HOW HIGH CAN WE GO?
EVALUATING MASSIVELY HIGH-DIMENSIONAL PROPENSITY SCORE MODELS IN LARGE-SCALE OBSERVATIONAL STUDIES

Yuxi Tian[1,5], Martijn J. Schuemie[2,5], and Marc A. Suchard[1,3,4,5]

[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

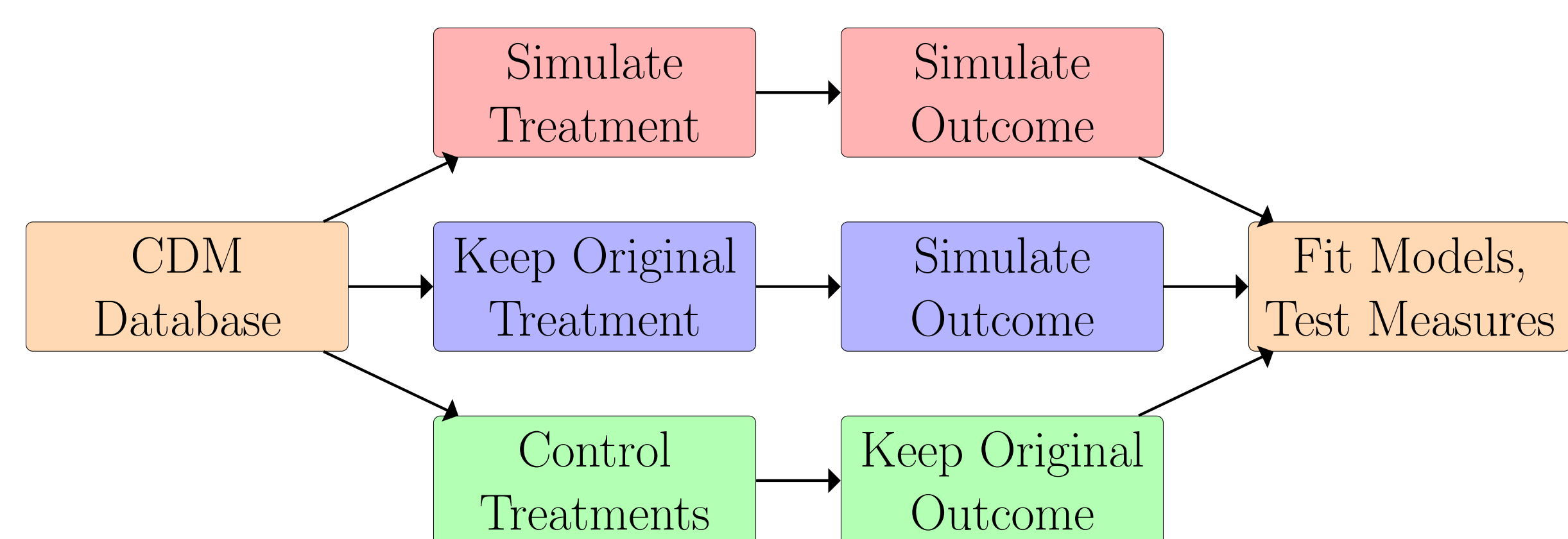[5]Observational Health Data Sciences and Informatics

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 high-dimensional scenarios pose challenges in developing propensity score and outcome models for conducting cohort studies to examine drug safety or comparative effectiveness. The **high-dimensional propensity score (hdPS)** algorithm and standard multivariate methods such as regularized regression are common methods for generating propensity score models to construct comparable patient populations, but their performance in large studies is not well characterized. We have developed OHDSI tools that implement hdPS, and we plan to compare its performance with recent OHDSI extensions for conducting **massive sample-size, regularized regression (MSSRR)**. We plan to evaluate the performance of both propensity score methods through measures of cohort balance and treatment effect estimation. 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.

Background

hdPS: **univariate screen**, **user-selected number of covariates**
The hdPS algorithm [1] assesses covariates based on their univariate association with the exposure and outcome. A specific number of covariates are included for use in the propensity score model.

MSSRR: **multivariate screen**, **covariates automatically selected**
Regularized regression uses a penalty term in a multivariate regression to automatically force many coefficients towards 0. MSSRR [2] can perform regularized regression in very large sample sizes.

While hdPS has been used for large-scale observational studies, its performance evaluations, including compared to multivariate methods, tend to occur at much smaller scales [3].

Methods

- We construct hdPS, covariate balance measures, and simulation methods to build upon the cohort study design implemented in the OHDSI COHORTMETHOD and PATIENTLEVELPREDICTION packages
- We also utilize MSSRR methods already implemented in the OHDSI CYCLOPS package

To evaluate the relative performance of hdPS and MSSRR in building a propensity score model at scale, we plan to compare the following measures:
1) Ability to predict the exposure status for each unit
2) Covariate balance among matched sets built on the propensity score
3) Estimation of overall treatment effect using a Cox outcome model

Methods - Simulations

We intend to analyze the propensity score models using several relevant drug safety issues based on real-world data, through three study designs as shown below. Our simulations build upon Franklin et al. [3], but we aim to construct studies larger in sample size by at least one order of magnitude.

In the third (bottom) simulation, we assess each model's ability to distinguish exposures that are known positive and negative controls (e.g. exposures that are known to affect, or not affect, the outcome).

Results

We have implemented the hdPS model and covariate balance measures in the OHDSI COHORTMETHOD and PATIENTLEVELPREDICTION packages.

The code below outlines the steps to employ hdPS in COHORTMETHOD, and can be used immediately in package studies, e.g. the celecoxib vs. diclofenac analysis described in the main COHORTMETHOD vignette example.

```r
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

- It's now possible to compare hdPS and MSSRR at scale using OHDSI tools
- Shortly, we will construct simulations to examine the relative performance of hdPS and MSSRR models in generating credible, population-level estimates of drug safety or comparative effectiveness

[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.