Automated Data-Driven Analytics to Improve Robustness of Causal Effectiveness Studies in Healthcare Databases

Richard Wyss, PhD, MSc

Division of Pharmacoepidemiology and Pharmacoeconomics
Department of Medicine, Brigham and Women's Hospital
and Harvard Medical School

Background

- Electronic healthcare databases generated from insurance claims and electronic health records can provide valuable information on the effectiveness and safety of pharmaceutical drugs as they are used in routine-care

- However, these data sources are not collected for research purposes and confounding arising from non-randomized treatment choices remains a fundamental challenge for extracting valid evidence to help guide treatment and regulatory decisions.

Background

- Many analytic decisions to mitigate confounding bias.
  - Study design and restriction criteria
  - Which variables to adjust for?
  - How to adjust for variables?
- Decisions can be subjective and are not always transparent
- How can automation supplement expert knowledge to improve healthcare database studies?
  - Improve validity and robustness by minimizing subjective decision making during the analytic process.
  - Improve transparency and reproducibility.

Background

- Decisions related to study design and cohort definitions are difficult to automate:
  - Semi-automated computer modules & reporting guidelines can increase transparency and reproducibility
  - However, these analytic decisions require expert clinical guidance
    - selection of comparator group, restriction criteria, etc.
    - unique to each study question
- Decisions to reduce confounding bias once cohorts have been defined, automation is more promising.
  - Decisions on identification and selection of confounders
  - Decisions on how to adjust for selected confounders

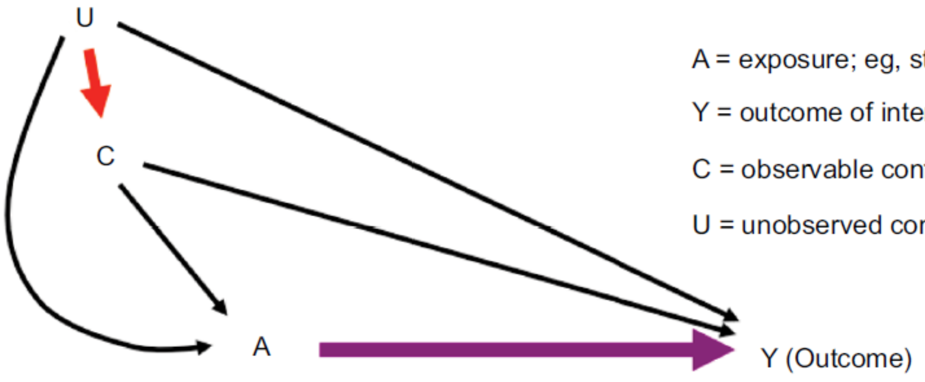
Limitations of Manual Confounder Selection

- Standard tools for confounding adjustment have typically relied on adjusting for a limited number of investigator specified variables.

- Adjusting for investigator-specified variables alone is often inadequate
  - Some confounders are unknown at the time of drug approval
  - Many confounders are not directly measured in routine-care databases.

Proxy Confounder Adjustment

- Healthcare databases may be understood and analyzed as a high-dimensional set of “proxy” factors that indirectly describe the health status of patients (Schneeweiss 2008, 2017).



A = exposure; eg, start of a new drug

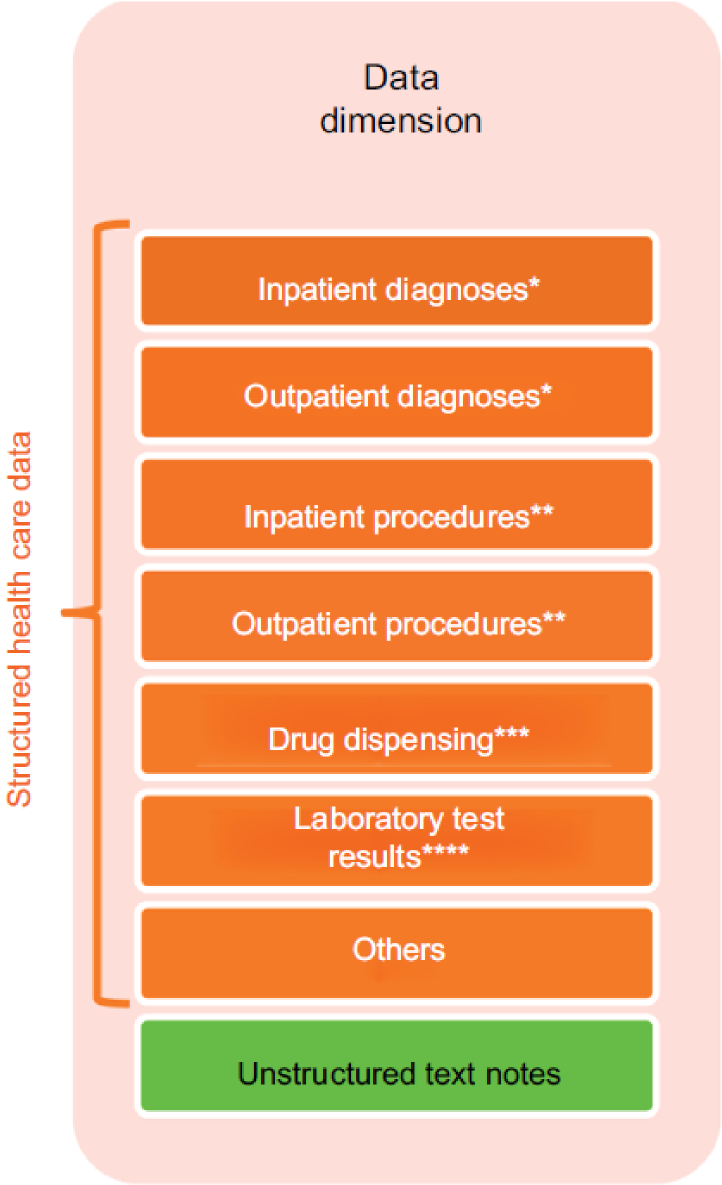
Y = outcome of interest

C = observable confounder (serves as proxy)

U = unobserved confounder

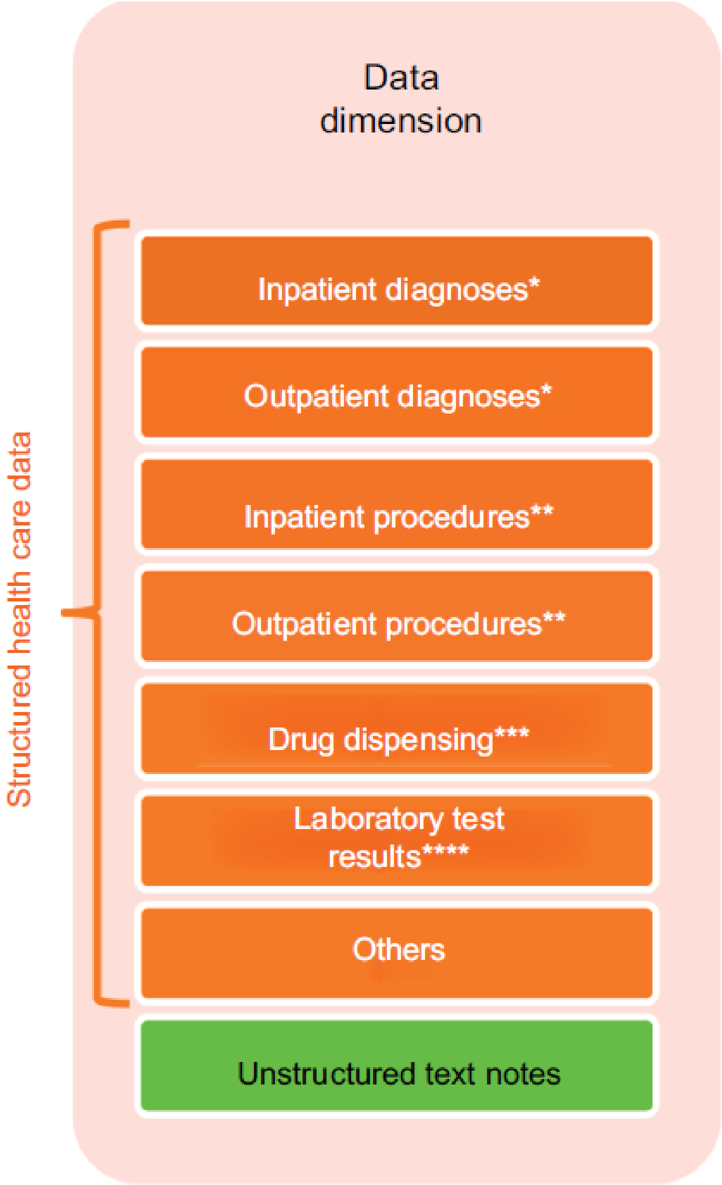
Unobserved confounder	Observable proxy measurement	Coding examples
Very frail health	Use of oxygen canister	CPT-4
Sick but not critical	Code for hypertension during a hospital stay	ICD-9, ICD-10
Health-seeking behavior	Regular check-up visit; regular screening examinations	ICD-9, CPT-4, #PCP visits

Proxy Confounder Adjustment



- How to generate/identify proxy variables?
  - **Manual:** expert knowledge to identify claims codes that are thought to capture important confounder information.
    - **Limitations:**
      - difficult to know all confounders a priori
      - difficult to determine which combination of codes best capture information on those confounders
  - **Kitchen sink approach:**
    - Balance all pre-treatment variables
      - Some dimension reduction is necessary in high-dimensional data
      - Harm efficiency and validity

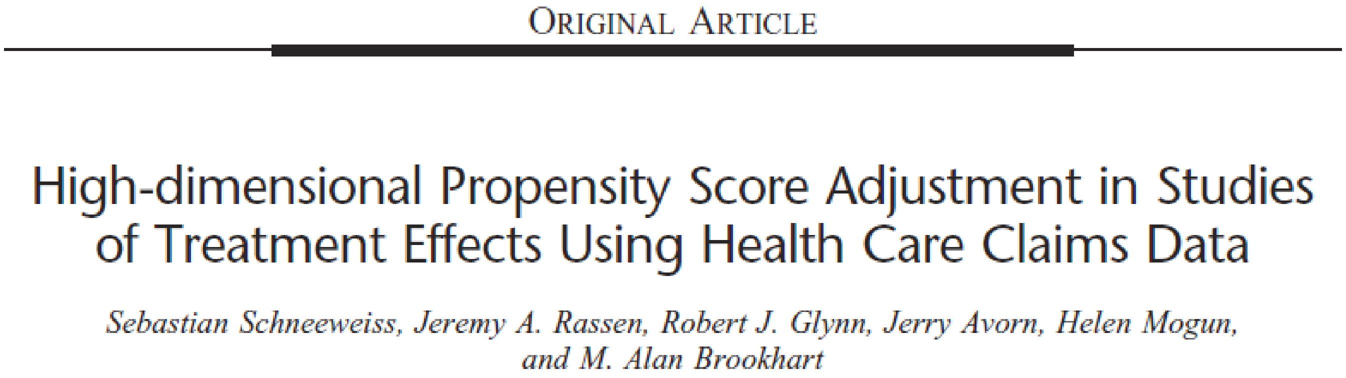
Proxy Confounder Adjustment



- How to generate/identify proxy variables?
  - **Can we use automated algorithms to identify "proxy confounders" based only on empirical associations observed in the data?**
  - **Automation:** use empirical associations and longitudinal coding patterns in the data to identify which codes likely contain important confounder information.

Automated Confounder Selection

- Schneeweiss et al. (2009) first to propose automated proxy confounder adjustment in healthcare claims databases.

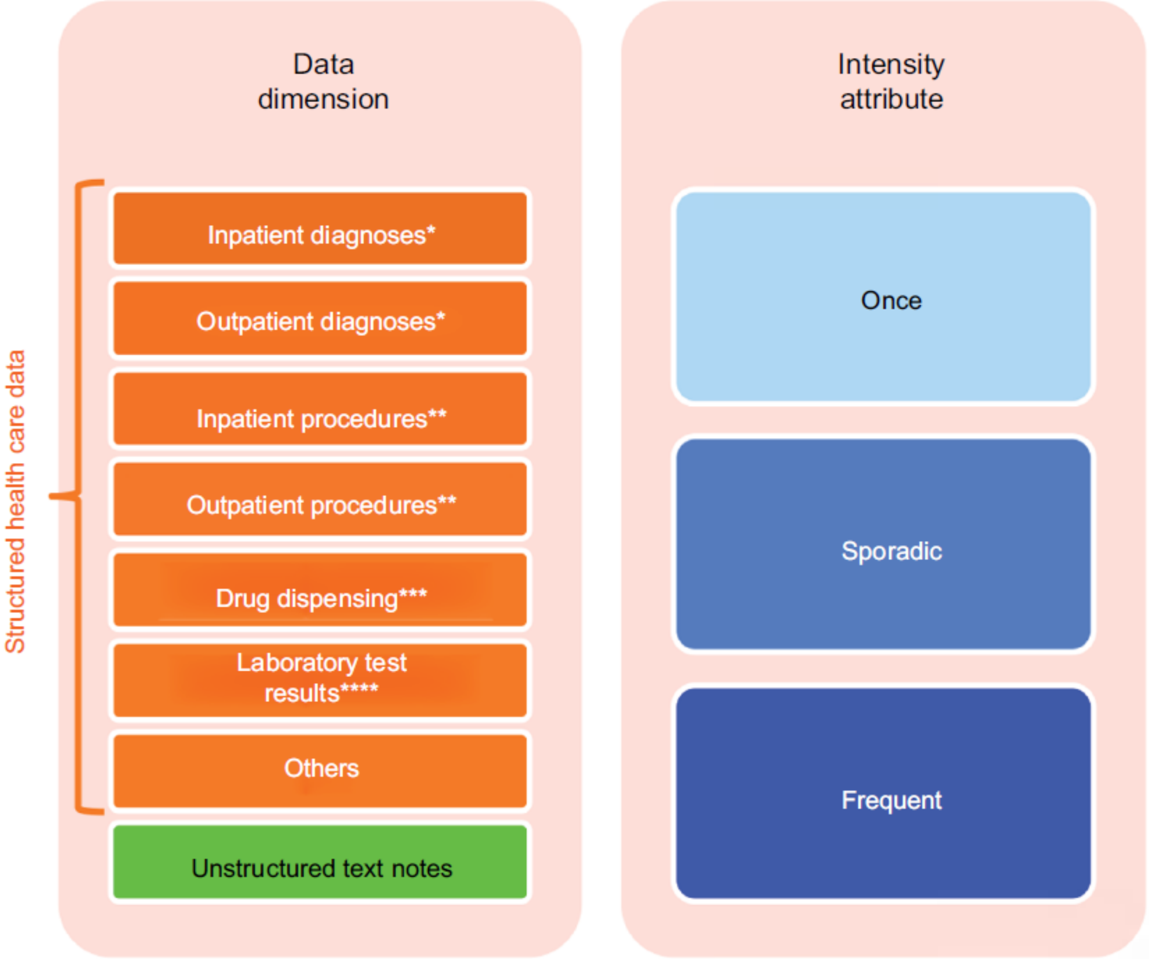


Original Article

High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data

Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart

High-dimensional propensity score



- Evaluates thousands of diagnostic and procedural claims codes.
- For each code creates three binary variables based on the frequency of occurrence
- Evaluates potential confounding impact of generated variables by looking at strength of association with both the treatment and outcome

Limitations of HDPS

- HDPS has proven useful across a wide range of applications (>700 citations on google scholar).
  - **Limitations**:
    - How many variables to adjust for?
    - Evaluates empirical associations based only on univariate associations with treatment and outcome.

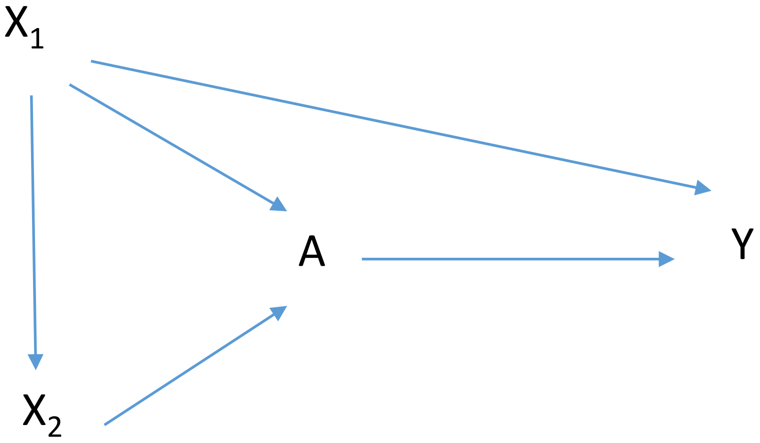


Figure. X_2 is independent of the outcome after conditioning on X_1

Limitations of HDPS

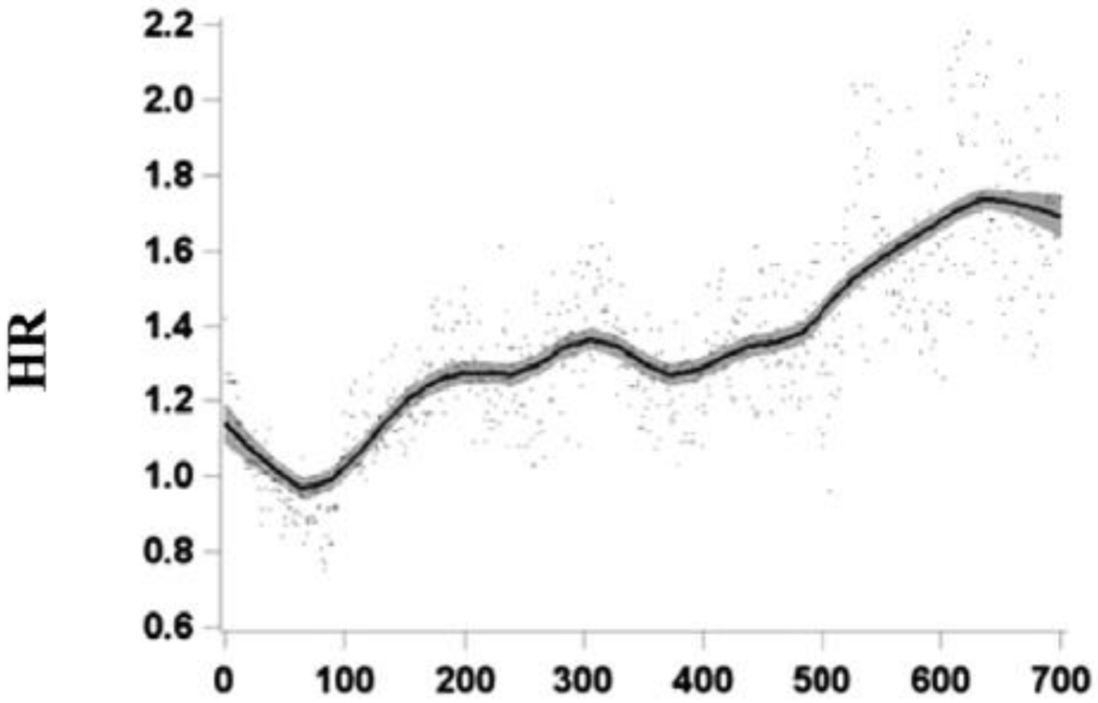


Figure. Published in Patorno et al. (2015) showing the relation of hazard ratio (HR) estimates to the number of empirical covariates sequentially added to the propensity score in a study comparing the effect of anticonvulsant medications on cardiovascular risk.

Limitations of HDPS

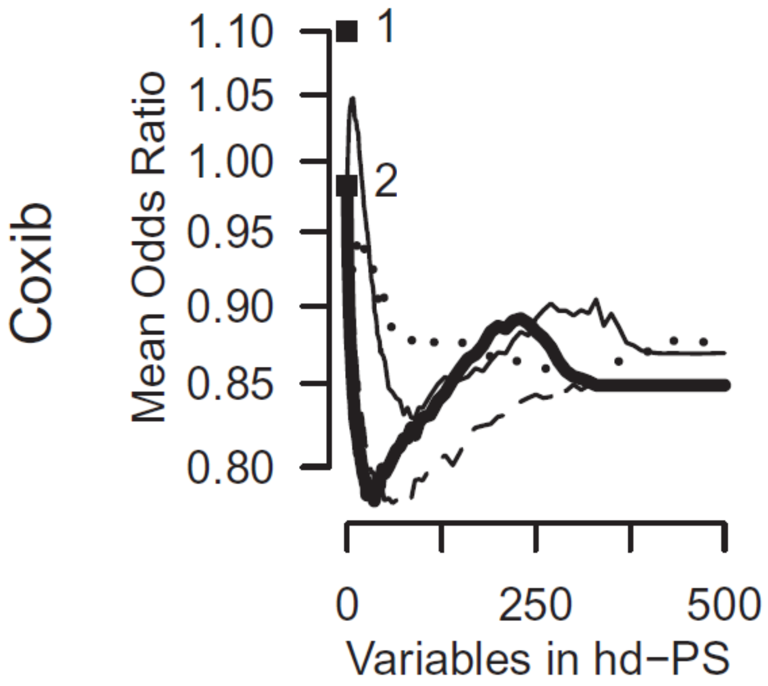


Figure. Published in Rassen et al. (2012) showing the relation of odds ratio estimates to the number of empirical covariates sequentially added to the propensity score in a study comparing the effect of Coxib vs NSAID on GI bleed.

Supplementing HDPS with ML

- Can data-adaptive algorithms (machine learning) help to improve proxy confounder adjustment?

- Machine learning algorithms have become very powerful for prediction modeling.
  - ML algorithms are good at identifying strong predictors of the outcome or treatment
  - Clear objective optimization rules for determining how algorithms should adapt to the data (e.g., cross-validated prediction error).

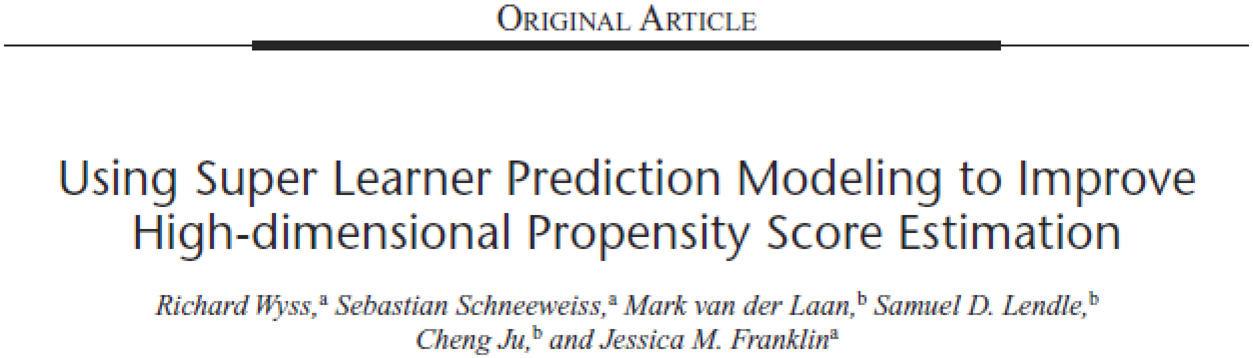
- Machine learning algorithms for causal inference is more challenging
  - No clear criteria on how to judge/optimize algorithm's performance:
  - Want to minimize bias in effect estimate but this cannot be directly measured

Machine Learning for Causal Inference

- Recent developments in machine learning for causal inference have been developed that can potentially improve automated confounding control.

- Common theme is that these methods use ML prediction algorithms in a way to consider the joint associations of covariates with both the treatment and outcome simultaneously.
    - Targeted Learning (TMLE and CTMLE)
    - Many variations of regularized regression

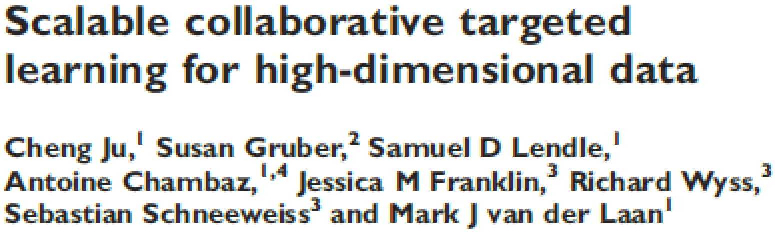
Using ML for proxy confounder adjustment



ORIGINAL ARTICLE

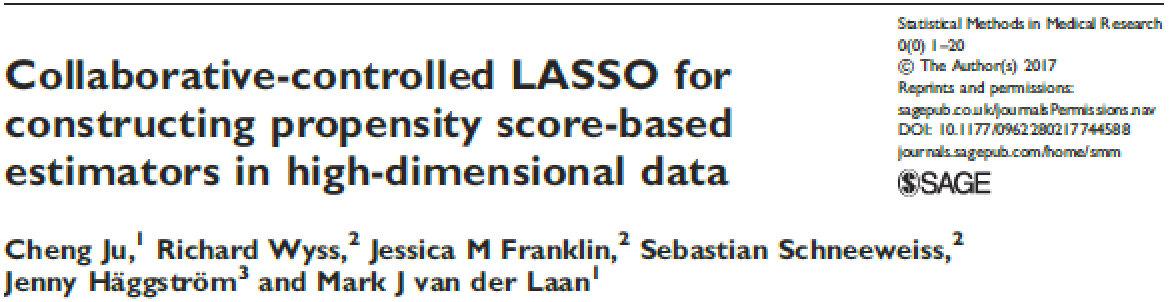
Using Super Learner Prediction Modeling to Improve High-dimensional Propensity Score Estimation

Richard Wyss,[a] Sebastian Schneeweiss,[a] Mark van der Laan,[b] Samuel D. Lendle,[b] Cheng Ju,[b] and Jessica M. Franklin[a]



Scalable collaborative targeted learning for high-dimensional data

Cheng Ju,[1] Susan Gruber,[2] Samuel D Lendle,[1] Antoine Chambaz,[1,4] Jessica M Franklin,[3] Richard Wyss,[3] Sebastian Schneeweiss[3] and Mark J van der Laan[1]



Statistical Methods in Medical Research
0(0) 1–20
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280217744588
journals.sagepub.com/home/smm
SAGE

Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data

Cheng Ju,[1] Richard Wyss,[2] Jessica M Franklin,[2] Sebastian Schneeweiss,[2] Jenny Häggström[3] and Mark J van der Laan[1]

Table 1. Methods for confounding control

Method #	Name	Description
1	HDPS 25	Logistic PS model controlling for 25 HDPS selected variables
2	HDPS 100	Logistic PS model controlling for 100 HDPS selected variables
3	HDPS 200	Logistic PS model controlling for 200 HDPS selected variables
4	HDPS 300	Logistic PS model controlling for 300 HDPS selected variables
5	HDPS 400	Logistic PS model controlling for 400 HDPS selected variables
6	HDPS 500	Logistic PS model controlling for 500 HDPS selected variables
7	HDPS SL	PSs were estimated by running the super-learner on the library of HDPS models listed above
8	CTMLE 10	CTMLE with HDPS pre-ordering and a patience parameter of 10
9	CTMLE 50	CTMLE with HDPS pre-ordering and a patience parameter of 50
10	HDPS Lasso	Ran a lasso regression on the outcome using 500 HDPS created variables as the predictors. Variables whose coefficients were shrunk to 0 were excluded. All other variables were included in a logistic PS model

Methods

- We used a "plasmode simulation" framework where empirical data is incorporated into the simulation structure (Franklin et al. 2014).
- We considered 3 datasets:
  - NSAID
    - Cox-2 inhibitor vs nonselective NSAID
  - NOAC
    - Oral anti-coagulant vs warfarin
  - STATIN
    - High vs. low intensity statin use
- We simulated the outcome as a function of 200 baseline variables and considered scenarios where we varied the sample size, outcome incidence, and treatment prevalence.

Plasmode Simulation Study

Table 2. Simulation Scenarios

Scenario	Description	Sample size	Tmt prevalence	Outcome incidence	Tmt effect
1	Base case	10,000	0.4	0.1	OR=1
2	Increased sample size	20,000	0.4	0.1	OR=1
3	Reduced sample size	5,000	0.4	0.1	OR=1
4	Rare outcome	10,000	0.4	0.02	OR=1
5	Reduced treatment prevalence	10,000	0.2	0.1	OR=1
6	Vary tmt effect	10,000	0.4	0.1	OR=2

Discussion

- Overfitting the PS model can affect confounding control in settings involving small samples or rare exposures
- Combining Super Learner with HDPS generated variables may be useful in improving the robustness of automated “proxy confounder adjustment”
- Variable selection using scalable CTMLE has nice theoretical properties, but is sensitive to the pre-ordering
  - Tended to underfit PS models
- **No single algorithmic approach for confounder selection and causal inference was optimal across all settings.**

Conclusions

- Empirical associations by themselves are not sufficient to completely characterize causal relations (Pearl 2013).
  - “causal inference from observational data requires prior causal assumptions or beliefs, which must be derived from subject matter knowledge, not from statistical associations detected in the data.” (Hernan et al. 2002)
- Data driven algorithms can help to reduce subjective decision making in healthcare database studies, but cannot completely eliminate it.
  - Expert knowledge remains a key component of healthcare database analyses.

R software

- Software for the methods discussed is available at
  - https://github.com/lendle/hdps
  - https://github.com/lendle/TargetedLearning.jl

- R code for producing plasmode simulations is available on CRAN
  - Package “Plasmode” (authors: Jessica Franklin, Younathan Abdia, Shirley Wang)

- Software is also available on our Division’s website
  - https://www.drugepi.org/dope/software#Pharmacoepidemiology

Acknowledgements

- Jessica Franklin
- Sebastian Schneeweiss
- Mark van der Laan
- Wesley Eddings
- Sam Lendle
- Cheng Ju

Thank you!