hdPS adjustment for analyzing electronic healthcare data:

Overview, recent advances & open questions

Ehsan Karim

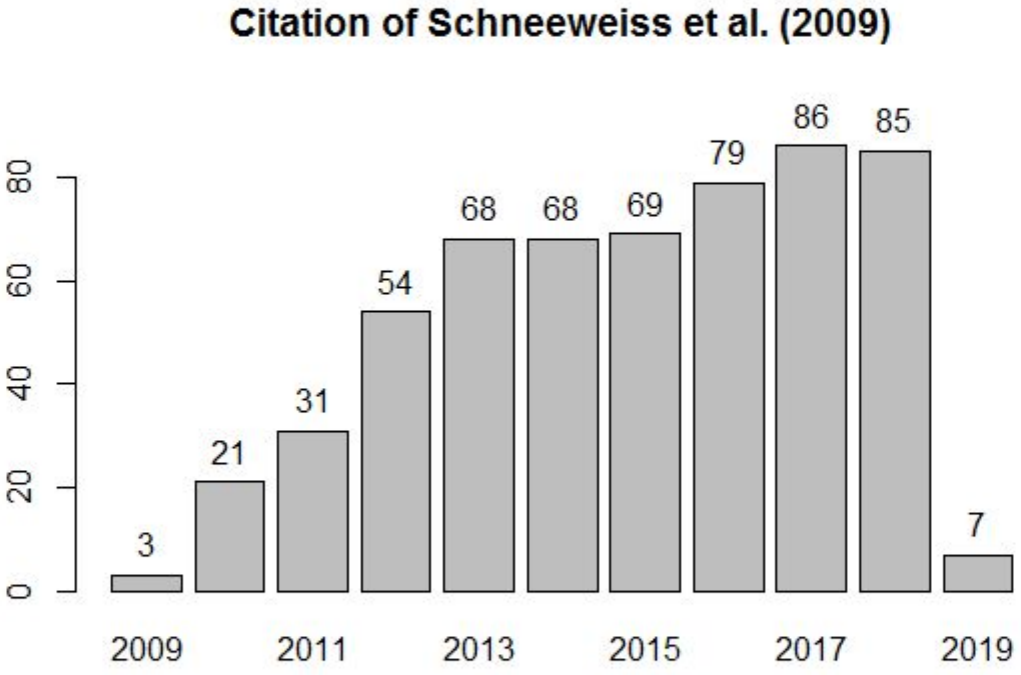
ehsan.karim@ubc.ca

Feb 19, 2019; BIRS

Popularity of hdPS

- High-dimensional propensity score (hdPS)
- Unmeasured confounding

High-dimensional propensity score adjustment in studies of treatment effects using health care claims data
S Schneeweiss, JA Rassen, RJ Glynn... - Epidemiology (..., 2009 - ncbi.nlm.nih.gov
Background Adjusting for large numbers of covariates ascertained from patients' health care claims data may improve control of confounding, as these variables may collectively be proxies for unobserved factors. Here we develop and test an algorithm that empirically ...
Cited by 353 Related articles All 14 versions Cite Save More



General Idea of hdPS

Administering health care data:

Longitudinal patient records: diagnostic and procedural information

Proxy measures of U

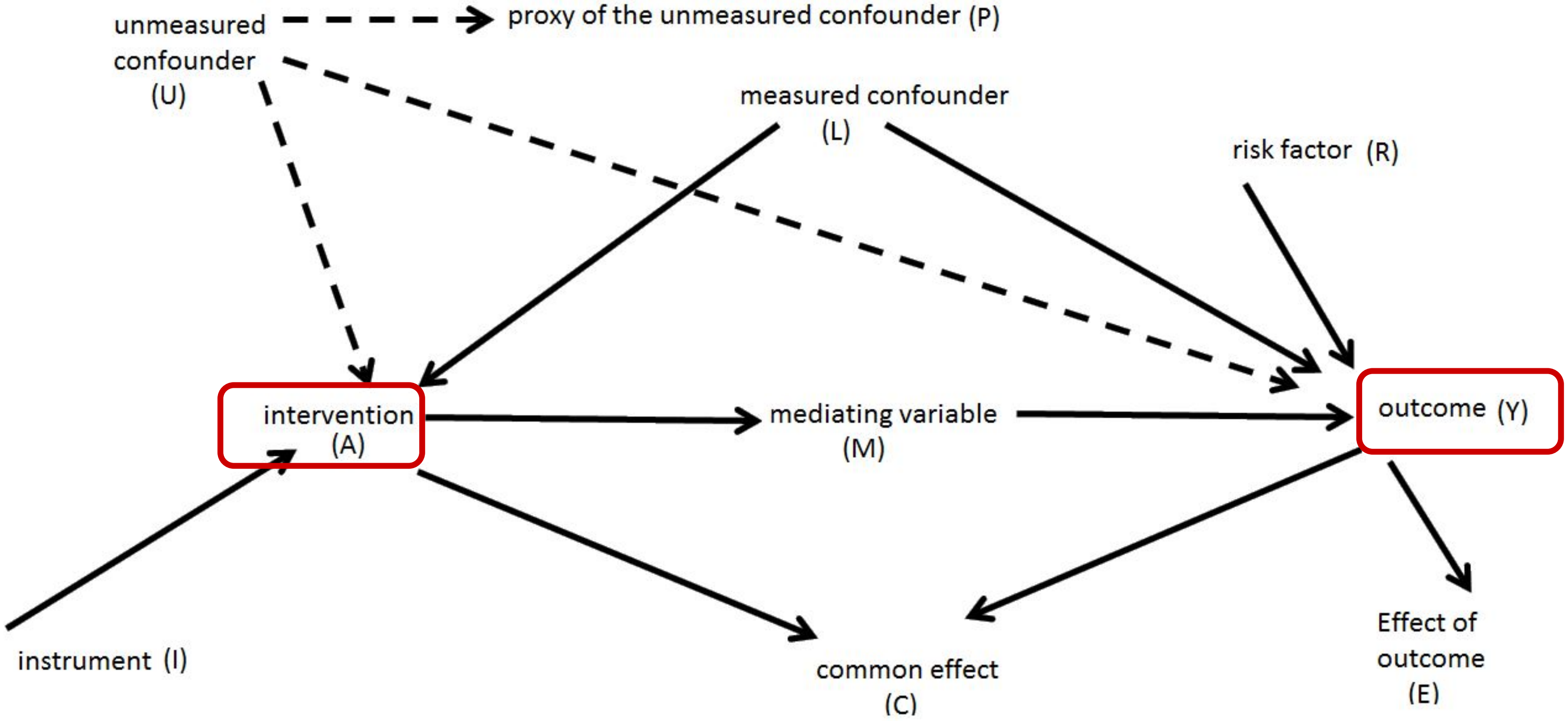
[HTML] Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects

S Schneeweiss - Clinical epidemiology, 2018 - ncbi.nlm.nih.gov

Background Decision makers in health care increasingly rely on nonrandomized database analyses to assess the effectiveness, safety, and value of medical products. Health care data scientists use data-adaptive approaches that automatically optimize confounding control to study causal treatment effects. This article summarizes relevant experiences and extensions. Methods The literature was reviewed on the uses of high-dimensional propensity score (HDPS) and related approaches for health care database analyses, including …

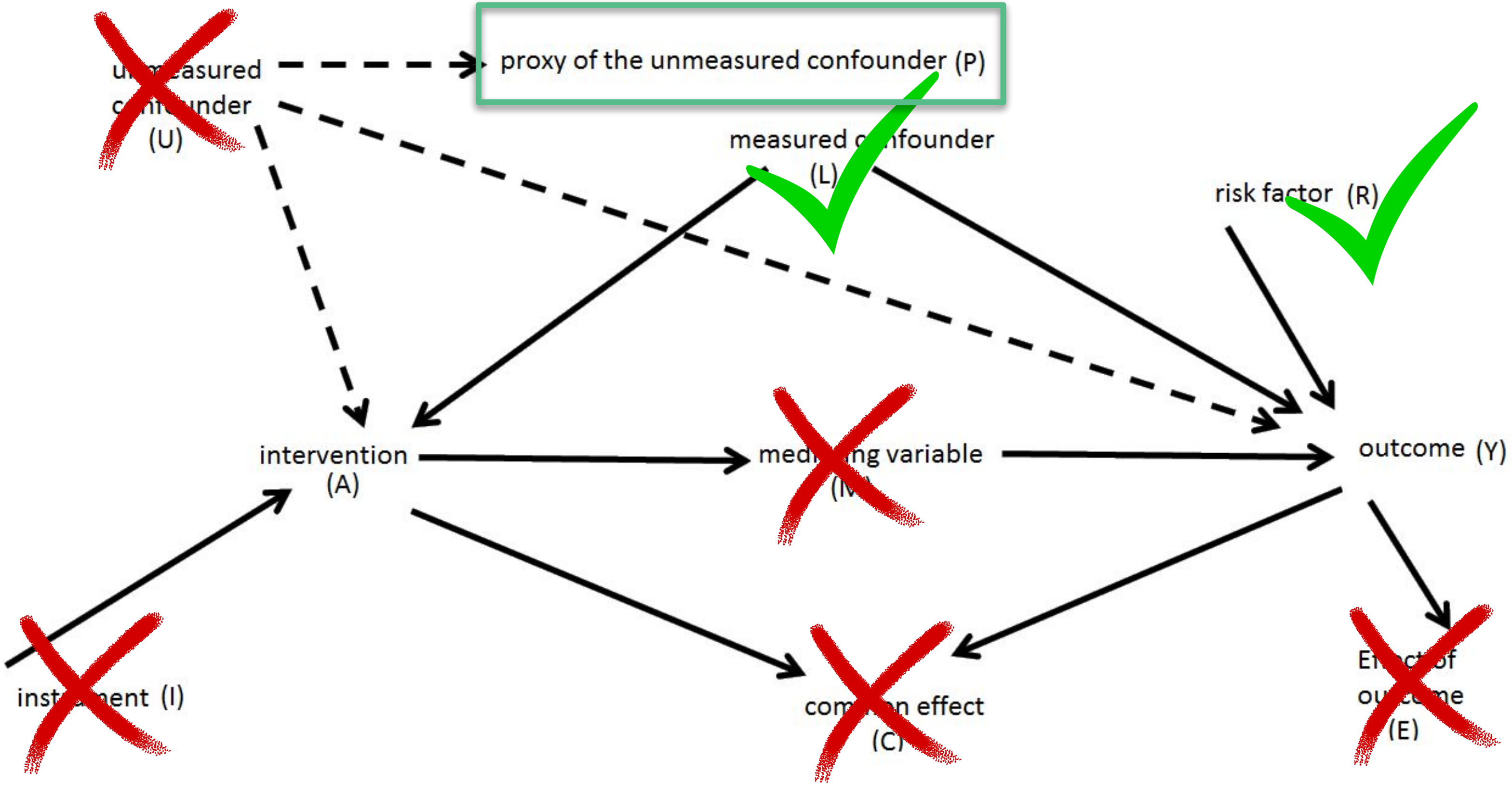
Unobserved confounder	Observable proxy measurement	Coding examples
Very frail health	Use of oxygen canister	CPT-4
Sick but not critical	Code for hypertension during a hospital stay	ICD-9, ICD-10
Health-seeking behavior	Regular check-up visit; regular screening examinations	ICD-9, CPT-4, #PCP visits
Fairly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

Type of variables



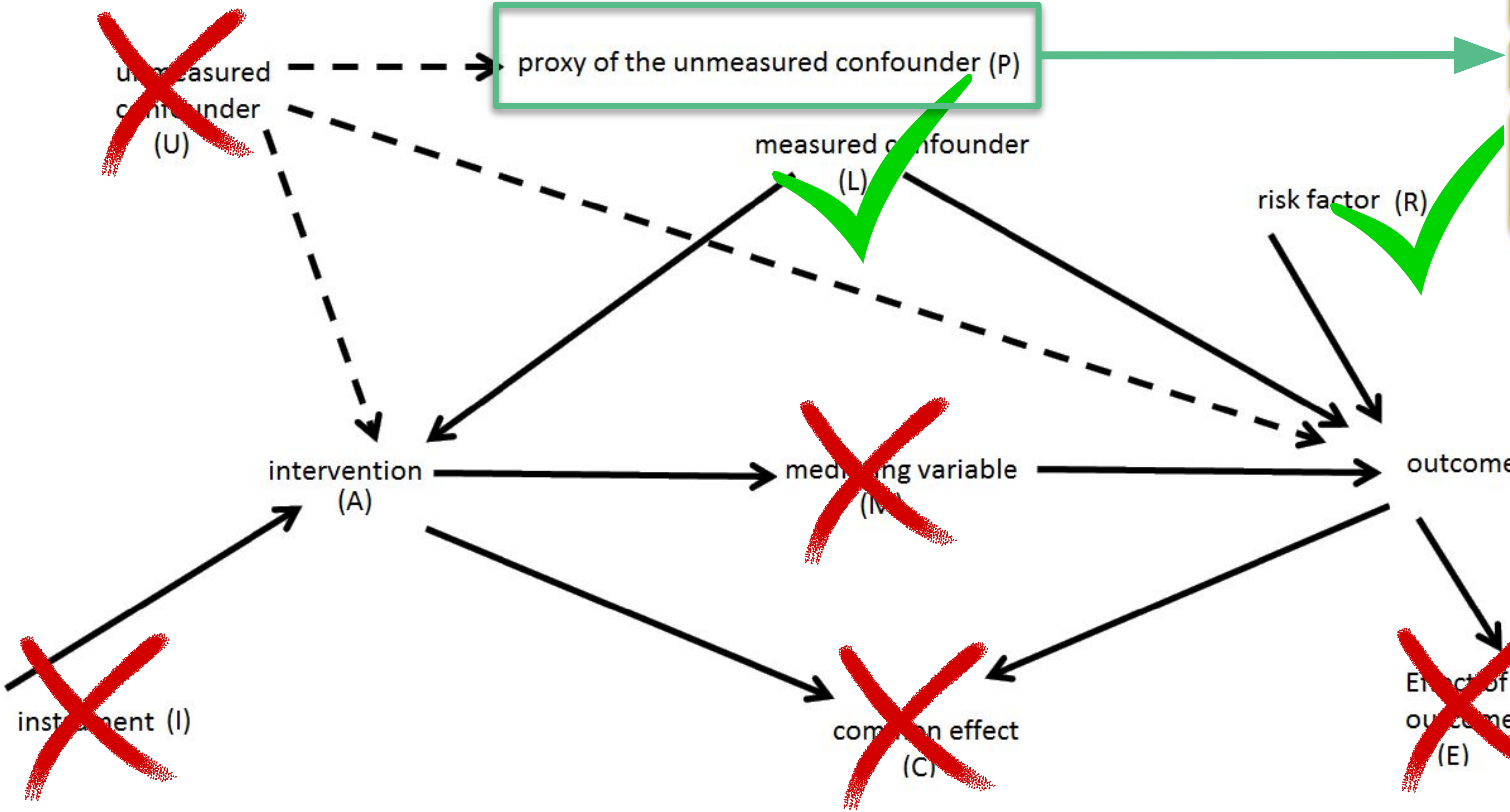
Investigator specified covariates: L + R

Type of variables



Type of variables

Investigator specified covariates: L + R
High-dimensional covariates: P
4 data dim×200×3=2,400 binary variables



Amount of confounding due to an unmeasured confounder

1. Assumptions:
   » p_{u1} = prevalence among treated
   » p_{u0} = prevalence among untreated
   » p_{uY1} = prevalence among dead
   » p_{uY0} = prevalence among alive
   » $RR_{uY} = p_{uY1} \,/\, p_{uY0}$
2. Adjusted RR:
   » $RR_{adj} = RR_{obs}\boxed{\frac{p_{u1}(RR_{uY}-1)+1}{p_{u0}(RR_{uY}-1)+1}}$
3. Amount of Bias / confounding due to u:
   » $Bias_M = \boxed{\frac{p_{u1}(RR_{uY}-1)+1}{p_{u0}(RR_{uY}-1)+1}}$

Let
Y = outcome (dead/alive)
A = treatment
RR_{obs} = Crude RR

u = unmeasured confounder (say, healthy eating)
RR_{adj} = ??

Bross formula 1966

Spurious effects from an extraneous variable
IDJ **Bross** - Journal of chronic diseases, **1966** - Elsevier
Abstract Spurious effects from an extraneous variable are a troublesome problem in many areas of research in the biological and behavioral sciences. While investigators have recognized intuitively that there is a relationship between the size of an effect and its chance of being spurious, current textbooks do not contain any explicit statement of this relationship. In this paper one such statement, the Size Rule, is developed. The application of this rule ...

Prioritization in hdPS

Erratum: High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data
R Wyss, B Fireman, JA Rassen, S Schneeweiss - Epidemiology, 2018 - journals.lww.com
Claims Data," the authors introduce a semiautomated variable selection algorithm for high-dimensional proxy adjustment within insurance health care claims databases. 1 The high-dimensional propensity score (HDPS) algorithm evaluates thousands of diagnostic, procedural, and medication claims codes and, for each code, generates binary variables based on the frequency of occurrence for each code during a defined pre-exposure covariate assessment period. The HDPS then prioritizes or ranks each variable based on its ...

1. ~~Assumptions:~~
   - » p_{u1} = prevalence among treated
   - » p_{u0} = prevalence among untreated
   - » p_{uY1} = prevalence among dead
   - » p_{uY0} = prevalence among alive
   - » $RR_{uY} = p_{uY1} / p_{uY0}$
2. Adjusted RR:
   - » $RR_{adj} = RR_{obs} \boxed{\frac{p_{u1}(RR_{uY}-1)+1}{p_{u0}(RR_{uY}-1)+1}}$
3. Amount of Bias / confounding due to u
   - » $Bias_M = \boxed{\frac{p_{u1}(RR_{uY}-1)+1}{p_{u0}(RR_{uY}-1)+1}}$

Bross formula 1966

Spurious effects from an extraneous variable
IDJ Bross - Journal of chronic diseases, 1966 - Elsevier
Abstract Spurious effects from an extraneous variable are a troublesome problem in many areas of research in the biological and behavioral sciences. While investigators have recognized intuitively that there is a relationship between the size of an effect and its chance of being spurious, current textbooks do not contain any explicit statement of this relationship. In this paper one such statement, the Size Rule, is developed. The application of this rule ...

Replace u by p (binary) and calculate

Amount of Bias / confounding due to p:

» $Bias_M = \boxed{\frac{p_{p1}(RR_{pY}-1)+1}{p_{p0}(RR_{pY}-1)+1}}$

Sort based on magnitude of rank-score (descending):

» $|\mathrm{Log}(Bias_M)|$ [biased-based, Bross 1966]

PS model based on
- L
- R
- Select p variables (top 500)

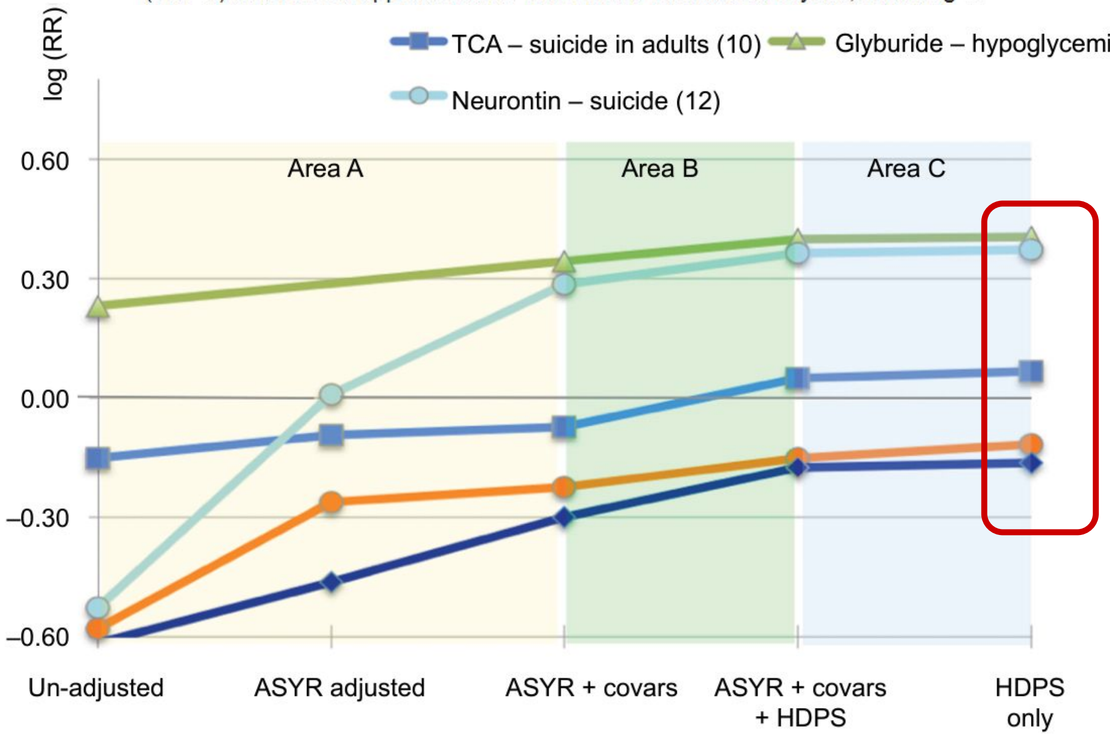
Performance of hdPS

[HTML] Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects

S Schneeweiss - Clinical epidemiology, 2018 - ncbi.nlm.nih.gov

Background Decision makers in health care increasingly rely on nonrandomized database analyses to assess the effectiveness, safety, and value of medical products. Health care data scientists use data-adaptive approaches that automatically optimize confounding control to study causal treatment effects. This article summarizes relevant experiences and extensions. Methods The literature was reviewed on the uses of high-dimensional propensity score (HDPS) and related approaches for health care database analyses, including …

Sequential addition of covariates vs. **change in effect estimate**



“This strongly suggests that even without the investigator-specifying covariates for adjustment, the algorithm alone optimizes confounding adjustment.”

Limitations / Extensions

- Bivariate adjustment
  - (multivariate adjustment instead of Bross?)
  - (Collinearity?: Ridge/LASSO)
- Mis-specification
  - (double robust/TMLE, SL)
- Time-varying covariates
  - (MSM)
- IV / collider

Regularized regression versus the **high-dimensional propensity score** for confounding adjustment in secondary database analyses

JM Franklin, W Eddings, RJ Glynn… - American journal of …, 2015 - academic.oup.com

… We present a simulation study that compares the **high-dimensional propensity score** algorithm for variable selection with approaches that utilize direct adjustment for all potential confounders via regularized regression, including **ridge** regression and lasso regression …

[PDF] Scalable collaborative targeted learning for large scale and high-dimensional data

C Ju, S Gruber, SD Lendle, JM Franklin… - UC Berkeley Division of …, 2016 - core.ac.uk

… **dimensional propensity score** (hdPS) algorithm is a method to extract information from electronic medical claims data that produces hundreds or even thousands of candidate covariates, increasing the dimension of the data dramatically. [16] In order to apply C- **TMLE** to large …

High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions

R Neugebauer, JA Schmittdiel, Z Zhu… - Statistics in …, 2015 - Wiley Online Library

… The **high-dimensional propensity score** (hdPS) algorithm was proposed 1 for automation of confounding adjustment in problems … of confounders 'by hand' is not practical because of the **high** dimensionality of … t. At each time point , expert-selected covariates (listed in **Table** 1) are …

Extensions of hdPS

- Only ~ 30% of the selected hdPS covariates were common.
- Statistical inefficiency

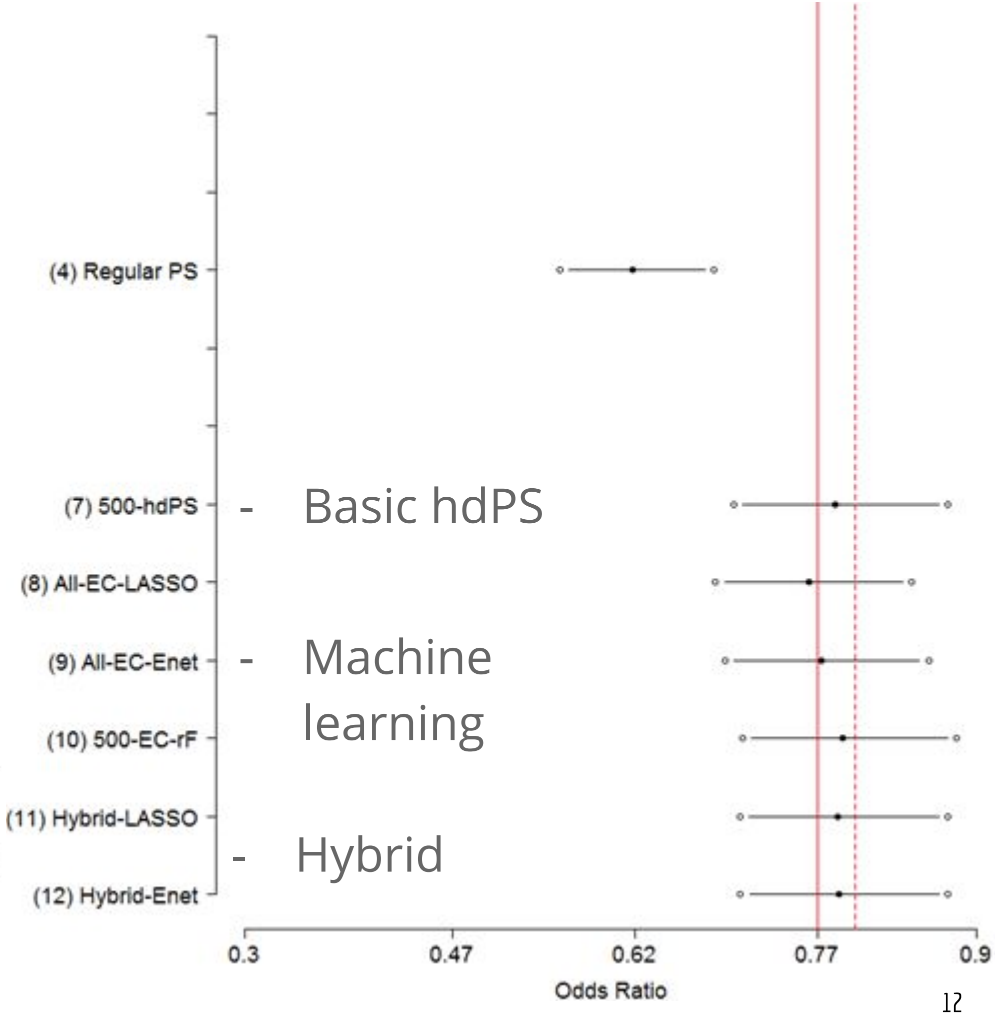
ORIGINAL ARTICLE

Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm?

Mohammad Ehsanul Karim,[a,b] Menglan Pang,[c,d] and Robert W. Platt[c,e,f]

Epidemiology • Volume 29, Number 2, March 2018

ISSN: 1044-3983/18/2902-0191
DOI: 10.1097/EDE.0000000000000787



Inflated SE

Propensity score model overfitting led to inflated variance of estimated odds ratios

T Schuster, WK Lowe, RW Platt - Journal of clinical epidemiology, 2016 - Elsevier

Objective Simulation studies suggest that the ratio of the number of events to the number of estimated parameters in a logistic regression model should be not less than 10 or 20 to 1 to achieve reliable effect estimates. Applications of propensity score approaches for confounding control in practice, however, do often not consider these recommendations. Study Design and Setting We conducted extensive Monte Carlo and plasmode simulation studies to investigate the impact of propensity score model overfitting on the performance in …

- "… overfitting of propensity score models can lead to inflated variance of effect estimates and therefore to estimation inaccuracy in situations where relatively many covariates are included in the propensity score model" (# of exposed vs. # of covariates)
- hdPS context

[HTML] Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children
HK Brown, JG Ray, AS Wilton, Y Lunsky, T Gomes... - Jama, 2017 - jamanetwork.com
Importance Previous observations of a higher risk of child autism spectrum disorder with serotonergic antidepressant exposure during pregnancy may have been confounded. Objective To evaluate the association between serotonergic antidepressant exposure during pregnancy and child autism spectrum disorder. Design, Setting, and Participants Retrospective cohort study. Health administrative data sets were used to study children born to mothers who were receiving public prescription drug coverage during pregnancy in ...

Application of hdPS

JAMA study (2017): **Serotonergic Antidepressant Use** during pregnancy vs. **Autism Spectrum Disorder** in Children

- Unadjusted: HR, 2.16 [95% CI, 1.64-2.86]
- Multivariable adjusted: HR, 1.59 [95% CI, 1.17-2.17]
- IPTW hdPS: HR, 1.61 [95% CI, 0.997-2.59] → "not associated"!!
- 1:1 hdPS matching: HR, 1.64 [95% CI, 1.07-2.53] (sensitivity analysis 1)
- Pre-pregnancy data: HR, 1.85 [95% CI, 1.37-2.51] (sensitivity analysis 2)

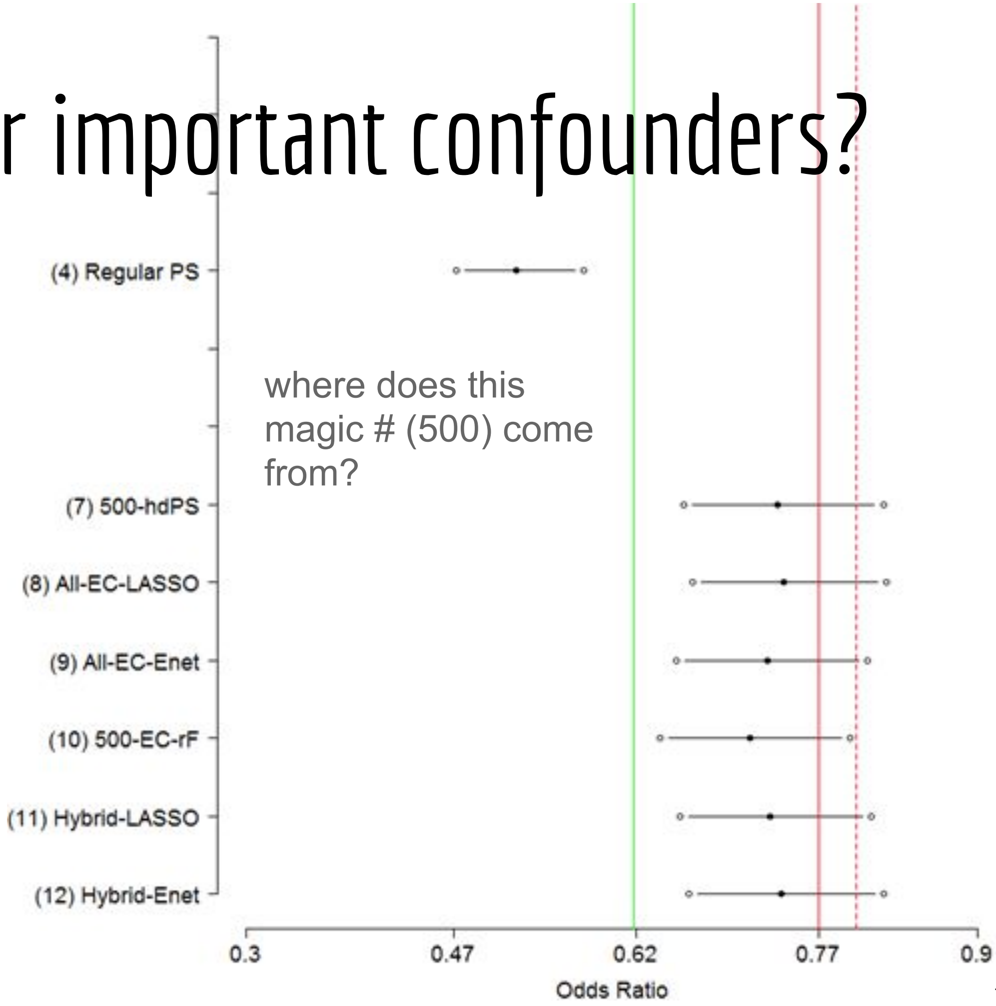
"Adjusting for too many pre-exposure covariates will lead to **collinearity and statistical inefficiency** ...."

[HTML] Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects
S Schneeweiss - Clinical epidemiology, 2018 - ncbi.nlm.nih.gov
Background Decision makers in health care increasingly rely on nonrandomized database analyses to assess the effectiveness, safety, and value of medical products. Health care data scientists use data-adaptive approaches that automatically optimize confounding control to study causal treatment effects. This article summarizes relevant experiences and extensions. Methods The literature was reviewed on the uses of high-dimensional propensity score (HDPS) and related approaches for health care database analyses, including ...

Collective substitute for important confounders?

- use of a hdPS "to adjust for 500 covariates that might **collectively contribute to confounding**"
- "association … may not be causal"
  (JAMA editorial)

- Most simulations based on plasmode

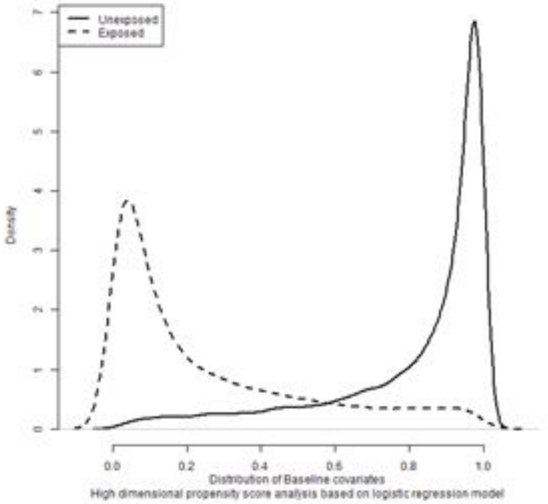


where does this magic # (500) come from?

Current practices and Open questions

Balance diagnostics

- PS analysis not reported, hdPS being main analysis!
- Deviation from PS
  - design vs. analysis stage; selective inference?
- Balance diagnostics in high-dimension
  - balance in p?
- Trimming:
  - practical/near positivity assumption violation
  - target population? bias-variance trade-off



Thank you!

ehsan.karim@ubc.ca