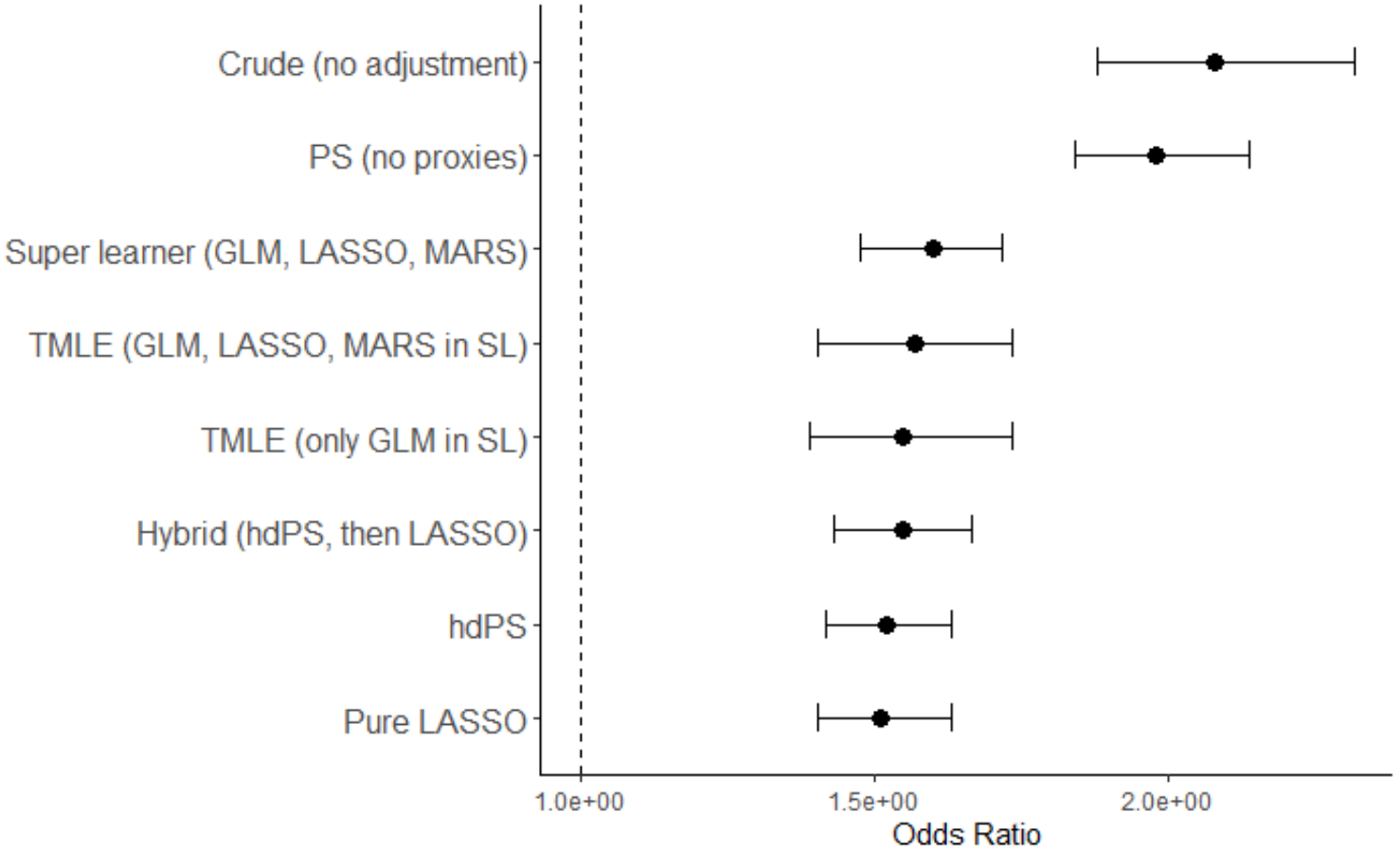
Tutorial on hdPS

The American Statistician: https://doi.org/10.1080/00031305.2024.2368794

(3) High Dimensional Propensity Score Idea



(3) hdPS vs ML extensions



Empirical analysis on NHANES data.

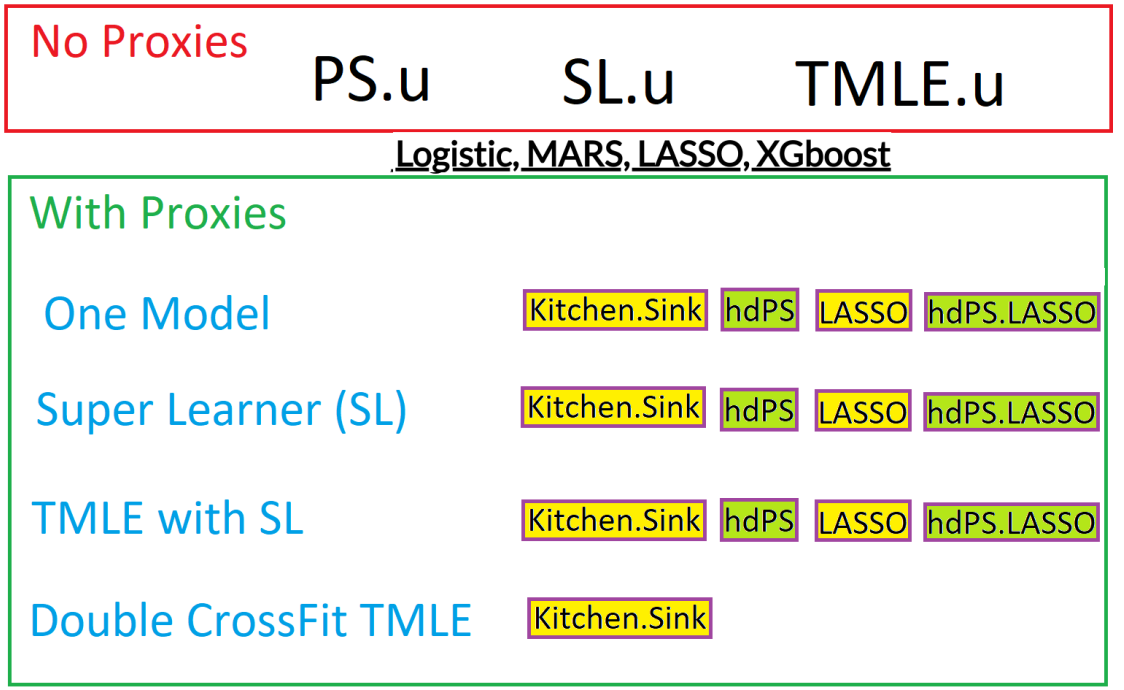
(4) Simulation Mechanism (Proxy and model specification)

1. Used imperfect proxy of **unmeasured confounder (problem 1)**.
2. **Transformed lab variables** (polynomials, interactions, complex functions, 10 converted to 6)

$$\begin{aligned}
\text{Transformed.var.1} &= \log(\text{globulin}) \\
\text{Transformed.var.2} &= \text{protein} \cdot \text{calcium} \\
\text{Transformed.var.3} &= \left(\frac{\text{diastolicBP}}{\text{systolicBP}}\right)^2 \\
\text{Transformed.var.4} &= \sqrt{\frac{\text{uric.acid} + \text{bilirubin}}{2}} \\
\text{Transformed.var.5} &= \frac{\text{phosphorus}^2}{\text{sodium} \cdot \text{potassium}} \\
\text{Transformed.var.6} &= \log(\text{systolicBP} + 10)
\end{aligned}$$

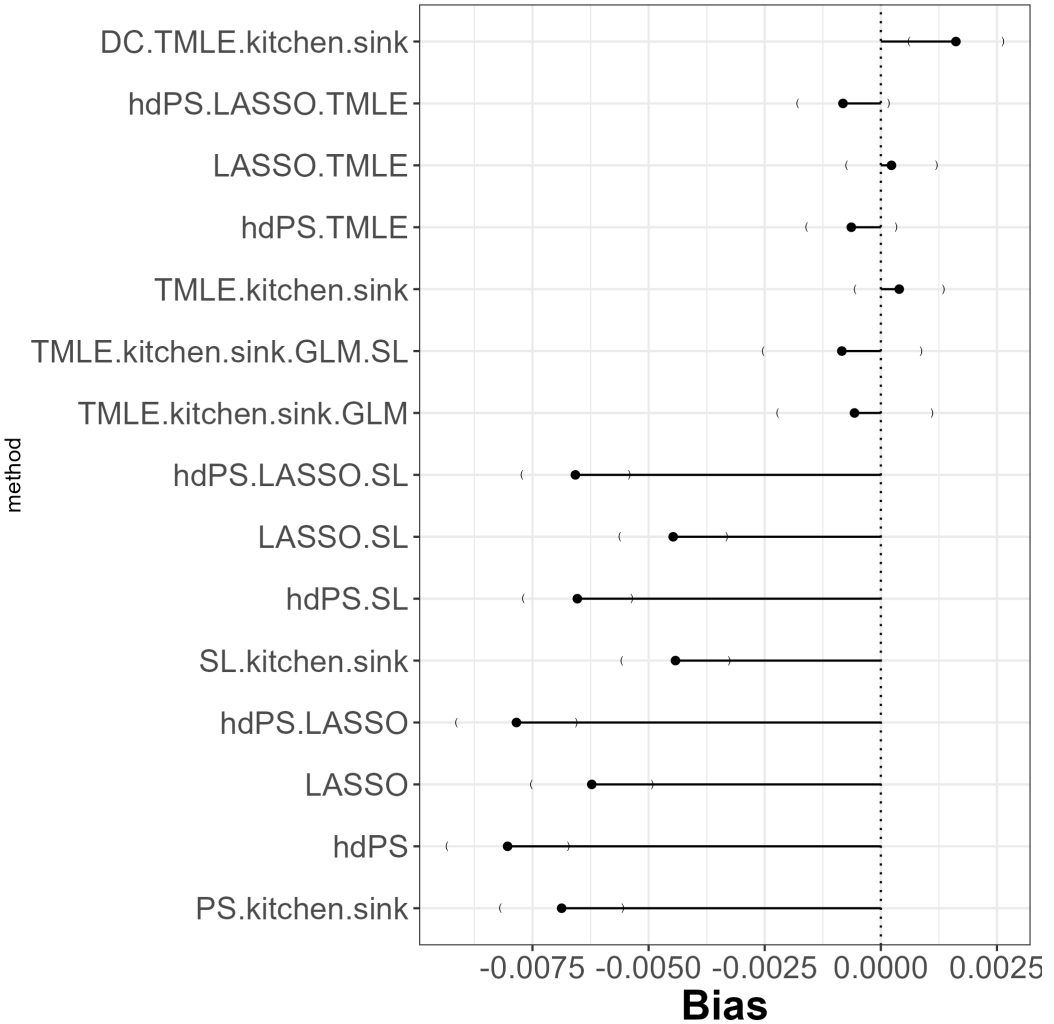
Original (untransformed 10) variables supplied in analysis, inducing **model-misspecification (problem 2)**.

(4) hdPS and TMLE (and choice of Super Learner)

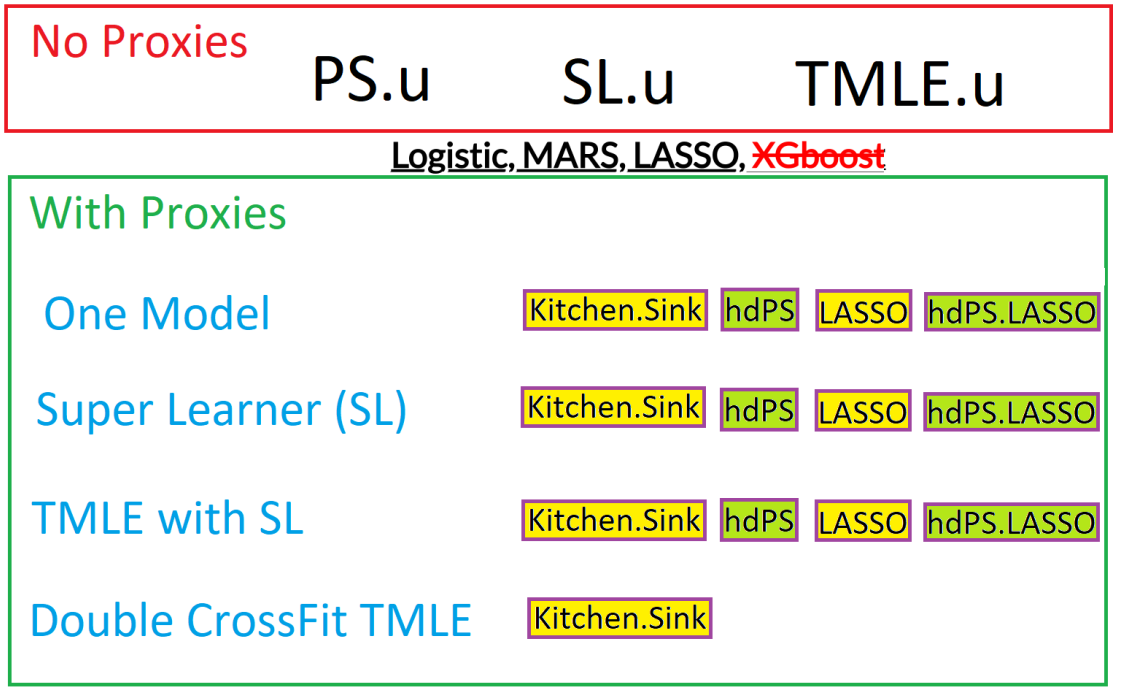


4 learners within Super learner.

(4) hdPS and TMLE (and choice of Super Learner)



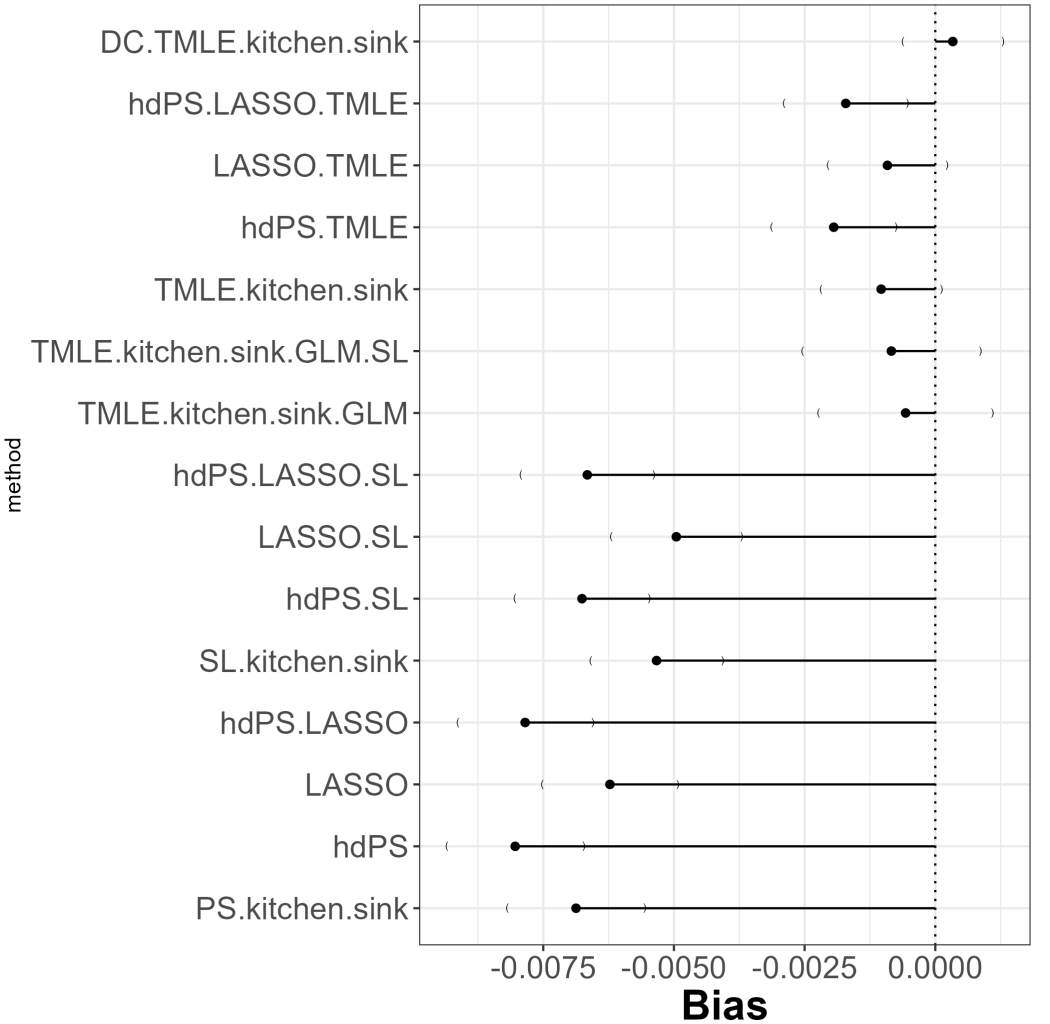
(4) hdPS and TMLE (and choice of Super Learner)



3 learners within Super learner.

(Phillips, van der Laan, Lee, and Gruber, 2023)

(4) hdPS and TMLE (and choice of Super Learner)



(4) hdPS and TMLE (and choice of Super Learner)

ShinyApp: https://ehsanx.shinyapps.io/hdPSsim/

Summarize

- Alsways good to perform a simulation to check validity of the plug-in ML method
  - Especially in terms of variance estimation in causal inference problems
- Double robust approaches are recommended.
  - Software for SL, TMLE, DCF-TMLE available.
- Always good to compare with results from regular regression.

References

Karim, E. M. E. (2021b). *Understanding Basics and Usage of Machine Learning in Medical Literature*. https://ehsanx.github.io/into2ML/. v1.1.

Karim, M. E. and pi-OER team (2024). "Advanced Epidemiological Methods". . Accessed on: April 20, 2024. URL: https://ehsanx.github.io/EpiMethods/.

Phillips, R. V., M. J. van der Laan, H. Lee, et al. (2023). "Practical considerations for specifying a super learner". In: *International Journal of Epidemiology*. DOI: 10.1093/ije/dyad023.

Weberpals, J., T. Becker, J. Davies, et al. (2021). "Deep learning-based propensity scores for confounding control in comparative effectiveness research: A large-scale, real-world data study". In: *Epidemiology* 32.3, pp. 378-388.