How high can we go? Evaluating massively high-dimensional propensity score and outcome 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 to implement the high-dimensional propensity score (hdPS) algorithm and massive sample-size, regularized regression (MSSRR) methods to construct comparable patient cohorts. We plan to evaluate performance, including cohort balance, of both propensity approaches and of MSSRR methods in building massive outcome models, while considering 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 specifications of propensity score models to identify comparable patients and outcome models to estimate treatment effects are crucial decisions 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; an arbitrary 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, and also as methods for outcome model generation 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 score and outcome models 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.

To evaluate the relative performance of hdPS and MSSRR in building a propensity score model at scale, we first plan to measure the mean squared error of propensity prediction through a simulation study 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 two orders-of-magnitude or more, in the millions of patients.

Correctly recovering propensity score estimates is only a single facet of the research process that consists of building propensity score models, using them to construct comparable patient cohorts and then estimating treatment effects between cohorts. We further plan to compare measures of cohort balance in propensity score stratifications generated using hdPS and MSSRR. In addition, we plan to report the differential treatment effect estimates based on outcome models that (1) are low-dimensional and stratify by hdPS estimates as recommended by hdPS developers [1], (2) use MSSRR, are high-dimensional and stratify by hdPS estimates, or (3) use MSSRR, are high-dimensional and stratify by MSRR propensity score estimates.

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(connection, 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.