Causal inference in analyzing administrative healthcare data:

Can we integrate machine learning approaches within this framework?

Ehsan Karim
http://ehsank.com/

Feb 23, 2021; CHSPR

Outline

1. Notations and Questions
   - Regression vs. Propensity score (PS)
2. Health care databases
3. High-dimensional propensity score (hdPS)
   - The original mechanism
   - Machine learning-based hdPS
   - AI-based hdPS
   - Future directions
4. Examples in Canadian data

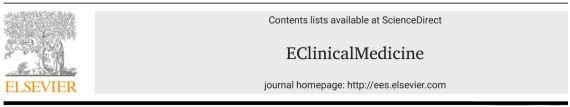
Reference Preview

Will be primarily discussing Schneeweiss (2018). Will also briefly mention Karim, Pang, and Platt (2018),Weberpals et al. (2021), Zivich and Breskin (2021).

- Schneeweiss (2018) Clinical Epidemiology
- Karim et al. (2018) Epidemiology
- Weberpals et al. (2021) Epidemiology
- Zivich and Breskin (2021) Epidemiology

Notations and Motivating Example

- Y = Outcome
  - Airway disease (among BC immigrants)
- A = Primary exposure
  - Respiratory tuberculosis
- $C = (C_1, \ldots, C_7)$ are covariates measured at baseline
  - age, sex, income, education, comorbidity score, TB incidence in birth country, year of residency in BC


Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: http://ees.elsevier.com

ELSEVIER

Research Paper

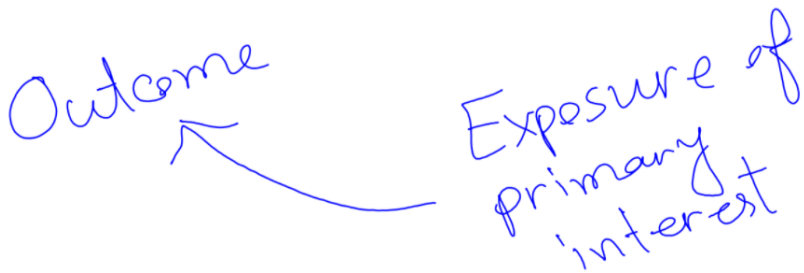
Post-tuberculosis airway disease: A population-based cohort study of people immigrating to British Columbia, Canada, 1985–2015

C.Andrew Basham [a,b,*], Mohammad E. Karim [a,c], Victoria J. Cook [b,d], David M. Patrick [a,b,d], James C. Johnston [a,b,d]

Questions

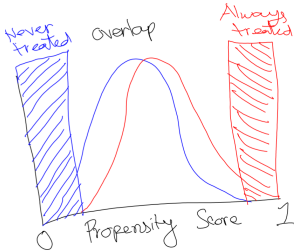
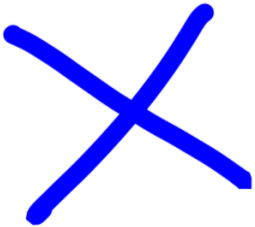
Inferential goals

1. Prediction, developing risk scores (predict Y)
2. Identifying important predictors (identify C_1, C_2 that are important to predict Y)
3. Descriptive, exploratory (is A and Y associated?)
4. Evaluating a predictor of primary interest (does A cause Y?)



Regression vs. Propensity score

- For RCT, adjustment may not be essential, as design takes care of bias sources.
- For RWE studies, more caution necessary. Adjustment of C is important (usually large #s).

Model / Method	Propensity score	Regression analysis
Exposure Modelling	$\text{PS} = P(A = 1\|C) = f(C)$	No design stage analysis
		
Outcome Modelling	$E(Y\|A, C) = g(A, C)$	$E(Y\|A, C) = g(A, C)$

Health care databases

Why use non-randomized data at all?

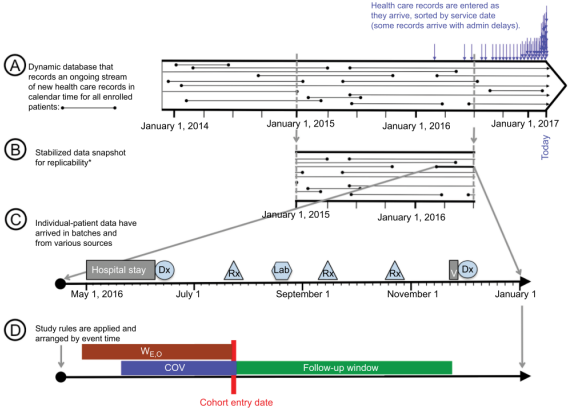
- clinical health records
  - medical facility
  - hospital
  - clinic and practice.
- administrative data
  - PharmaNet
  - Medical Services Plan
- clinical registries
  - TB registry
  - MS registry

Health care databases: Advantages

- larger sample size
- diverse population
- longitudinal records over many years
- detailed
  - health encounters,
  - comorbidity history,
  - drug exposure history
- possibility to link other databases
  - Immigration
  - Vital Statistics

Health care databases: Study implementation

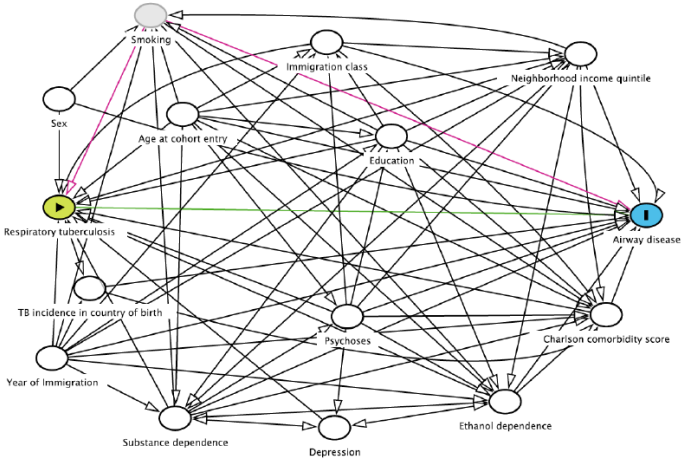
Schneeweiss (2018): freeze a data cut from dynamic data stream, encounters collected at covariate assessment period and follow-up, apply rules/algorithms to define variables.



Health care databases: Limitations

- Not specifically designed for answering a particular research question.
- Data sparsity:
  - no defined or routine interview dates
  - data collection relies on visits and encounters
- Investigators have no control over which factors were measured during the data collection stage.
  - smoking in TB
  - MRI data in MS

Health care databases: Limitations



Methods: Expectation vs. Reality (Part I: unmeasured confounder)

Exposure Model

$$P(A = 1|C) = f(C)$$
$$= \frac{1}{1 + exp[\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \alpha_6 C_6 + \alpha_7 C_7]}$$

Outcome Model

$$E(Y|A, C) = g(A, C)$$
$$= \beta_0 + \psi A + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4 + \beta_5 C_5 + \beta_6 C_6 + \beta_7 C_7$$

Assumption

$$Y|A, C \sim N[E(Y|A, C), \sigma^2]$$

Methods: Expectation vs. Reality (Part 2: model misspecification)

Exposure Model

$$P(A = 1|C) = f(C)$$
$$= \frac{1}{1 + exp[\alpha_0 + \alpha_1(C_2 \times C_3) + \alpha_2(C_4^2 \times \frac{\exp C_5}{5 \times C_7}) + \alpha_3 C_6]}$$

Outcome Model

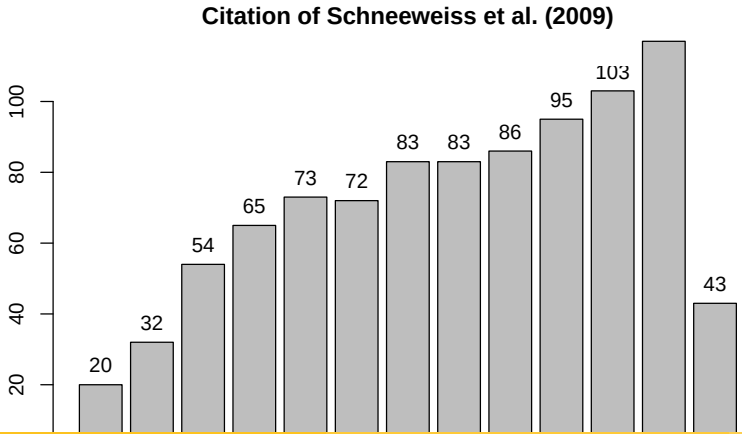
$$E(Y|A, C) = g(A, C)$$
$$= \beta_0 + \psi A + \beta_1 C_1^2 + \beta_2(C_2 \times C_3 \times C_7) + \beta_3 \frac{\exp C_4}{C_5 \times 2}$$

Assumption

$$Y|A, C \sim N[E(Y|A, C), \sigma^2]$$

hdPS

```
## Warning: package 'scholar' was built under R version 4.1.3
```



hdPS proxy collection

Schneeweiss (2018):

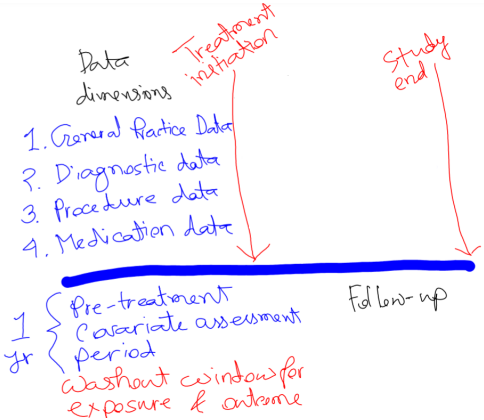
Unobserved confounder	Observable proxy measurement	Coding examples
Very frail health	Use of oxygen canister	CPT-4
Sick but not critical	Code for hypertension during a hospital stay	ICD-9, ICD-10
Health-seeking behavior	Regular check-up visit; regular screening examinations	ICD-9, CPT-4, #PCP visits
Fairly healthy senior	Receiving the first lipid-lowering medication at age 70 years	NDC, ATC, Read
Chronically sick	Regular visits with specialist, hospitalization; many prescription drugs	#specialist visits, NDC, ATC
Outcome surveillance intensity	General markers for health care utilization intensity	#visits, #different drugs

They were the first to propose adjusting for something that **may not be interpretable** directly with the context of the research question. Logic was two fold:

1. variables from same subject should be **correlated** = has relevant information
2. adjust items that are **predictive of outcome** (as established in PS literature)

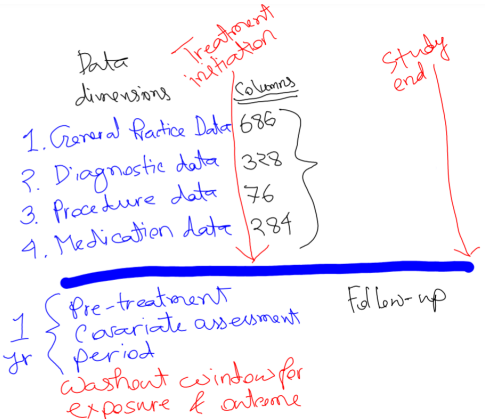
hdPS proxy collection from same subjects (1)

Collection of proxy data for the unmeasured + mis-measured variables (Karim, Pang, and Platt 2018)



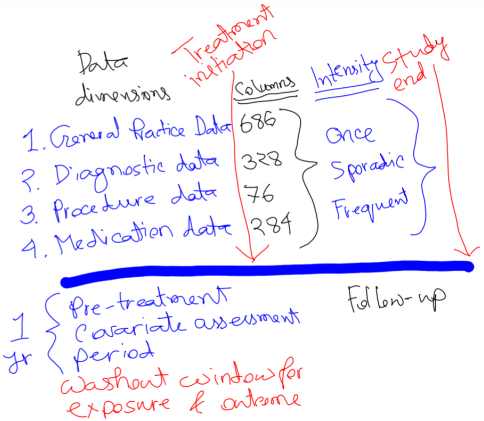
hdPS proxy collection from same subjects (1)

Collection of proxy data for the unmeasured + mis-measured variables (Karim, Pang, and Platt 2018)



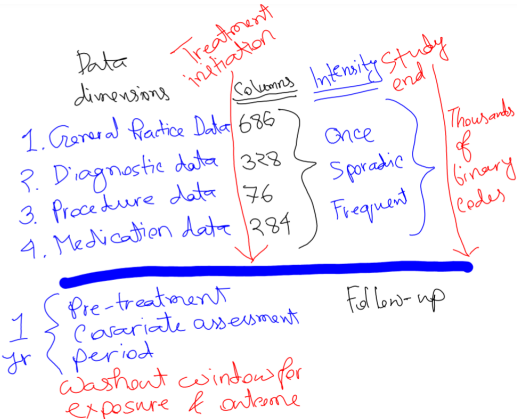
hdPS proxy collection from same subjects (1)

Collection of proxy data for the unmeasured + mis-measured variables (Karim, Pang, and Platt 2018)



hdPS proxy collection from same subjects (1)

Collection of proxy data for the unmeasured + mis-measured variables (Karim, Pang, and Platt 2018): but restricted to 2,400 empirical covariates EC (based on high prevalence)



hdPS proxy collection from same subjects (1)

List of additional proxy variables:

Dimension 1 Practice	Dimension 2 Diagnostic	Dimension 3 Procedure	Dimension 4 Medication
EC-dim1-1-once	EC-dim2-1-once	EC-dim3-1-once	EC-dim4-1-once
EC-dim1-1-sporadic	EC-dim2-1-sporadic	EC-dim3-1-sporadic	EC-dim4-1-sporadic
EC-dim1-1-frequent	EC-dim2-1-frequent	EC-dim3-1-frequent	EC-dim4-1-frequent
...	...	...	
EC-dim1-686-frequent	EC-dim2-328-frequent	EC-dim3-76-frequent	EC-dim4-284-frequent

4 dimension $\times$ 3 intensity $\times$ 200 most prevalent codes = 2,400 ECs

hdPS mechanism: find EC as h(outcome, exposure prevalence) (2)

Assumption:

- $p_{u=1,a=1}$ = prevalence of unmeasured confounder among treated ($A = 1$)
- $p_{u=1,a=0}$ = prevalence of unmeasured confounder among untreated ($A = 0$)
- $p_{u=1,y=1}$ = prevalence of unmeasured confounder among dead ($Y = 1$)
- $p_{u=1,y=0}$ = prevalence of unmeasured confounder among alive ($Y = 0$)

Bross (1966) formula says, the amount of bias due to u is

$$\text{Bias}_M = \frac{p_{u=1,a=1} \times \left(\frac{p_{u=1,y=1}}{p_{u=1,y=0}} - 1\right) + 1}{p_{u=1,a=0} \times \left(\frac{p_{u=1,y=1}}{p_{u=1,y=0}} - 1\right) + 1}$$

hdPS mechanism: find EC as h(outcome, exposure prevalence) (2)

~~Assumption~~ Calculate:

- $p_{EC=1,a=1}$ = prevalence of ~~unmeasured confounder~~ EC among treated ($A = 1$)
- $p_{EC=1,a=0}$ = prevalence of ~~unmeasured confounder~~ EC among untreated ($A = 0$)
- $p_{EC=1,y=1}$ = prevalence of ~~unmeasured confounder~~ EC among dead ($Y = 1$)
- $p_{EC=1,y=0}$ = prevalence of ~~unmeasured confounder~~ EC among alive ($Y = 0$)

Bross (1966) formula says, the amount of bias due to EC is

$$\text{Bias}_M = \frac{p_{EC=1,a=1} \times \left(\frac{p_{EC=1,y=1}}{p_{EC=1,y=0}} - 1\right) + 1}{p_{EC=1,a=0} \times \left(\frac{p_{EC=1,y=1}}{p_{EC=1,y=0}} - 1\right) + 1}$$

hdPS: select hdPS variables from ECs

1. Rank (descending) each EC by the magnitude of log-bias: $|\log \text{Bias}_M|$

Rank by bias	$\lvert\log \text{Bias}_M\rvert$	EC
1	0.42	EC-dim1-21-once
2	0.32	EC-dim2-95-once
3	0.25	EC-dim4-289-once
...	...	...
2,400	0.01	EC-dim4-64-frequent

2. Take top 500 of these ECs. These are hdPS variables.

hdPS: estimate treatment effect

Exposure Model (investigator-specified covariates)

$$P(A = 1|C) = f(C)$$
$$= \frac{1}{1 + exp[\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \alpha_6 C_6 + \alpha_7 C_7]}$$

Outcome Model (in the PS matched or weighted data)

$$E(Y|A, C) = g(A, C)$$
$$= \beta_0 + \psi A$$

hdPS: estimate treatment effect

Exposure Model (investigator-specified covariates + hdPS variables [EC])

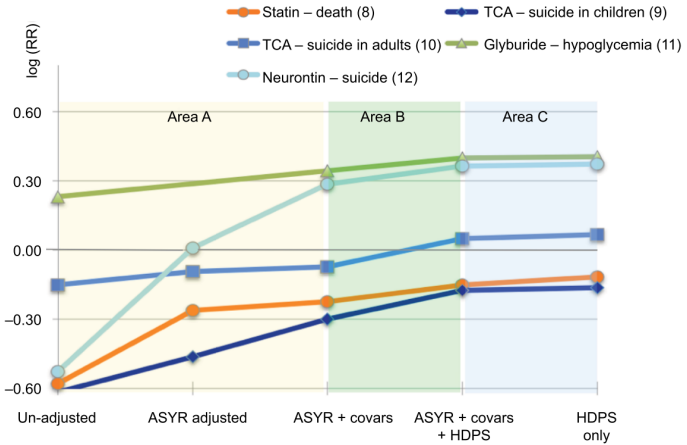
$$P(A = 1|C) = f(C)$$
$$= \frac{1}{1 + exp[\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \alpha_6 C_6 + \sum_{i=1}^{500} \alpha_i' EC_i]}$$

Outcome Model (in the PS matched or weighted data)

$$E(Y|A, C) = g(A, C)$$
$$= \beta_0 + \psi A$$

hdPS: estimate treatment effect

Performance of hdPS. Schneeweiss (2018):



ML-hdPS: deal with collinearity

Kitchen SSink Exposure Model (A ~ C + EC)

$$P(A = 1|C, EC) = \frac{1}{1 + exp[\alpha_0 + \alpha_1 C_{\text{important}} + \alpha_2 C_{\text{potential confounder}} + \sum_{i=1}^{2,400} \alpha_i' \text{EC}_i]}$$

Outcome Model for EC selection (Y ~ C + top-500 ECs)

$$f(Y|C, EC) = \alpha_0 + \alpha_1 C_{\text{important}} + \alpha_2 C_{\text{potential confounder}} + \sum_{i=1}^{500} \alpha_i' \text{EC}_i$$

ML-hdPS: deal with collinearity

hdPS Exposure Model (A ~ C + top-ranked EC)

$$P(A=1|C,EC) = \frac{1}{1+exp[\alpha_0 + \alpha_1 C_{\text{important}} + \alpha_2 C_{\text{potential confounder}} + \sum_{i=1}^{\text{top } 500} \alpha_i' EC_i]}$$

Outcome Model for EC selection (Y ~ C + ECs)

$$f(Y|C,EC) = \alpha_0 + \alpha_1 C_{\text{important}} + \alpha_2 C_{\text{potential confounder}} + \sum_{i=1}^{2{,}400} \alpha_i' EC_i$$

ML-hdPS: deal with collinearity

Exposure Model (investigator-specified covariates + hdPS variables [EC])

$$P(A=1|C) = \frac{1}{1+exp[\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \alpha_6 C_6 + \sum_{i=1}^{500} \alpha_i' EC_i]}$$

ML extensions: exposure model vs. outcome model

- ECs selected separately
  - are usually highly correlated, and
  - has inflated variance.
- Karim, Pang, and Platt (2018) used elasic net and LASSO to reduce the number of selected hdPS variables (EC variables).
  - found that the hybrid method (elastic net of hdPS) performs better than ML or hdPS.
  - quality of proxy information matters.

ML-hdPS: deal with model misspecification

Exposure Model

$$P(A=1|C) = \frac{1}{1 + exp[\alpha_0 + \alpha_1(C_2 \times C_3) + \alpha_2(C_4^2 \times \frac{\exp C_5}{5 \times C_7}) + \alpha_3 C_6]}$$

Outcome Model

$$E(Y|A,C) = \beta_0 + \psi A + \beta_1 C_1^2 + \beta_2(C_2 \times C_3 \times C_7) + \beta_3 \frac{\exp C_4}{C_5 \times 2}$$

More ML extensions

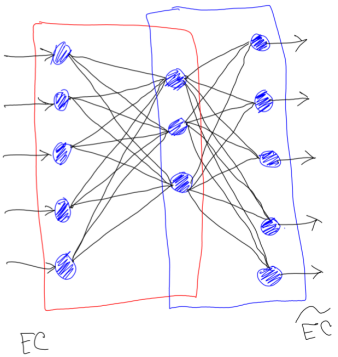
- Tree-based models: automatically detect function form
  - Karim, Pang, and Platt (2018) used random forest, and chose top important ECs
- Superlearner (ensemble learner) with tree-based + parametric models
- Double robust methods with Superlearner

AI-hdPS: towads dimension reduction

- Principal component (PC) analysis
  - compute linear transformation of all covariates to PCs, and top few PCs are selected to reduce dimension.
  - extract components of each variable resoponsible for most variance
  - incapable of learning non-linear feature representations

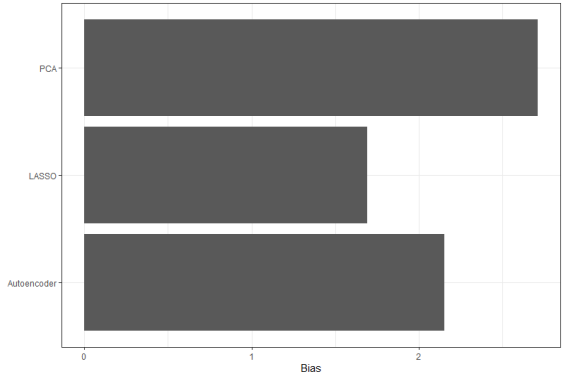
AI-hdPS: towads dimension reduction

- Weberpals et al. (2021) proposed to use autoencoders (3, 5, 7 layers) to reduce EC dimensions:



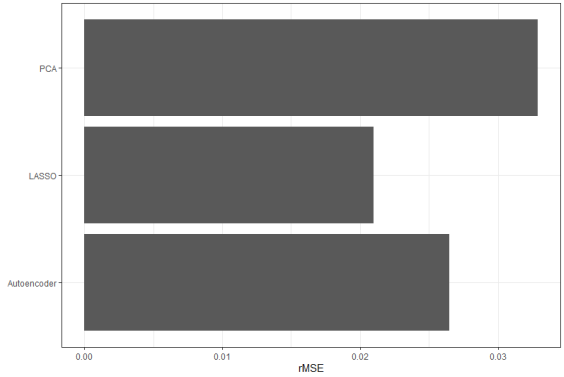
AI-hdPS: towads dimension reduction

% Bias



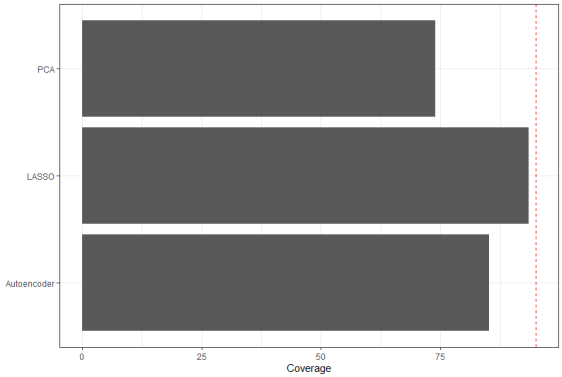
AI-hdPS: towads dimension reduction

rMSE



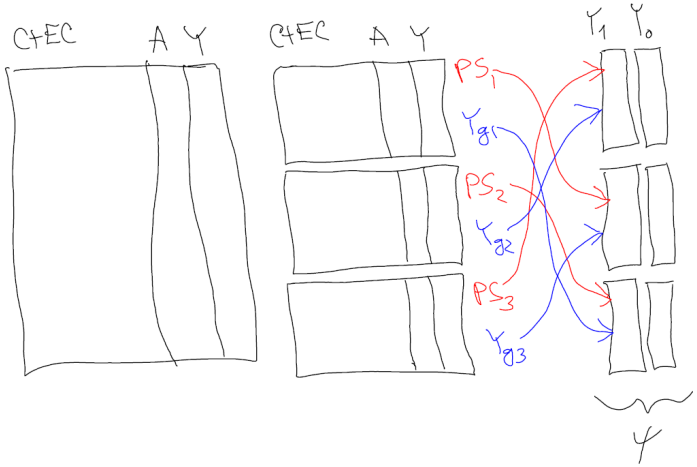
AI-hdPS: towads dimension reduction

Coverage: slower rate of convergence for non-parametric methods



Future directions

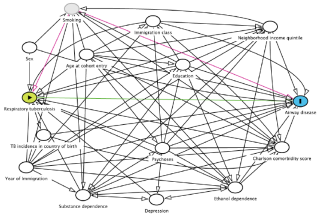
Zivich and Breskin (2021): Cross-fitting + together with double-robust approaches



A purist view

- Bias-based hdPS is not really PS!
  - design stage vs. analysis stage
  - hdPS uses corr(EC, outcome).
- Motivation of PS and hdPS are different to begin with
  - PS requires no unmeasured confounding
- I prefer to use hdPS as a sensitivity analysis.

Examples of hdPS (1)



Airway example: analysis approaches that were compared: hdPS as sensitivity analyses:

- PS decile-adjusted (investigator-specified covariates)
- hdPS decile-adjusted (investigator-specified + EC)
- LASSO-hdPS decile-adjusted (investigator-specified + LASSO-reduce EC)
- Others: E-value, other proxy variable adjustment

Conclusions were similar.

Examples of hdPS (2)

A 2017 JAMA paper

JAMA | Original Investigation

Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children

Hilary K. Brown, PhD; Joel G. Ray, MD, MSc, FRCPC; Andrew S. Wilton, MSc; Yona Lunsky, PhD, CPsych; Tara Gomes, MHSc; Simone N. Vigod, MD, MSc, FRCPC

Method	HR	CI 95%
Unadjusted	2.16	1.64-2.86
Multivariable adjusted	1.59	1.17-2.17
IPTW hdPS	1.61	0.997-2.59
1-1 hdPS matching	1.64	1.07-2.53
Pre-pregnancy data	1.85	1.37-2.51

Conclusion: **not associated**!

Reference

Karim, Mohammad Ehsanul, Menglan Pang, and Robert W Platt. 2018. "Can We Train Machine Learning Methods to Outperform the High-Dimensional Propensity Score Algorithm?" *Epidemiology* 29 (2): 191–98.

Schneeweiss, Sebastian. 2018. "Automated Data-Adaptive Analytics for Electronic Healthcare Data to Study Causal Treatment Effects." *Clinical Epidemiology* 10: 771–78.

Weberpals, Janick, Tim Becker, Jessica Davies, Fabian Schmich, Dominik Rüttinger, Fabian J Theis, and Anna Bauer-Mehren. 2021. "Deep Learning-Based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-Scale, Real-World Data Study." *Epidemiology, Doi: 10.1097/EDE.0000000000001338*.

Zivich, Paul N, and Alexander Breskin. 2021. "Machine Learning for Causal Inference: On the Use of Cross-Fit Estimators." *Epidemiology, Doi: 10.1097/EDE.0000000000001332*.

Thank you!