Big Data Method 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

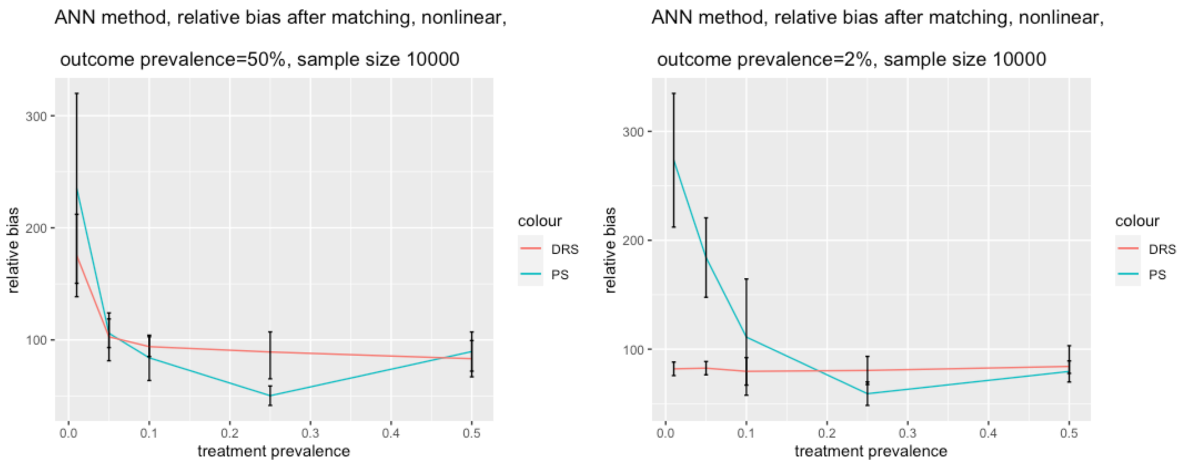
Authors: Yuchen Guo, Sarah Khalid, Victoria Strauss, M Sanni Ali and Daniel Prieto-Alhambra

Background: Observational research often involves data that are routinely collected in healthcare practice. While having the advantage of being representative of real-world, these data often come from a variety of sources and can include a large number of patient-level variables. Propensity score (PS) and disease risk score (DRS) are frequently used to minimize confounding in real world studies and machine learning methods are considered to be potentially suitable for calculating the score from multi-modality and multi-dimensional data (Patrick G. Arbogast, 2011). It is important to understand how they perform under different scenarios such as with varying degrees of treatment and outcome risk.

Objectives: To assess the performance of data driven methods for minimising confounding under different treatment and outcome risk scenarios, given a large number of patient-level variables.

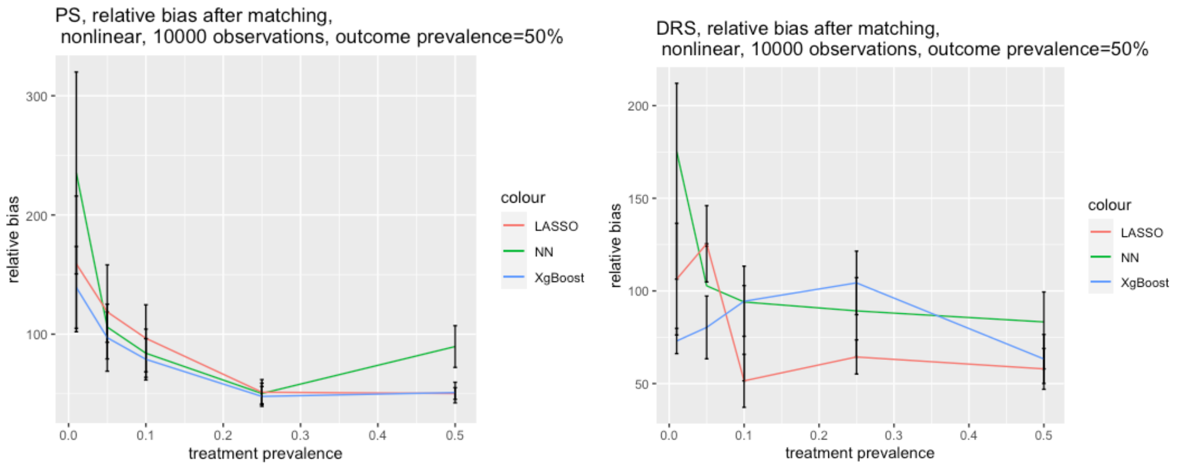
Methods: DRS and PS were computed using three machine learning methods: least absolute shrinkage and selection operator (LASSO), artificial neural network (ANN) and eXtreme Gradient Boosting (XgBoost), all of them hyperparameter tuned. Matching at 1:1 ratio for exposed and unexposed (PS) and outcome and without outcome (DRS) was then used to obtain estimated treatment effect. Finally, we used relative bias and the corresponding confidence intervals to assess the estimation. Data we generated has 50 patient confounders (25 standard normal and 25 binomial distribution), 5 instrumental variables (binomial distribution) and 1 risk factor (binomial distribution), as well as 30% two-way interaction terms and 10% nonlinear terms. These settings of nonlinearity and non-additivity were inspired by (Setoguchi, 2008).

Results: Part of the plots are shown below for demonstration and details of all results in plot can be found in appendix**.** Bias of treatment effect estimation under different scenarios were reported. Bias of treatment effect estimation under different scenarios were reported. Results demonstrated that for both common and high outcome risk, when decreasing treatment prevalence, bias increase. When treatment prevalence is lower than 10%, for both common and high outcome risk, our data suggest using DRS as it gives lower bias, while PS gives more than 100% relative bias. On the other hand, if treatment prevalence is common (between 10% and 50%), our data suggest use PS instead of DRS. Among the three machine learning methods, all of them have similar performances, LASSO and XgBoost can both give us lowest bias under different scenarios.


ANN method, relative bias after matching, nonlinear, outcome prevalence=50%, sample size 10000

ANN method, relative bias after matching, nonlinear, outcome prevalence=2%, sample size 10000

Figure 1: PS DRS comparison


PS, relative bias after matching, nonlinear, 10000 observations, outcome prevalence=50%

DRS, relative bias after matching, nonlinear, 10000 observations, outcome prevalence=50%

Figure 2: Machine learning methods comparison

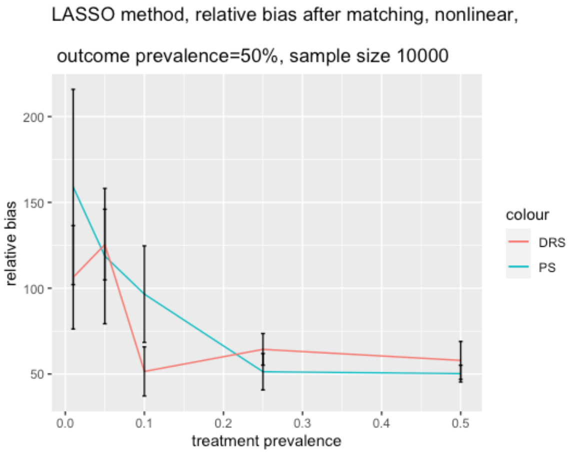
Further Exploration: We are currently testing both methods in Clinical Practice Research Datalink (CPRD) data to demonstrate our results and provide more insights.

Bibliography

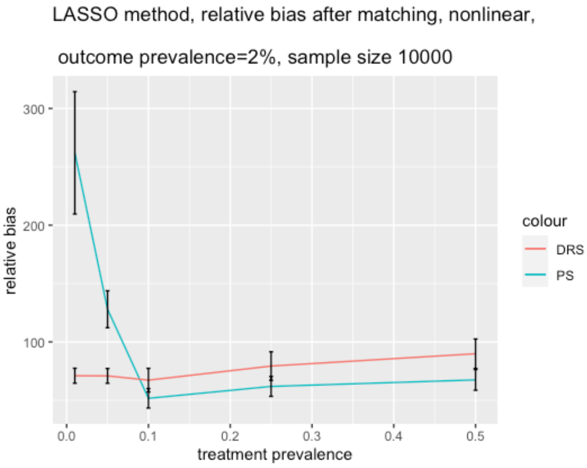
Patrick G. Arbogast, W. A. (2011). Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders, . *American Journal of Epidemiology*.

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

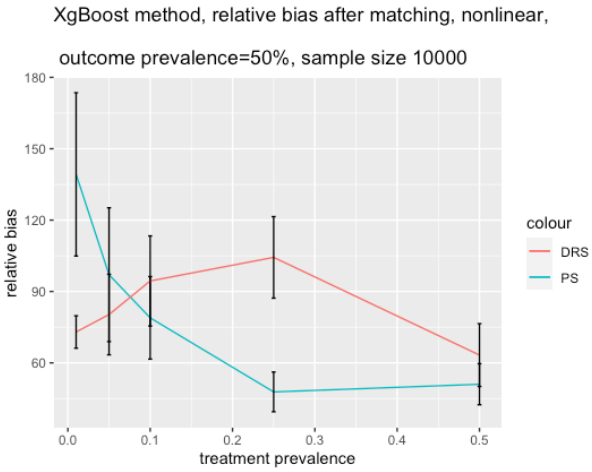
Appendix


LASSO method, relative bias after matching, nonlinear,

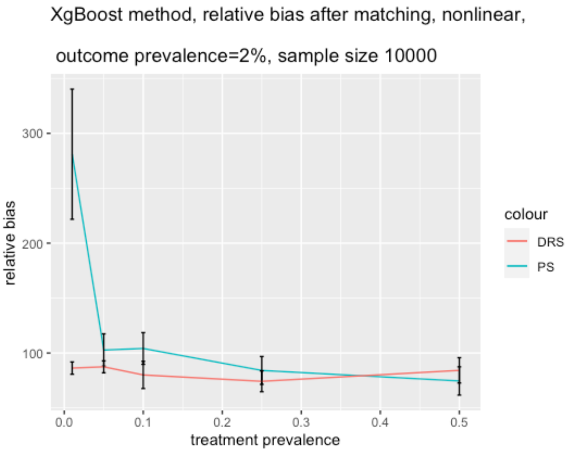
outcome prevalence=50%, sample size 10000


LASSO method, relative bias after matching, nonlinear,

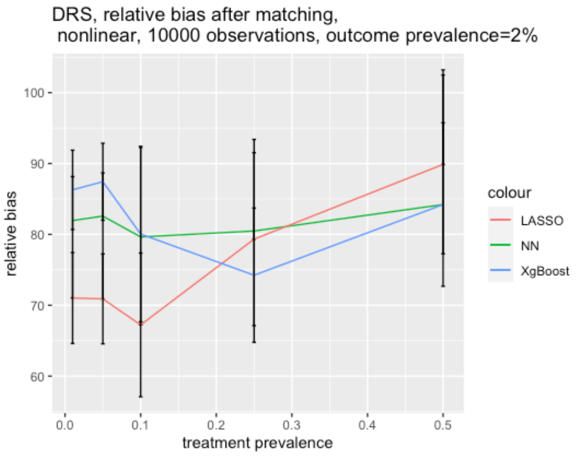
outcome prevalence=2%, sample size 10000

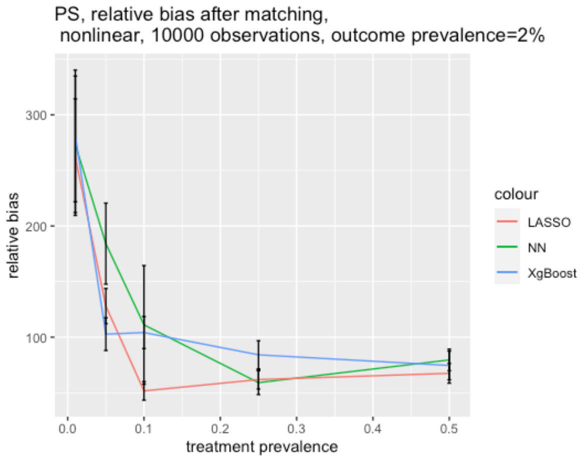

XgBoost method, relative bias after matching, nonlinear,

outcome prevalence=50%, sample size 10000


XgBoost method, relative bias after matching, nonlinear,

outcome prevalence=2%, sample size 10000


DRS, relative bias after matching,
nonlinear, 10000 observations, outcome prevalence=2%


PS, relative bias after matching,
nonlinear, 10000 observations, outcome prevalence=2%