Name:	Yuxi Tian
Affiliation:	University of California, Los Angeles
Email:	ytian13@ucla.edu
Presentation type (s):	Lightning talk (preferred), else poster

Synthetic and Negative Control Evaluation Framework for Large-Scale Propensity Score Survival Analysis

Yuxi Tian[1], Martijn J. Schuemie[2], Marc A. Suchard[1]
[1]University of California, Los Angeles, California, USA;
[2]Janssen R&D, Titusville, New Jersey, USA

Abstract

Propensity score (PS) adjustment is a popular method for confounding control in observational studies, but reliable frameworks for evaluating PS performance are lacking. We compare the performance of L1-regularized regression and the high-dimensional propensity score (hdPS) algorithm as propensity score estimators using synthetic and real-world negative control outcomes experiments based on two published drug cohort studies. We find that L1-regularized regression achieves higher model fit and covariate balance than the hdPS algorithm, while relative performance on bias reduction is mixed.

Introduction

Propensity score adjustment has become a predominant method for confounding control in retrospective observational studies. By estimating the treatment assignment probability, propensity score adjusted studies mimic a randomized study, in which compared cohorts are balanced by design. Successful propensity score adjustment leads to bias reduction in estimating comparative drug effects.

Traditionally, investigators construct propensity scores by including established or suspected confounding covariates in the model. Modern access to longitudinal health databases for research use offers investigators an enormity of data that can overwhelm a manual selection decision process. Automated methods for propensity score model selection offer improved utilization of all available covariates. Among these, the high-dimensional propensity score (hdPS) algorithm[1], which conducts a univariate screen for potential confounders, has gained widespread use in pharmaco-epidemiology. Alternatively, L1-regularized logistic regression is a workhorse of statistical model selection, and performs model selection on all covariates simultaneously[2]. This study aims to compare the performance of the hdPS algorithm with L1-regularized regression as propensity score estimators.

We detail a synthetic and negative control outcome experiment framework for our propensity score performance evaluation. We construct simulated, synthetic data using an extension of the "plasmode" framework[3]. We also conduct negative control outcome experiments[4], which utilize real-world data and complement the shortcomings of solely relying on simulation results.

Evaluation Framework

Our synthetic experiments utilize realistic simulated data under known comparative drug hazard ratios. We operate under the Cox proportional hazards framework, and fashion each simulation after empirical study data. In the synthetic survival model, we use the observed covariates and treatment exposure data, and use estimated covariate

effects and baseline survival functions. Our negative control outcome experiments use control outcomes identified via a published procedure[5].

Empirical Studies

We reproduce two published observational studies using the Truven MarketScan Medicare Supplemental (MDCR) database. The first study[6] is a comparison of anticoagulant medications – dabigatran and warfarin – on intracranial hemorrhage adverse events. The second study[7] is a comparison of the Cox-2 inhibitor celecoxib and the traditional NSAID diclofenac on upper gastrointestinal complications.

Results

On outcome-independent metrics, L1-regularization outperforms the hdPS algorithm. The c-statistic, also known as the area-under-curve (AUC), measures treatment prediction accuracy. In both studies, L1-regularization has an AUC more than 0.05 greater than the evaluated hdPS algorithm methods. L1-regularization also produces improved covariate balance between compared drug groups compared to the hdPS algorithm.

In our simulations, we find mixed results on estimation bias reduction. No propensity score method is predominantly superior to the other methods. Additionally, our simulation results are strongly influenced by simulation parameters such as the true hazard ratio and the simulated outcome prevalence. These effects are endemic to the simulation design and complicate the comparison of propensity score methods.

In our negative control experiments, we judge propensity score methods by their ability to validate a set of negative controls as having statistically insignificant effect sizes. Unadjusted estimation is susceptible to bias and validates few negative controls, while propensity score adjustment increases the proportion of validated negative controls. We find mixed results in the propensity score methods, and neither L1-regularization or the hdPS algorithm performs significantly better than the other.

Conclusion

In this study, L1-regularization offers clear benefit compared to the hdPS algorithm on exposure prediction and covariate balance. Evaluation on outcome-dependent metrics such as bias reduction and negative control validation produce mixed results. Our evaluation methods extend existing survival simulation methods and demonstrate the utility of negative control experiments in propensity score evaluation.

References

1. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology (Cambridge, Mass.). 2009 Jul;20(4):512.
2. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996 Jan 1:267-88.
3. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational statistics & data analysis. 2014 Apr 30;72:219-26.
4. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Statistics in medicine. 2014 Jan 30;33(2):209-18.
5. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. Journal of biomedical informatics. 2017 Feb 28;66:72-81.
6. Graham DJ, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, Sheu TC, Mott K, Goulding MR, Houstoun M, MaCurdy TE. Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for non-valvular atrial fibrillation. Circulation. 2014 Oct 30:CIRCULATIONAHA-114.
7. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. European journal of clinical pharmacology. 2013 Mar 1;69(3):549-57.