How high can we go? Evaluating massively high-dimensional propensity score models in large-scale observational studies

Yuxi Tian[1], Martijn J. Schuemie[2], PhD, Marc A. Suchard[1,3,4], MD, PhD

[1] Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
[2] Janssen Research and Development LLC, Titusville, NJ, USA
[3] Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, CA, USA
[4] Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Abstract

Large-scale observational studies that fully utilize the information available in healthcare databases can include millions of patients and unique measurements of their health. These massively high-dimensional scenarios pose challenges in developing propensity score and outcome models for conducting cohort studies to examine drug safety or comparative effectiveness. We have developed novel OHDSI tools that implement the high-dimensional propensity score (hdPS) algorithm and massive sample-size, regularized regression (MSSRR) methods in constructing comparable patient cohorts. We plan to evaluate the performance of both propensity score approaches through measures of cohort balance and through estimation of treatment effect when coupled with an outcome model. Comparison studies are conducted through data simulation and through analyzing several real-world drug safety issues at scale. We wish to characterize the capabilities of different propensity score and outcome models on the largest scales necessitated by observational healthcare data analysis.

Introduction

The specification of propensity score models to identify comparable patients is a crucial decision in conducting observational studies. In dealing with healthcare claims databases where the number of patients and variables alike can range in the millions or more, an investigator cannot know based on expert knowledge alone the exact covariates to include in a propensity score or outcome model. Variable selection techniques are needed to facilitate this process.

The high-dimensional propensity score (hdPS) algorithm is one method for selecting potential confounders for inclusion in a propensity score [1]. Covariates are ranked by their prevalence and by their univariate association with the outcome and/or the treatment; a certain number are then used in the propensity score model. While hdPS has been used for large-scale observational studies, its actual performance compared to standard multivariate methods, such as regularized regression and its more recent OHDSI extensions for massive sample-size, regularized regression (MSSRR) [2], has only been investigated on much smaller scales [3].

MSSRR methods stand as useful alternatives to hdPS for propensity score models in massive observational healthcare settings. In regularized regression, all potential covariates are included in a multivariate regression; a penalty term shrinks coefficients with extreme values towards 0, leaving a subset of the original covariates for inclusion in the final model. The performance of MSSRR in generating propensity scores has not been thoroughly evaluated for large-scale observational studies.

Methods

We have recently implemented the hdPS algorithm within the OHDSI CohortMethod package for reproducible usage across OHDSI studies utilizing the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). Our hdPS implementation can serve as a drop-in substitute for the MSSRR-based propensity score model provided through the Cyclops package. We also plan on implementing additional measures to test the performance of propensity score methods in creating covariate balance.

To evaluate the relative performance of hdPS and MSSRR in building a propensity score model at scale, we plan to compare several measures, including (1) estimation of overall treatment effect using a Cox regression outcome model conditioned on matching on propensity score, (2) covariate balance within strata built on the propensity score, (3) covariate balance among matched sets built on the propensity score. We intend to assess these metrics in simulation studies that builds upon Franklin et al. [3]. The chief difference in design lies in the size of our simulated samples; while the aforementioned study uses simulations of 30,000 patients, we intend to perform simulations larger by one order-of-magnitude or more, in the millions of patients. Finally, we intend to analyze several relevant drug safety issues using real-world data. We will assess each propensity score's ability to estimate a zero treatment effect size for treatments that are known negative controls and a non-zero treatment effect size upon signal injection with known positive controls.

Results

The diagram below outlines the steps necessary to employ hdPS in the CohortMethod package and can be employed immediately in package studies, e.g. the celecoxib vs. diclofenac analysis described in the main CohortMethod vignette example.

```
library(CohortMethod) # establish connection and CohortMethod settings (omitted)

# HDPS implementation
screenedData = runHdps(cohortMethodData)              # univariate screen
hdPs <- createPs(screenedData, outcomeId = 3,         # fit logistic regression
                 prior = createPrior("none"))         # turn-off regularization
hdpsPropensityModel <- getPsModel(hdPs, screenedData) # return fitted model
```

Conclusions

We have recently implemented the hdPS model in the OHDSI CohortMethod package. This implementation provides an open-source, reproducible mechanism for constructing hdPS models against any dataset held in OMOP CDM, and for employing these models to construct patient cohorts for down-stream studies. Shortly, we plan to examine the relative performance of hdPS and MSSRR models in generating credible, population-level estimates of drug safety or comparative effectiveness.

References

[1] S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart, "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data," *Epidemiology*, vol. 20, no. 4, pp. 512 – 522, 2009.

[2] M. A. Suchard, S. E. Simpson, I. Zorych, P. Ryan, and D. Madigan, "Massive parallelization of serial inference algorithms for a complex generalized linear model," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 23, no. 1, p. 10, 2013.

[3] J. M. Franklin, W. Eddings, R. J. Glynn, and S. Schneeweiss, "Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses," *American Journal of Epidemiology*, p. Adv access: kwv108, 2015.