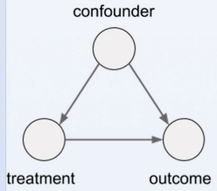
# Big Data Methods for Real World Data: A Simulation Study to Assess Suitability of Machine Learning Methods to Account for Confounding Under Various Treatment and Outcome Prevalence Scenarios

Yuchen Guo, PhD student at University of Oxford, yuchen.guo@ndorms.ox.ac.uk
Supervised by Prof. Daniel Prieto-Alhambra, Dr. Sara Khalid, Dr. Victoria Strauss and Dr. M Sanni Ali

## Abstract

Motivated by the surge of treatments for COVID-19 in the second quarter of 2020, with low prevalence of multiple emerging treatments, previous research suggests that disease risk score (DRS) could be a useful alternative to propensity scores (PS) for treatment effect estimation studies using observational data (Glynn, 2012). We tested DRS against PS under high and common outcome prevalence, as well as a range of different treatment prevalence. Three machine learning methods were applied to obtain PS and DRS: least absolute shrinkage and selection operator (LASSO), artificial neural network (ANN) and eXtreme Gradient Boosting (XgBoost). Average treatment effect was estimated by matching with both PS and DRS and relative bias were reported. Results demonstrated that for both common and high outcome prevalence, when decreasing treatment prevalence, PS and DRS matching bias increase. When treatment prevalence is lower than 10%, for both common and high outcome prevalence, we suggest using DRS as it gives lower bias. On the other hand, if treatment prevalence is common (between 10% and 50%), we suggest use PS matching instead. Among the three machine learning methods, all of them have similar performance, LASSO and XgBoost can both give us lowest bias under different scenarios.

## Background: confounding




To tackle confounding problem, traditionally, multivariable logistic regression informed by previous knowledge on causal structure and including pre-specified confounder/s is used for PS/DRS estimation. Machine learning (ML) methods are promising alternatives for automatic, more "agnostic" variable selection and PS/DRS calculation.

## Background: PS, DRS

- Both PS and DRS summarize confounding information into single measure.
- PS is the probability of each individual is exposed to treatment, as a function of observed covariates
- DRS is the probability of event occurrence, conditional on being unexposed and observed covariates . There are two ways calculating it:
  - Full cohort DRS (The one we are using)
  - Unexposed population DRS

## Objective

To test data driven machine learning methods and compare DRS and PS methods in treatment effect estimation, under different treatment prevalence and outcome prevalence

## Methods

We conducted simulation studies to compare full cohort DRS method and PS method for treatment effect estimation under different treatment prevalence and different prevalence of outcome and multiple confounders.

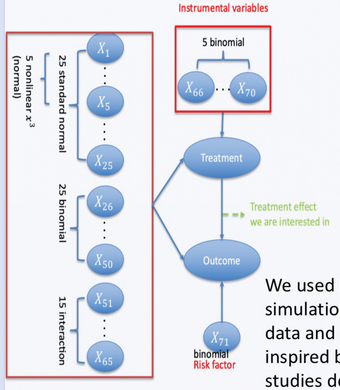
First, we simulated different dataset for different scenarios using Monte Carlo simulation. Then, DRS and PS were computed using three different kinds of machine learning methods:

- neural network: ANN
- Tree-based ensemble method: XgBoost
- Logistic regression with LASSO regularization

All of them were hyperparameter tuned using random search. Then, matching was used to obtain estimated treatment effect. Finally, we used relative bias and corresponding 95% confidence interval to assess the estimation.

## Data generation

Figure 1: Data generating process




We used Monte Carlo simulation, distribution of data and data settings were inspired by simulation studies done by Setoguchi(Setoguchi, 2008).

Data settings:

- 100 iterations. For each iteration, we generated 10000 observations, which has 50 confounders, 10 instrumental variables, a risk factor.
- We also included 15 interaction terms and 5 quadratic terms to represent non-additivity and non-linearity in data.
- Among 50 confounders, we set 25 of them to normal distribution and the other 25 of them to binomial distribution.
- Binary exposure/treatment were generated with prevalence 1%, 5%, 10%, 25% and 50%.
- Binary outcome were generated with prevalence 2% and 50%.

Table 1: Simulation scenarios

| Scenarios # | Treatment prevalence | Outcome risk |
|---|---|---|
| 1 – 5 | 0.01, 0.05, 0.1, 0.25, 0.5 | 0.5 |
| 6-10 | 0.01, 0.05, 0.1, 0.25, 0.5 | 0.02 |

## Results

- For all three machine learning methods, when decreasing treatment prevalence, bias generated by PS method increases.
- When treatment prevalence is lower than 10%, we observed that the treatment effect estimated by DRS has much lower bias comparing to PS.
- When treatment prevalence is greater than 10%, PS outperforms DRS for most scenarios.

Figure 2: PS vs DRS results

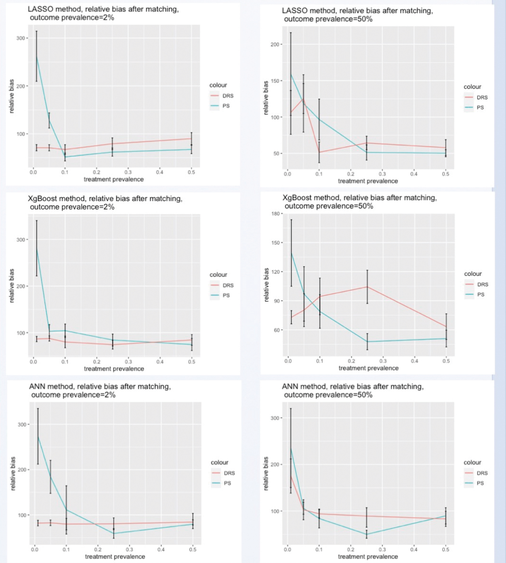
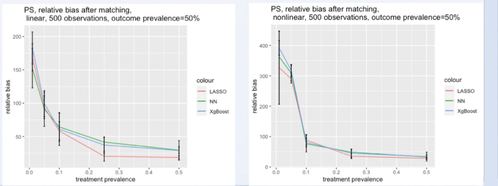
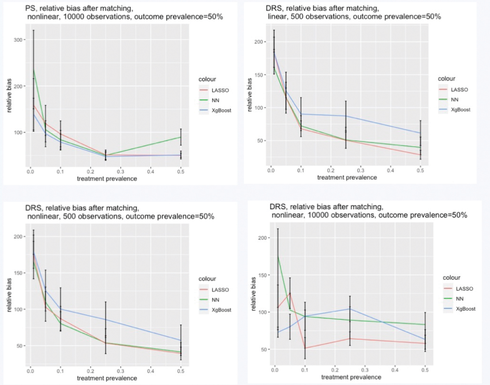



Figure 3: Machine learning comparison results




## Conclusions

- Under both high outcome prevalence(50%) or common outcome prevalence(2%), and multiple confounders setting, DRS outperforms PS when treatment prevalence is rare (lower than 10%).
- When treatment prevalence is common (between 10% and 50%), our data suggest use PS instead of DRS, as most of the relative bias we obtained from PS matching is lower than matching on DRS.
- Among the three machine learning methods, LASSO, XgBoost and Neural Network, all of them have similar performance as we have quite complicated data setting. Our data would recommend trying both LASSO and XgBoost as they both can give us lowest bias under certain scenarios.

## Reference

Glynn, R. G. (2012). Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf.

Setoguchi, S. S. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety, 546–555.

www.PosterPresentations.com