Automated data-adaptive analytics to study causal treatment effects

Sebastian Schneeweiss, MD, ScD
Professor of Medicine and Epidemiology

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

April 9, 2019

This work was funded by multiple grants from the NIH and PCORI

Papers that cover most of this talk:

Schneeweiss S, Rassen JR, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009;20:512–22. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol, 2018:10;771-88 .

Disclosures

- Co-PI, Harvard-Brigham & Women's Hospital Drug Safety Research Center (FDA)
- Co-Chair, Methods Core of the FDA Sentinel System
- Co-Chair, Partners Center for Integrated Healthcare Data Research
- PI of research grants awarded to BWH by Bayer, Vertex, Boehringer Ingelheim
- Consulting fees from WHISCON, LLC, and Aetion (incl. equity)
- Grants/contracts from NIH, PCORI, FDA, IMI, JLAF

Real-World Data

RCT data

Non-interventional data

10% **90%**

Research data Transactional data Data collected PRIMARILY for research Data used SECONDARILY for research

For purpose Other purpose Data specifically for study purpose Data intended for other studies

Other purpose

documentation Administrative Clinical

Digital health devices

- Framingham Study
- Cardiovas Health Study
- Slone Birth Defects Study
- ***Some registries***
- Nurses' Health Study 1
- ***Some registries***
- EHR-based studies
- NDI linkage
- Lab test databases
- ***Some registries***
- Claims data studies
- Geocoding/census
- Fit Bit devices
- Glucose monitors
- PRO apps

(Some records arrive with admin delays)

Dynamic database that records an ongoing stream of new healthcare records in ***Calendar Time*** for all enrolled patients:

1/1/2014 1/1/2015 1/1/2016 1/1/2017

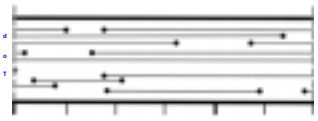
3 *Franklin, Schneeweiss. CPT 2017*

From transactional data to study implementation

Database Studies

Healthcare records are entered as they arrive, sorted by service date.

Stabilized data snapshot [redacted] for research purposes*



1/1/2015 1/1/2016

[redacted]

Individual-patient data has arrived in batches and from various sources

Dx Rx Rx Lab Dx Rx

Study rules are applied and arranged by ***Event Time***

From Real-World Data to Causal Treatment Effects

Achieve causal *REDUCTION*

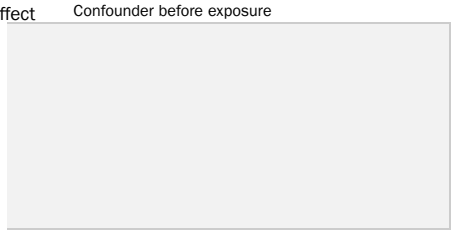
Design *DESIGN CHOICE BIAS*

study design 1) Controlled 2) self-controlled 3) scanning window • New users, active comparators A C Y

pop[n] • Patient-informed outcomes • • Causal temporality Exposure before outcome

Layer

• Medically-informed target rmed effect Confounder before exposure



Labs, stage

Physician notes
Dx, Px, Rx codes
Monitors
Physician notes
Behavior, econ, social

Measures La
Achieve fit-for purpose measurement

ch Patient

Dx codes Rx codes (DDI) Genomics, proteomics

• sensitivity
• specificity,
• completeness • mean sqr diff

NLP, bioinformatics *TREATMENT OUTCOME*

CONFOUNDERS TARGET POP[N]

BALANCE ROBUSTNESS

Achieve causal
Longitudinal record

processing, Machine learning,

Regression, PS analysis
Proxy adjustment: HDPS,

• Neg./pos. control endpoints •

Analytics Layer
analysis

Considering: • Confounders •

• Sensitivity analyses of design • Quantitative bias analysis

Balance in unmeasured confounders

Achieve balance:

• Follow-up model quality

• Measurement c-stat

With ML-informed

• Multiple comparisons
covariate prioritization

• Check balance: SD, residuals,

Causal study design (target trial)

Cohort Entry Date
(New use of drug A or B)

Follow-up Window Days [1, 365]
Days ([0, 365]

Time

Exclusion Assessment Window (Age ≤ 65) Days [0, 0]

Follow up Window

  Schneeweiss S et al. Ann Intern Med 2019

A = Exposure; e.g. start of a new drug Y = Outcome of interest
C = observable confounder

Confounding

C A

Y (Outcome)

Confounders are independent risk factors for the outcome and associates of treatment choice:
Comorbidity, severity, co-medication, etc. = prognosis

Confounding

Outcome of interest
C = observable confounder **(serves as proxy) U = unobserved confounder**

A

Y (Outcome)

A = Exposure; e.g. start of a new drug Y =

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 exams	ICD-9, CPT-4, # PCP visits
Fairly healthy senior	Receiving the first lipid-lowering medication at age 70	NDC, ATC, Read
Chronically sick	Regular visits with specialist, hospitalization; many prescription drugs	Outcome surveillance # specialist visits, NDC, ATC
intensity	General markers for healthcare utilization intensity	# visits, # different drugs

We now have a causal study design with clear temporality: -
Exposure before outcome
- Confounders before exposure
- No causal intermediates, no reverse causation

We have measured confounders and proxies of unmeasured confounders

Wouldn't it make sense to include many proxies as covariates in our analysis? We hope that they collectively will adjust for unmeasured confounding.

Do we need to know the specific meaning of each proxy variable?
The presence of a hosp. d/c dx of ICD-9 410 can be a new variable

Data-adaptive Causal Analyses (1)

Data dimension

Unstructured text notes

Temporality attribute

Intensity attribute

Inpatient Diagnoses *
Outpatient Diagnoses *
Inpatient Procedures **
Outpatient Procedures ** Drug dispensing ***
Lab test results****

Coding examples: *ICD: International classification of disease; **CPT: Current procedure terminology, Read codes; ***NDC: National Drug Code, ATC: Anatomical Therapeutic Classification; **** LOINC codes: Logical Observation Identifiers Names and Codes

Analyses (2) Intensity

Cohort entry date

t

Data-adaptive Causal

Temporality pattern

Sporadic

Frequen

Even

Distal

COV FUP

Proximal

Data-adaptive Causal Analyses (3)

1) Include variables that are related to the exposure and the outcome 2) Include variables that are unrelated to the exposure but related to the outcome

⇒ "Bias" covariate prioritization:
- Compute potential bias (Bross)
- Rank by decreasing potential bias

  Brookhart et al. AJE 2006; Bross JD. J Chronic Dis 1966

Data-adaptive Causal Analyses (4)

NLP/ imputatio n

2

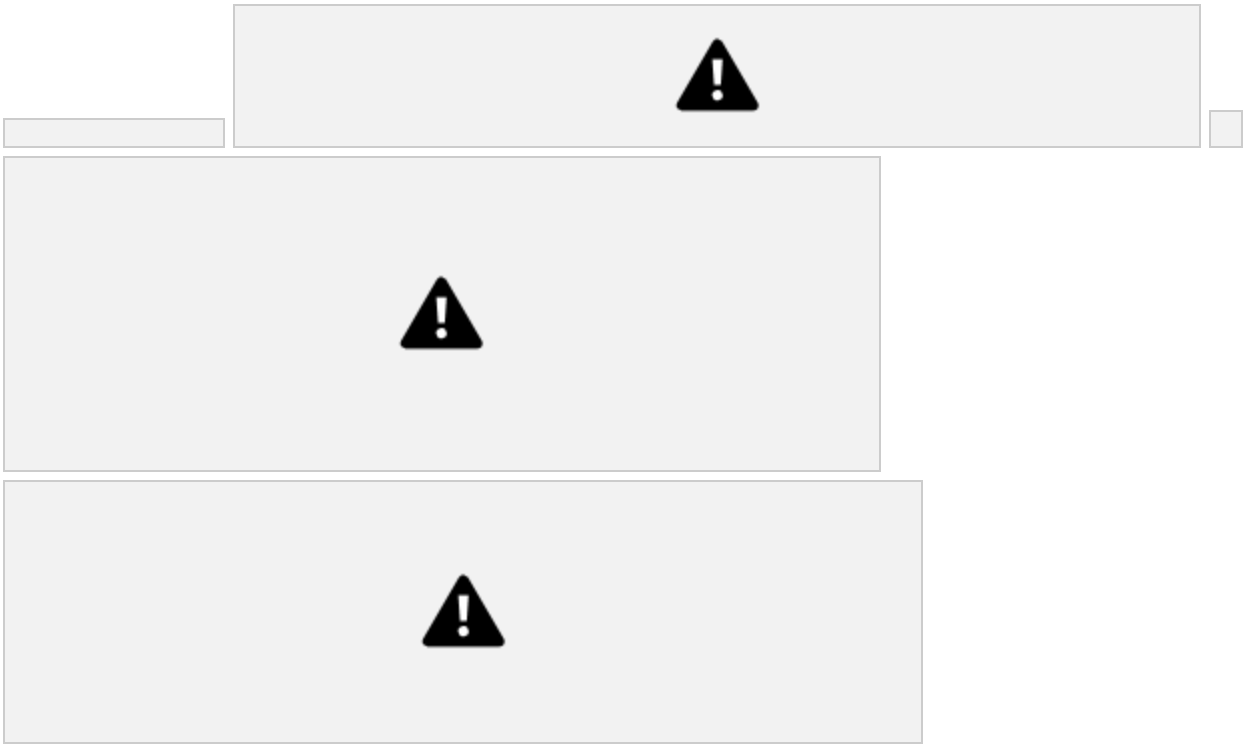
Prevalence filter

Frequency,

temporal patterns

Interactions

3



SSRI vs.

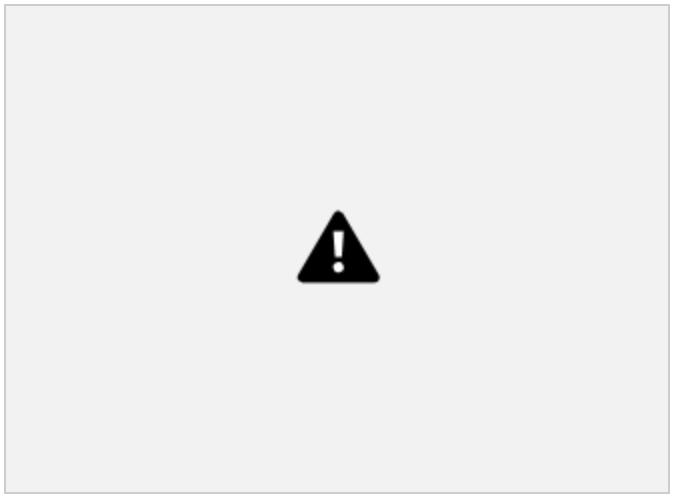
Tricyclic appears falsely protective because of their contraindication in acutely suicidal patients

Data-adaptive Causal Analyses: Performance (1)

Data sources

Claims databases:

U.S. Medicare U.S. commercial Canada Germany



ASYR = Age, sex, year, race

Area A Area B Area C

(1) Schneeweiss et al. High-dimensional propensity score adjustment in studies of treatment effects using healthcare claims data. Epidemiol 2009;20:512–22.

(2) Garbe et al. High-dimensional vs conventional PS in a comp. effectiveness study of coxibs and reduced upper GI complications. Eur J Clin Pharmac2013;69:549-57.
(3) Le et al. Effects of aggregation of drug and diagnostic codes on the performance of the hdPS algorithm. BMC Med Res Methodology 2013;13:142.
(4) Hallas, et al. Performance of the High-dimensional Propensity Score in a Nordic Healthcare Model. Basic Clin Pharmacol Toxicol 2017;120:312-17.

(5) Toh et al. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm. Pharmacoepidemiol Drug Safety. 2011;20:849-57.
(6,7) Rassen et al. Cardiovascular outcomes and mortality in patients using clopidogrel with PPIs after percutaneous coronary intervention. Circ 2009;120:2322-9.

Data-adaptive Causal Analyses: Performance (2)

Data sources

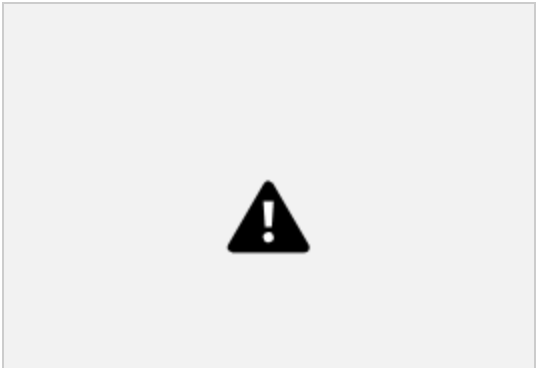
ASYR = Age, sex, year, race

Area A Area B Area C

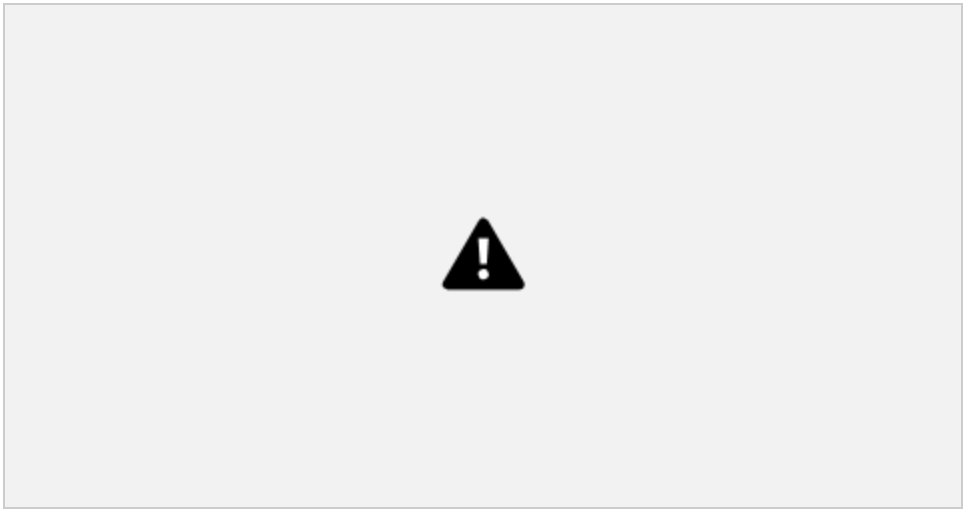
(8) Schneeweiss et al. High-dimensional propensity score adjustment in studies of treatment effects using healthcare claims data. Epidemiol 2009;20:512–22.
(9) Schneeweiss, et al. The comparative safety of antidepressant agents in children regarding suicidal acts. Pediatrics 2010;125: 876–88.
(10) Schneeweiss et al. Variation in the risk of suicide attempts and completed suicides by antidepressant agent in adults. Arch Gen Psychiatry 2010;67:497-506
(11) Zhou et al. Sentinel Modular Program for Propensity Score–Matched Cohort Analyses. Epidemiology 2017;28:838-46.

  (12) Patorno et al. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. JAMA 2010;303:1401-9

Data-adaptive Causal Analyses: Performance (3)



17



Data-adaptive Causal Analyses with free-text notes

1 Word:

leukocytosi
oxycontin
haptic
extracrani
scleral
splenomengali
valium

2 Words:

site cervix
categori within
specimen categori
peripher edema
maxillari sinus
differenti diagnos

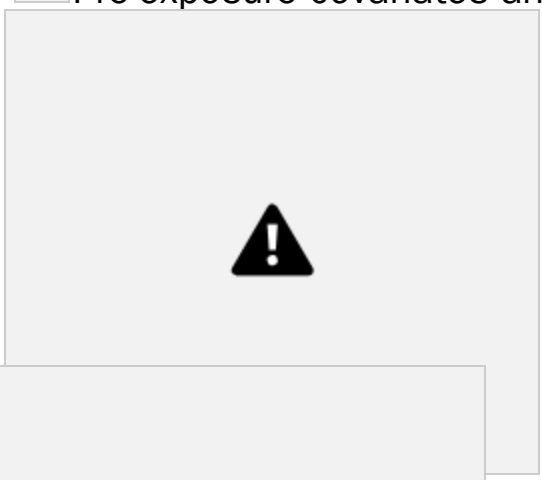
3 Words:

specimen site cervix
site cervix endocervix
categori within normal
impress ct abdomen
or 3 view
white female a
exam ct abdomen

Rassen et al. 2013

Pre-exposure covariates and the risk of increasing bias*

Instrumental variable or Z-bias

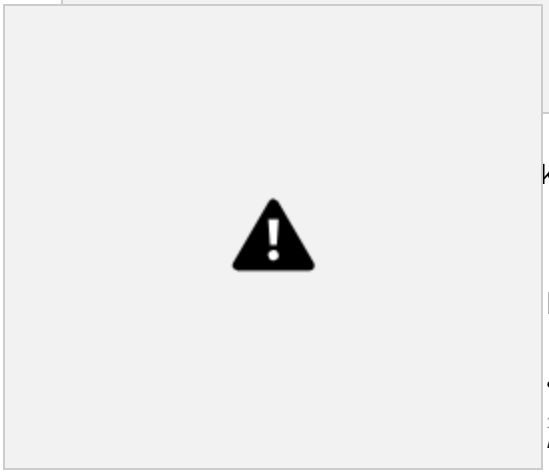


Trt

Outcome

Z-bias is a bit more likely:

Conditioning on treatment will open a back-door path and an IV-like variable will amplify residual confounding by U



k

Trt Outcome

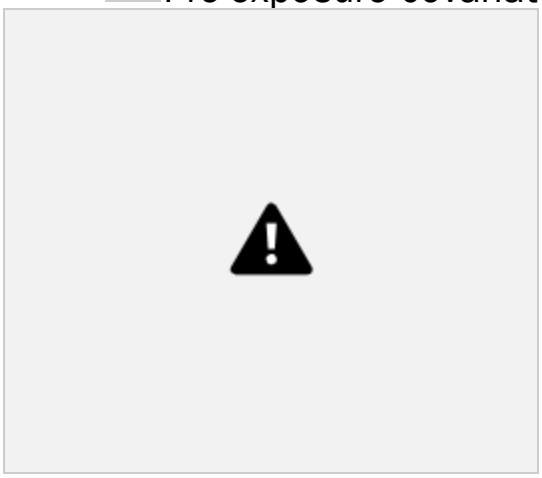
In practice we do not know whether variables are

- M or Z-bias candidates that should not be adjusted for or
- True confounders that should be adjusted for

  ** e.g. VanderWeele Eur J Epidemiol 2019*

Pre-exposure covariates and the risk of increasing bias*

Instrumental variable or Z-bias



IV **U**

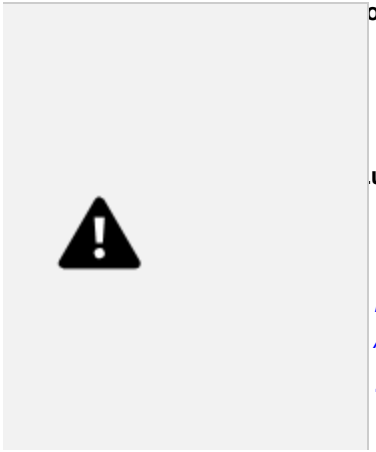
collider bias

Trt

Outcome

Liu WM, et al. AJE 2012:

In most realistic settings, it is really hard to construct meaningful



oke (U) RR=1.6

RR=15 **RR=27**

ung Cancer (Y)

Trt Outcome

Myers JA et al. AJE 2011:

…when in doubt

whether a pre-exposure variable is an IV or a confounder, then adjust for that covariate

AJE 2012;176:938-48.
Myers JA et al. Effects of adjusting for instrumental variables on bias and prediction of effect estimates. AJE 2011;174:1213-22.

Liu WM, et al. Implications of M bias in epidemiologic studies: A simulation study.

screening tools

for IVs

Arthritis

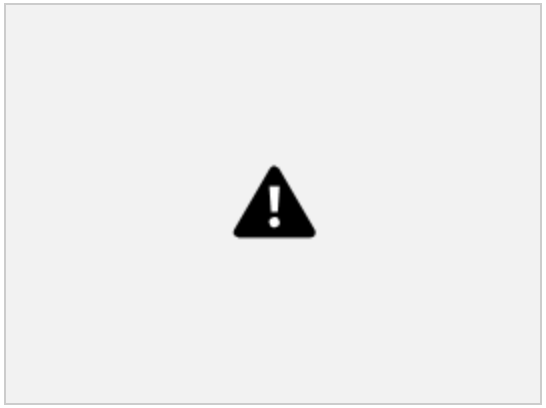
Empirical

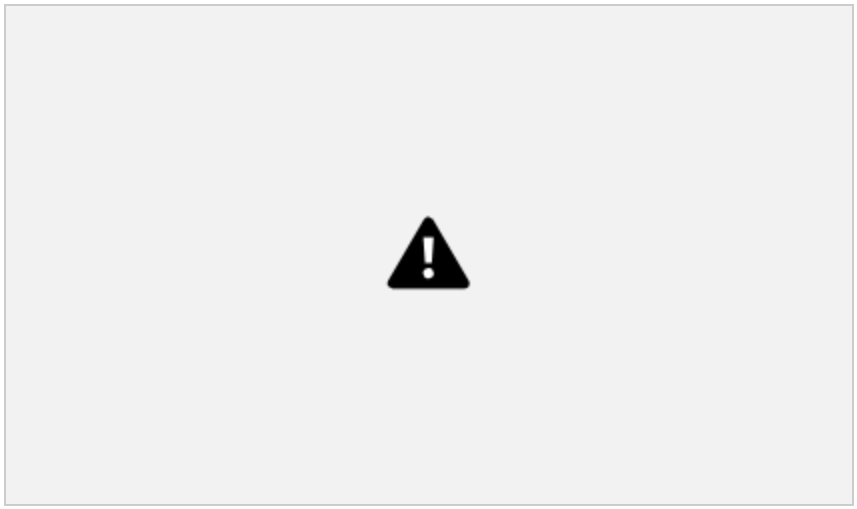
Hybrid models: HDPS augmented with machine learning

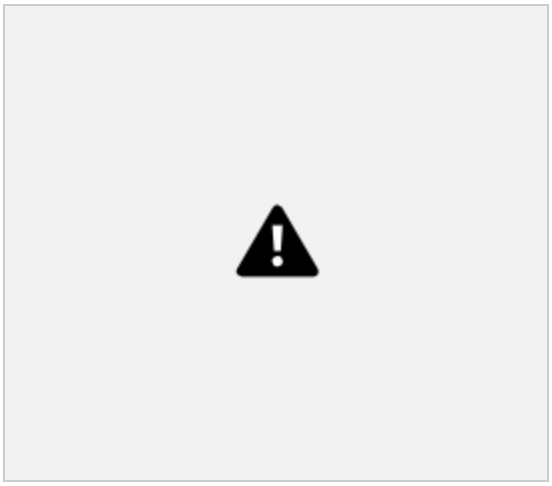
Our weak spot:

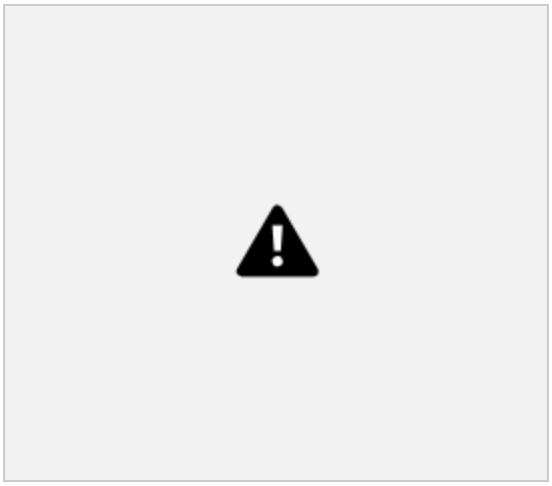
Automatically identifying independent risk factors of the outcome

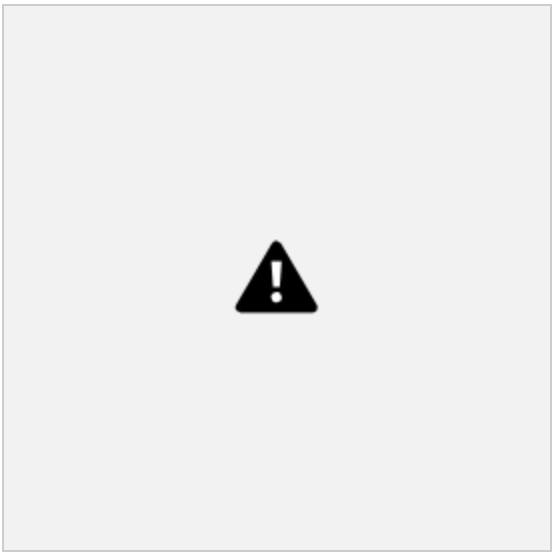
Even harder when outcomes are infrequent!











So, have we found the holy crail of database epidemiology?

Of course not, but …

- Have a generalizable framework to optimized confounding adjustment in a given healthcare database
- The importance of a causal study designs cannot be underestimated
- Automated confounding adjustment may reduce human biases in confounding adjustment

Some more references

Schneeweiss S, Rassen JR, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009;20:512–22.

Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol 2011;173:1404-13.

Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacoepidemiol Drug Safety 2012;21 S1:41-9. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis.* Apr 2014;72:219-226.

Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses. *Am J Epidemiol.* Oct 1 2015;182(7):651-659.

Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen RA, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analysing healthcare databases. Epidemiology 2017;28:237–48.

Wyss R, Schneeweiss S, Eddings W, van der Laan M, Lendle SD, Ju C, Franklin JM. Optimizing high-dimensional propensity score estimation through super learner prediction modeling. Epidemiology 2019;29:96-106.

Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol, 2018:10;771-88.

