Challenges of Univariate Proxy Variable Screening in High-Dimensional Propensity Score Analysis - An Argument for Multivariate Approaches

Dr. Mohammad Ehsanul Karim MSc, PhD[1,2*] and Yang Lei[3*]

[1*]School of Population and Public Health, University of British Columbia, Vancouver, Canada, V6T 1Z3, BC, 2206 East Mall.

[2]St. Paul's Hospital, Vancouver, Canada, V6Z 1Y6, BC, 588 - 1081 Burrard Street.

[3]Department of Statistics, University of British Columbia, Vancouver, Canada, V6T 1Z4, BC, Room 3182 Earth Sciences Building, 2207 Main Mall.

*Corresponding author(s). E-mail(s): ehsan.karim@ubc.ca;

Abstract

Purpose:
Methods:
Results:
Conclusion:

Keywords: Machine learning, Propensity score, Deep learning, Causal inference

JEL Classification: C18

MSC Classification: 92D30 , 62P10

1 Background

Aim:

2 Methods

Data and Simulation

Right Heart Catheterization dataset:

Plasmode simulation: See Table 1 for the description of the scenarios under consideration.

Table 1: Simulation Scenarios for plasmode simulation based on the Right Heart Catheterization (RHC) study.

Plasmode Simulation Scenario	Exposure Prevalence	Outcome Prevalence	True Odds Ratio	Sample Size
(i) Frequent Exposure and Outcome (Base)	30%	30%	1	3,500
(ii) Rare Exposure and Frequent Outcome	5%	30%	1	3,500
(iii) Frequent Exposure and Rare Outcome	30%	5%	1	3,500

True data generating mechanism used in plasmode simulation:

Performance measures: From this simulation, we derived several performance metrics: (1) bias, (2) average model standard error (SE; the average of estimated SEs obtained from a model over repeated samples), (3) empirical SE (the standard deviation of estimated treatment effects across repeated samples), (4) MSE, (5) coverage probability of 95% confidence intervals, (6) bias-eliminated coverage, and (7) Zip plot [1, 2].

Estimators under consideration

3 Results

4 Real-world analysis

Computing time:

5 Discussion

Contextualizing the literature:

Summary of the simulation findings:

Data analysis findings:

Future Direction:

Conclusion:

List of abbreviations

1. MSE - Mean Squared Error
2. SE - Standard Error
3. PS - Propensity Score

4. AE - Autoencoders
5. DL - Deep Learning
6. MARS - Multivariate Adaptive Regression Splines
7. SMD - Standardized Mean Difference
8. TMLE - Targeted Maximum Likelihood Estimation
9. RHC - Right Heart Catheterization
10. SUPPORT - Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments

Declarations

Ethics approval and consent to participate

The analysis conducted on secondary and de-identified data is exempt from research ethics approval requirements. Ethics for this study was covered by item 7.10.3 in University of British Columbia's Policy #89: Research and Other Studies Involving Human Subjects 19 and Article 2.2 in of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2).

Consent for publication

Availability of data and materials

Competing interests

Over the past three years, MEK has received consulting fees from Biogen Inc. for consulting unrelated to this current work. MEK was previously supported by the Michael Smith Foundation for Health Research Scholar award.

Funding

This work was supported by MEK's Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants and Discovery Accelerator Supplements.

Authors' contributions

MEK: Conceptualization, Writing – Original Draft, Review & Editing YL: Formal Analysis, Review & Editing