Reliability in Observational Research: Assessing Covariate Imbalance in Small Studies

George Hripcsak, MD, MS

Biomedical Informatics, Columbia University

COLUMBIA | COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

CIO Disclosure Information
George Hripcsak

- Leadership position: OHDSI Coordinating Center Director
- Research funding from: NIH, FDA

I have no other financial relationships to disclose.

Observational research

- Subjects **observed** in their natural settings
  - **Real-world evidence**
    - Often using data collected for other purposes
  - Administrative claims data
    - Merative CCAE (10M's)
  - Electronic health record (EHR) data
    - Columbia clinical data warehouse (6M)
  - Other sources
    - Census, social media, mobile sensors, imaging
- Versus experimental
  - Randomized clinical trials (RCTs)

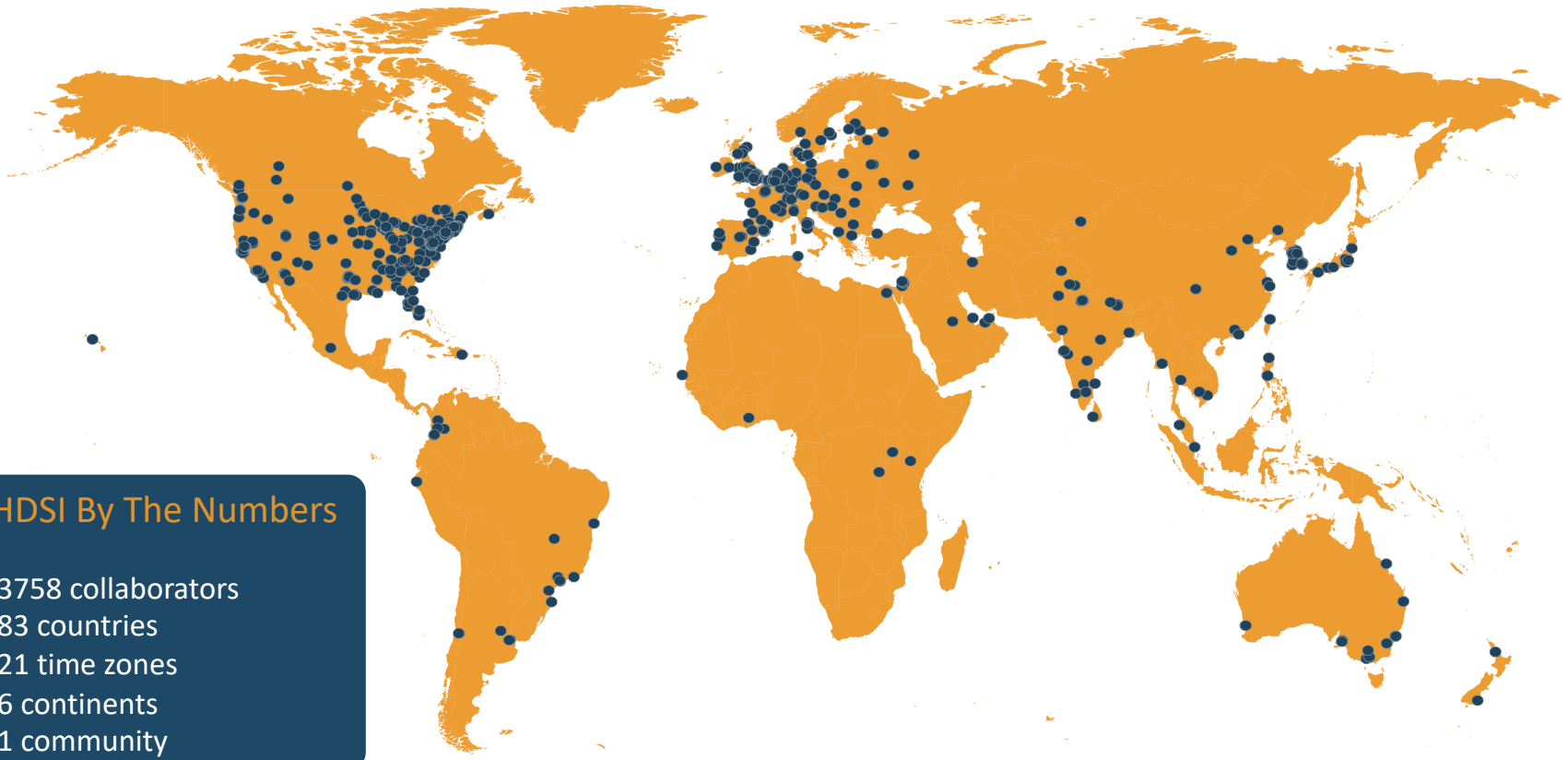
Observational Health Data Sciences and Informatics (OHDSI, as “Odyssey”)

Mission: To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care

A multi-stakeholder, interdisciplinary, international collaborative with a coordinating center at Columbia University

http://ohdsi.org

OHDSI

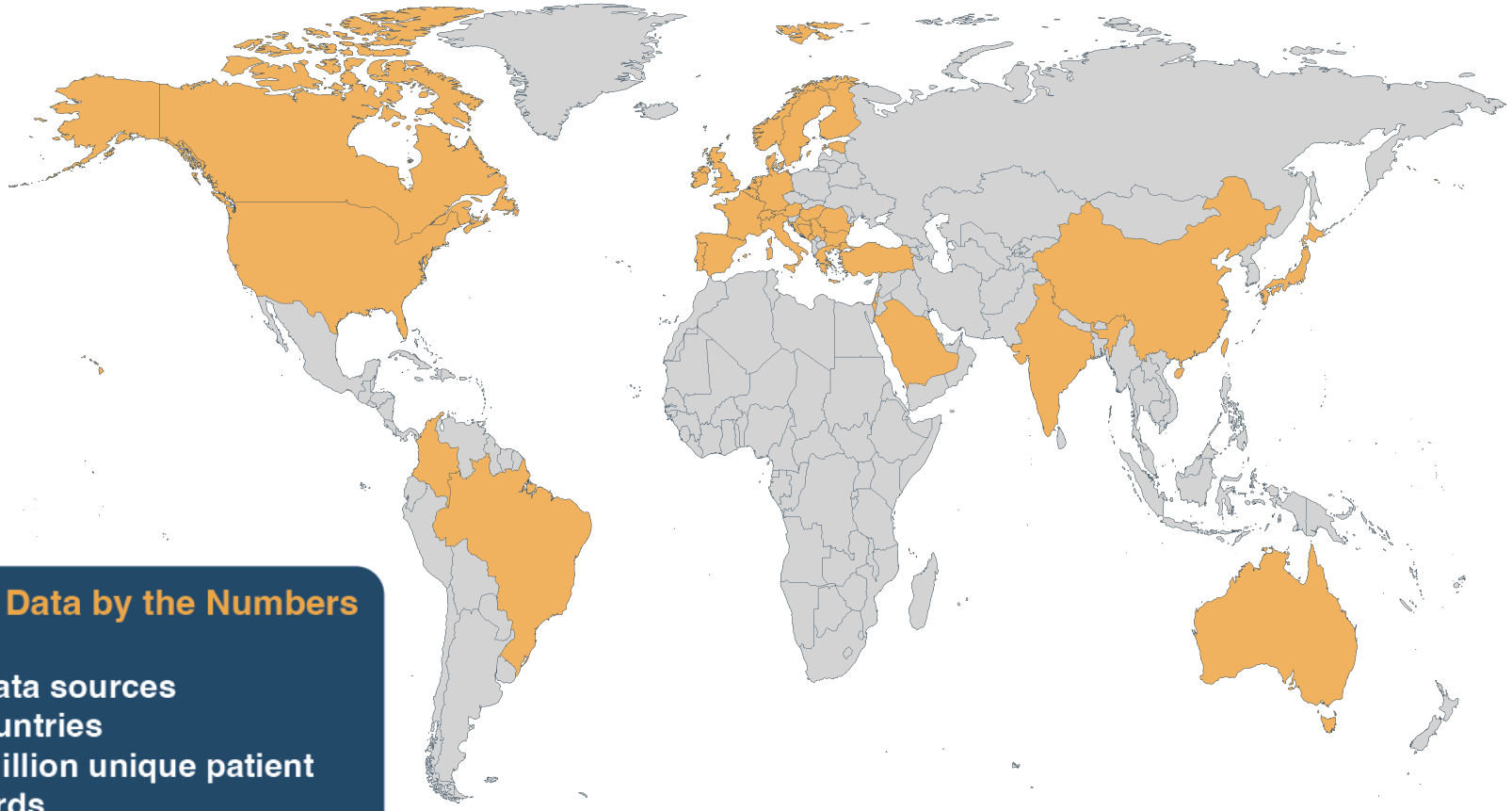


OHDSI By The Numbers

- 3758 collaborators
- 83 countries
- 21 time zones
- 6 continents
- 1 community

- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- >600 papers, specific influence on EMA and FDA for COVID-19

OHDSI data partners



OMOP Data by the Numbers

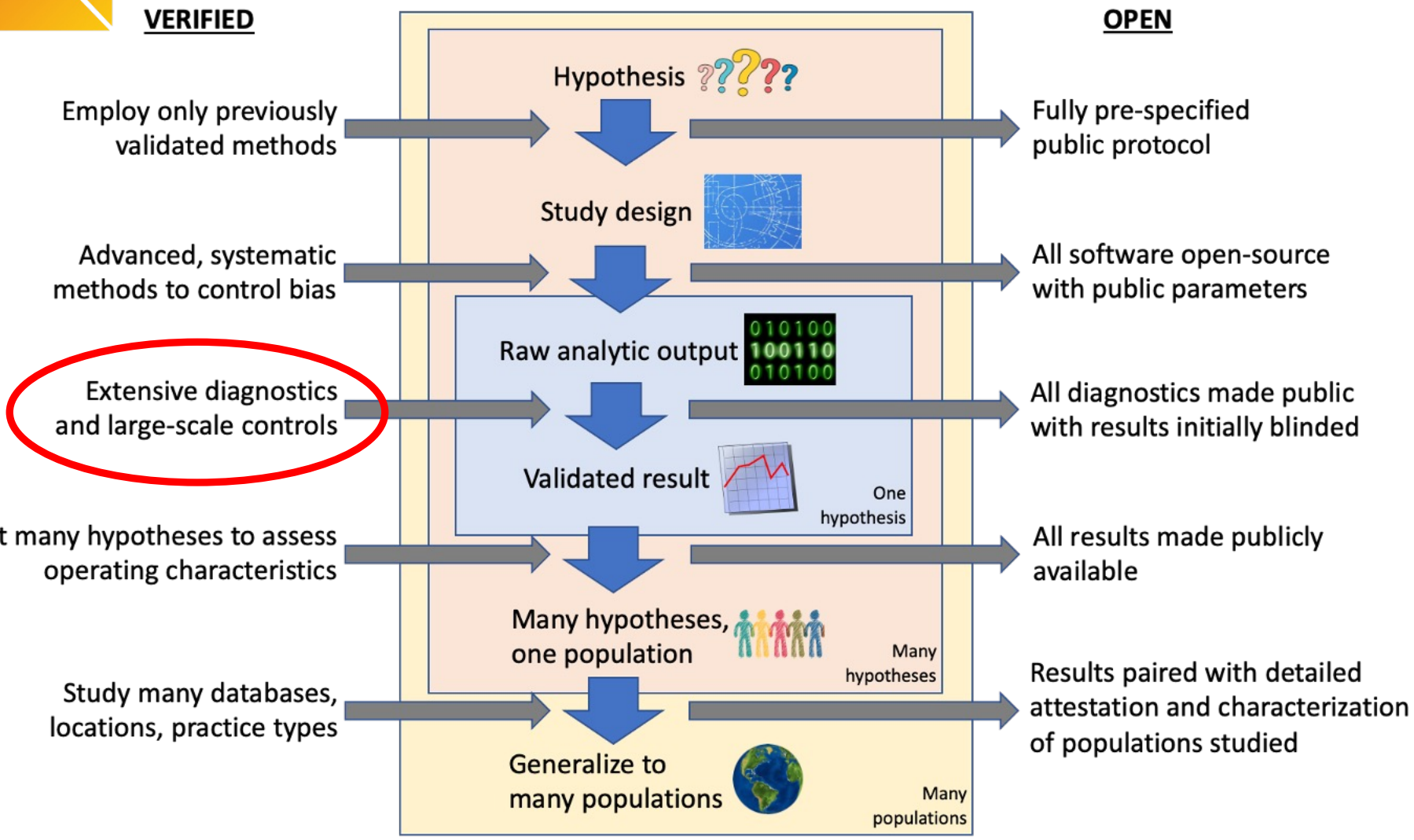
- 534 data sources
- 49 countries
- 956 million unique patient records
- approximately 12% of the world's population

OHDSI's 10 LEGEND Principles for generating **reliable evidence**

- LEGEND will generate evidence at a large scale
- Dissemination of the evidence will not depend on the estimated effects
- LEGEND will generate evidence using a prespecified analysis design
- LEGEND will generate evidence by consistently applying a systematic process across all research questions
  - No thumb on the scale
- LEGEND will generate evidence using best practices
- LEGEND will include empirical evaluation through the use of control questions
- LEGEND will generate evidence using open-source software that is freely available to all
- LEGEND will not be used to evaluate new methods
- LEGEND will generate evidence across a network of multiple databases
- LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network

Schuemie JAMIA 2020

Verified and open



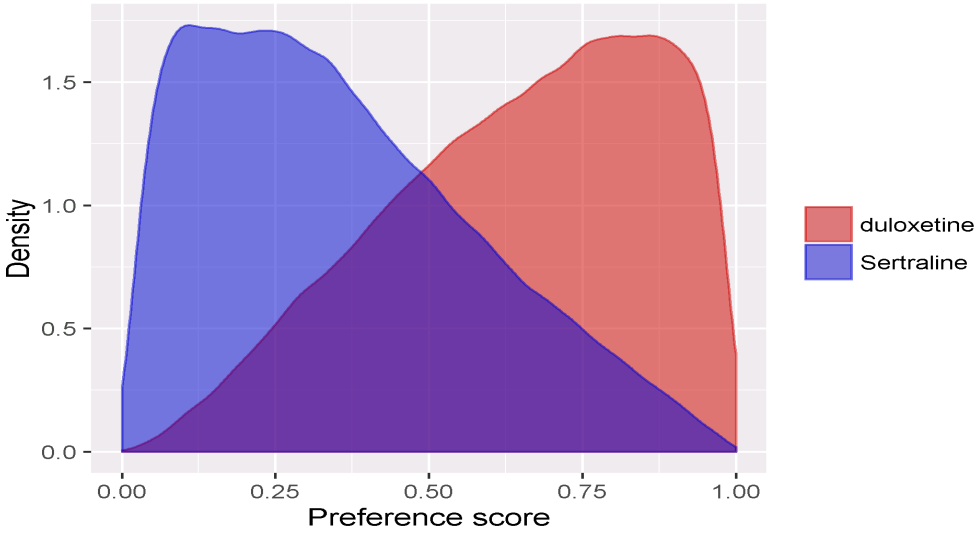
Motivation for the study

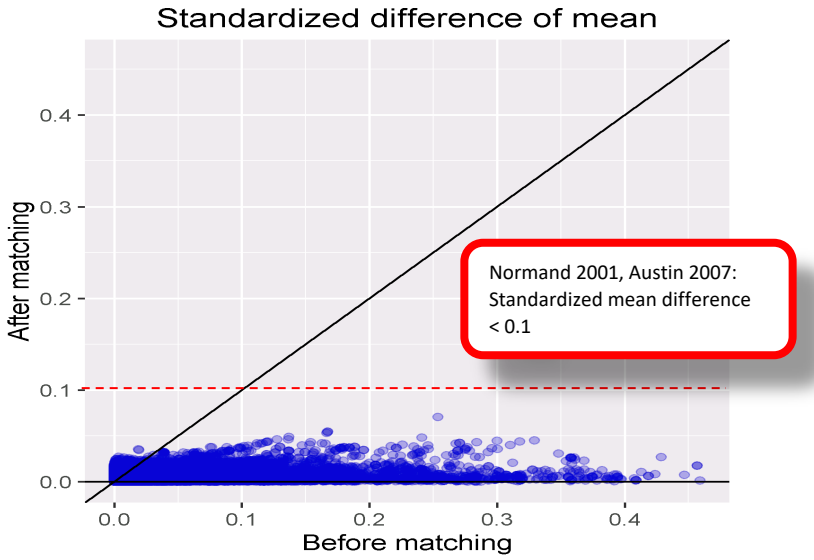
- Was working on confounding bias in comparative cohort studies
- Was using many covariates
- Diagnostics were failing for smaller studies
- Carried out this study
- Found the results apply even for handful of covariates

Addressing reproducibility

Propensity score adjustment with large-scale covariate set: measured confounding (and some unmeasured?)

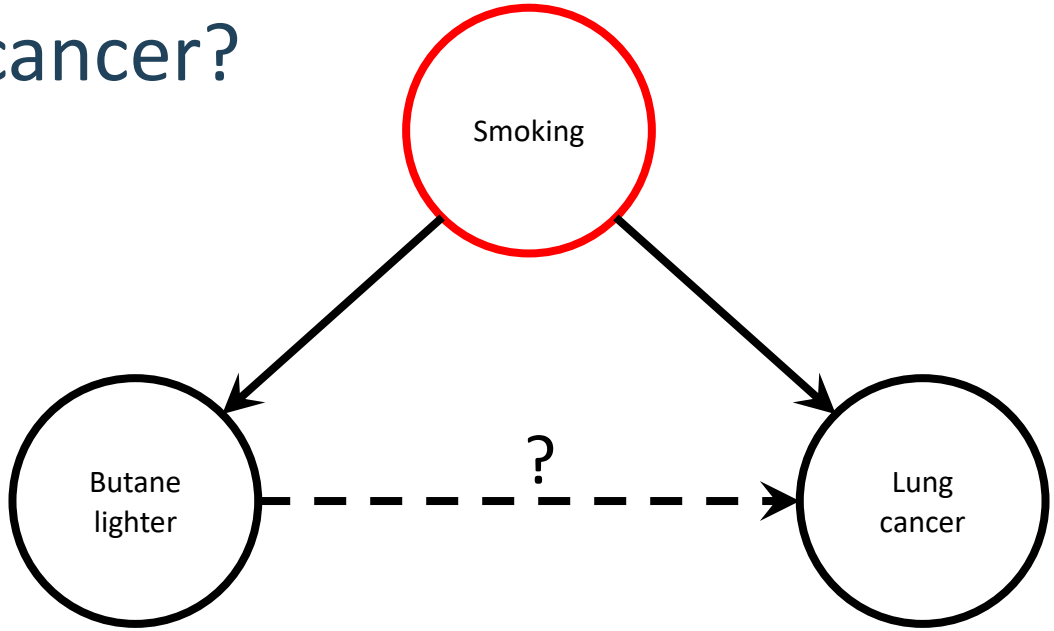
- Take advantage of the huge databases and balance on tens of thousands of covariates, pulling in other variables (BP)
- Mimic balance of randomization (imperfect)
- Don't rely on human expertise to select confounders: **systematic**
- Diagnostics





Confounding

- Does butane gas cause lung cancer?



Propensity score for confounding in comparative cohort studies

- Propensity score = patient's probability of belonging to the target cohort vs. the comparator cohort, given the baseline covariates
- Propensity score can be used as a 'balancing score'
  - if the two cohorts have similar propensity score distribution, then the distribution of covariates should be the similar
- Balance the propensity -> balance the covariates
- Balance the covariates -> the comparisons are similar
  - Make a causal assertion: must be due to the treatment

Rubin Biometrika 1983

How to select the confounders

- Manual selection -> poor agreement
  - Chien 2015: age, month, gender, #visits, income urbanization, #drugs, specific drugs, Charlson, comorbidities (16), +HDPS variables
  - Hicks 2018: age, sex, year of cohort entry, body mass index, smoking status, alcohol related disorders (including alcoholism, alcoholic cirrhosis of the liver, alcoholic hepatitis, and hepatic failure), and history of lung diseases (including pneumonia, tuberculosis, and chronic obstructive pulmonary disease), duration of HTN Rx, statin use, #drugs
  - Ku 2018: age, sex, race, income status, baseline HF, baseline myocardial infarction, baseline peripheral artery disease, baseline stroke, baseline eGFR, baseline proteinuria, and time-dependent covariates including diabetes mellitus, obesity, systolic blood pressure, statin use, aspirin use, diuretic use, and concurrent use of other antihypertensive agents for the outcome of HF
  - Magid 2010: age, gender, days on thiazide prior to 2nd agent start, # of visits prior to thiazide, Mean Systolic BP, Mean Diastolic BP, Chronic Obstructive Pulmonary Disease, Hyperlipidemia, Cancer, Dementia, Chronic liver disease, Depression
  - Hasvold 2014: age, gender, elevated blood glucose, overweight and low socio-economic status are known risk factors for diabetes, High cholesterol and hypertension are additionally known risk factors for CVD
- Empirical selection

Large-scale propensity score (LSPS)

- A **systematic** approach to propensity adjustment
- Use a large set of covariates (10,000 < n < 100,000)
- But don't want to balance *everything*
  - Mediators – pre-treatment
  - Simple colliders – pre-treatment
  - Instruments – diagnostics, domain knowledge
  - M-bias – correlation with underlying causes
- Fit a propensity model
  - LASSO (regularized regression) because #variables > #cases
- Match or stratify on propensity score
- Diagnostic: check that covariate balance is achieved on all observed variables

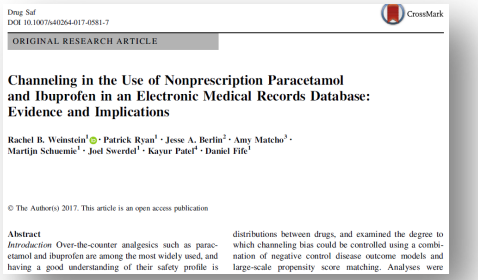
Zhang JBI 2022
Tian Int J Epi 2018

Notes about LSPS

- Goal is to adjust for all pre-treatment variables
  - Usually try to select confounders; **not** here
  - Differs from HDPS …, which attempt to select confounders
- More variables than cases, so use LASSO
  - Merely doing dimension reduction by picking the informative variables, but attempting to convey **all** the information
- When adjust for many variables, may adjust for ones that are not directly measured
    - Baseline blood pressure gets pulled in for HTN studies
    - Related to Wang and Blei Deconfounder
    - Ok on colliders, mediators, instruments, M-bias

LSPS

- Reduce bias if balance on many covariates instead of a few human-selected covariates (bias measured via negative controls)



Drug Saf
DOI 10.1007/s40264-017-0581-7

ORIGINAL RESEARCH ARTICLE

Channeling in the Use of Nonprescription Paracetamol and Ibuprofen in an Electronic Medical Records Database: Evidence and Implications

Rachel B. Weinstein[1] · Patrick Ryan[1] · Jesse A. Berlin[2] · Amy Matcho[3] · Martijn Schuemie[1] · Joel Swerdel[1] · Kayur Patel[4] · Daniel Fife[1]

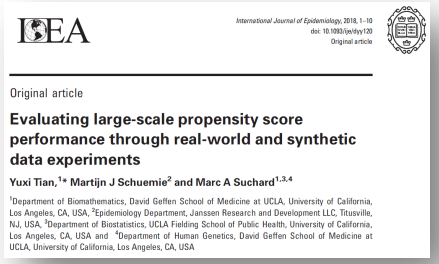
© The Author(s) 2017. This article is an open access publication

Abstract

Introduction Over-the-counter analgesics such as paracetamol and ibuprofen are among the most widely used, and having a good understanding of their safety profile is

distributions between drugs, and examined the degree to which channeling bias could be controlled using a combination of negative control disease outcome models and large-scale propensity score matching. Analyses were

- LSPS performs better than confounder selection like high-dimensional propensity score (HDPS)



IJEA

International Journal of Epidemiology, 2018, 1–10
doi: 10.1093/ije/dyy120
Original article

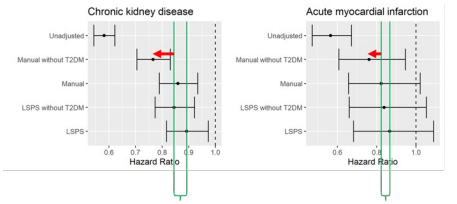
Original article

Evaluating large-scale propensity score performance through real-world and synthetic data experiments

Yuxi Tian,[1]* Martijn J Schuemie[2] and Marc A Suchard[1,3,4]

[1]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, [2]Epidemiology Department, Janssen Research and Development LLC, Titusville, NJ, USA, [3]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA and [4]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

- LSPS does better than manual selection if a confounder is missed

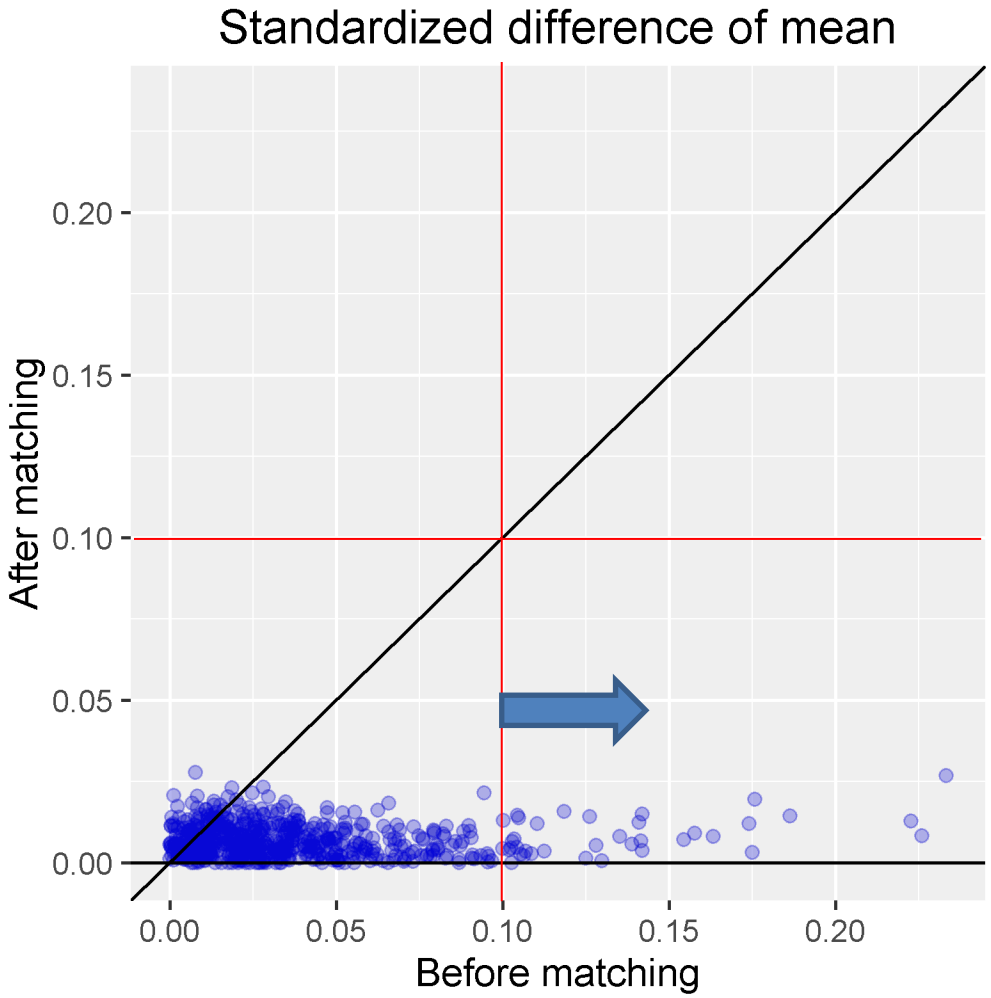


Zhang JBI 2022

How do you know you succeeded?

- Whether you balance 5 or 50,000 covariates that are potential confounders, how do you know it worked?

Diagnostic: Covariate balance



Plot 60,000 covariates;
most are binary:

$$\frac{\text{abs}(P_{\text{target group}} - P_{\text{comparator group}})}{\text{standard deviation}}$$

Normand 2001, Austin 2007:
Standardized mean difference
< 0.1

Problem for today

- As sample size falls, you always fail your diagnostics with chance imbalance
  - What to do different?

Covariate balance review

- Covariate balance is an important diagnostic for PS adjustment in cohort studies (1/3rd) [Granger 2020]
- The goal is not to detect imbalance, but to detect substantial imbalance [Austin 2009, …]
  - Else as sample size rises and therefore precision of SMD rises, all studies will be rejected
- The most common solution is to check for |SMD| over 0.1 (or 0.25) [Austin 2009, …]

Hripcsak medRxiv 2024

Reject small cohorts for chance imbalance

- Imbalance by chance

$$P(false\ rejection) = 1 - \left(2\Phi\left(\frac{\sqrt{N}}{20}\right) - 1\right)^{J}$$

  - Total sample of 250 and 5 covariates, 90% chance of rejecting study as imbalanced (SMD>0.1)
  - Total sample of 1000 and 20 covariates, 90%
  - As covariates increase, more chance rejection

Hripcsak medRxiv 2024

Idea

- Check not for nominally exceeding a threshold, but for statistically significantly exceeding the threshold
  - As sample size falls, the threshold allows more imbalance but the corresponding wider effect CI tolerates more bias
    - Confounding could shift effect estimate 1.2 to 1.4 but CI is 0.7 to 3
    - The CI is designed to accommodate chance imbalance, so no reason to reject studies with chance imbalance
- Try this new rule in simulation and RWD

Standardized mean difference (SMD)

- $sd_j = \sqrt{\frac{\left(\frac{s_{1,j}}{n_1}\right)\left(\frac{1-s_{1,j}}{n_1}\right)+\left(\frac{s_{0,j}}{n_0}\right)\left(\frac{1-s_{0,j}}{n_0}\right)}{2}}$
- $smd_j = \frac{\frac{s_{1,j}}{n_1} - \frac{s_{0,j}}{n_0}}{sd_j}$
- $varsmd_j = \frac{n_1+n_0}{n_1 n_0} + \frac{smd_j^2}{2(n_1+n_0-2)}$

Hripcsak medRxiv 2024

Three primary rules

- **All** – accept all studies (ignore imbalance)
  - Imbalance commonly ignored
- **Nominal** – reject studies with any covariate |SMD| is greater than 0.1
  - Most common threshold when one is used
- **Signif** – reject studies with any covariate |SMD| statistically significantly greater than 0.1 after Bonferroni correction for #covariates
  - Our proposal

Hripcsak medRxiv 2024

Base case simulation conditions

- Sample size combined cohorts 250 to 4000 per site = database (also 20,000)
  - network has 80 to 5 databases, keeping total sample size constant over network
- 1 treatment
  - binary with base positive rate 0.5
- 1 outcome
  - varied true effect size from 0
  - base prevalence 0.25 (0.01)
- 1000 covariates (also 20 and 100,000)
  - one confounder with varied strength from 0 (also distributed confounders)
  - nine outcome predictors
  - rest independent
  - covariates binary with base positive rate 0.5 (0.1)

Hripcsak medRxiv 2024

Simulation base case

- $x_{i,j} \sim Bernoulli(0.5)$
- $t_i \sim Bernoulli\left(0.5 + c_t\left(1 - 2x_{i,1}\right)\right)$
- $y_i \sim Bernoulli\left(0.25 + c_e(1 - 2t_i) + c_y\left(1 - 2x_{i,1}\right) + c_x\left(1 - 2x_{i,2}\right) + \cdots + c_x\left(1 - 2x_{i,10}\right)\right)$

Hripcsak medRxiv 2024

Simulation analysis

- Logistic regression (R glm) of treatment on outcome
- Purposely not include covariates, because the goal is to see the effect of different degrees of residual confounding
  - E.g., as if PS adjustment had been done and may not have been fully effective
  - We should see imbalance and possibly an effect estimate biased by confounding
- Run analysis on each database, and then do meta-analysis on effect estimate (R rma)

Three rules, two levels

- Rules
  - **All** – accept all studies (ignore imbalance)
  - **Nominal** – reject studies any |SMD| > 0.1
  - **Signif** – reject studies any |SMD| statistically significantly > 0.1 after Bonferroni
- Levels
  - Database
    - Apply rule to each covariate, reject some databases
  - Network
    - Random effects model (R rma) on the SMDs for each covariate across non-rejected databases
    - Apply the rule to the meta-analytic estimates, potentially reject whole network study

Hripcsak medRxiv 2024

Rules can be applied at the database or network level

- <network> on <database>
  - All-On-All
  - All-On-Nominal
  - All-On-Signif
  - Nominal-On-All
  - Nominal-On-Nominal
  - Nominal-On-Signif
  - Signif-On-All
  - Signif-On-Nominal
  - Signif-On-Signif

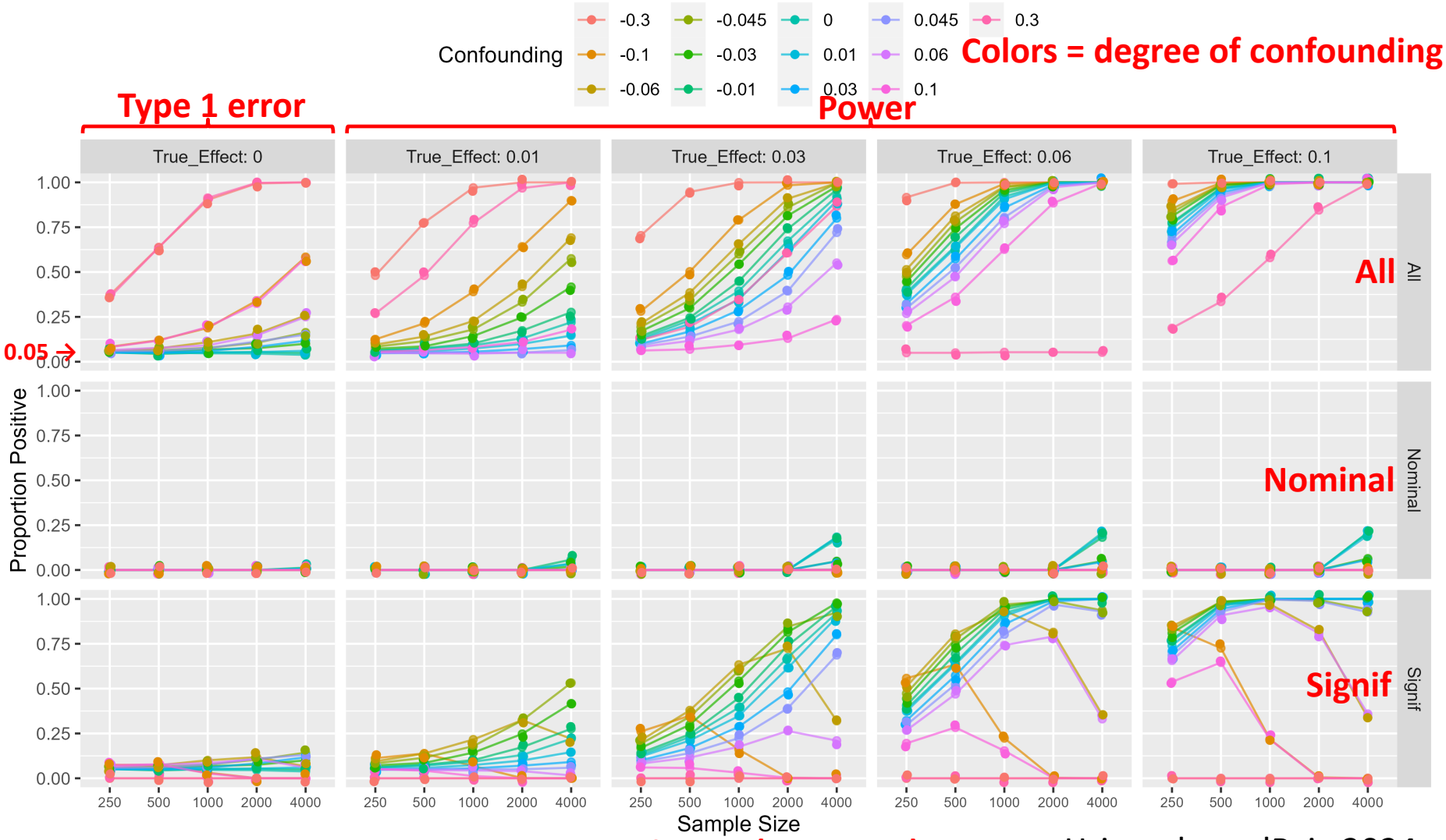
Hripcsak medRxiv 2024

Metrics

- Type 1 error rate
  - Among studies with no true effect
  - Numerator – # not rejected and effect $p<0.05$
  - Denominator – total number of studies
- Power
  - Among studies with a true effect
  - Numerator – # not rejected and effect $p<0.05$
  - Denominator – total number of studies

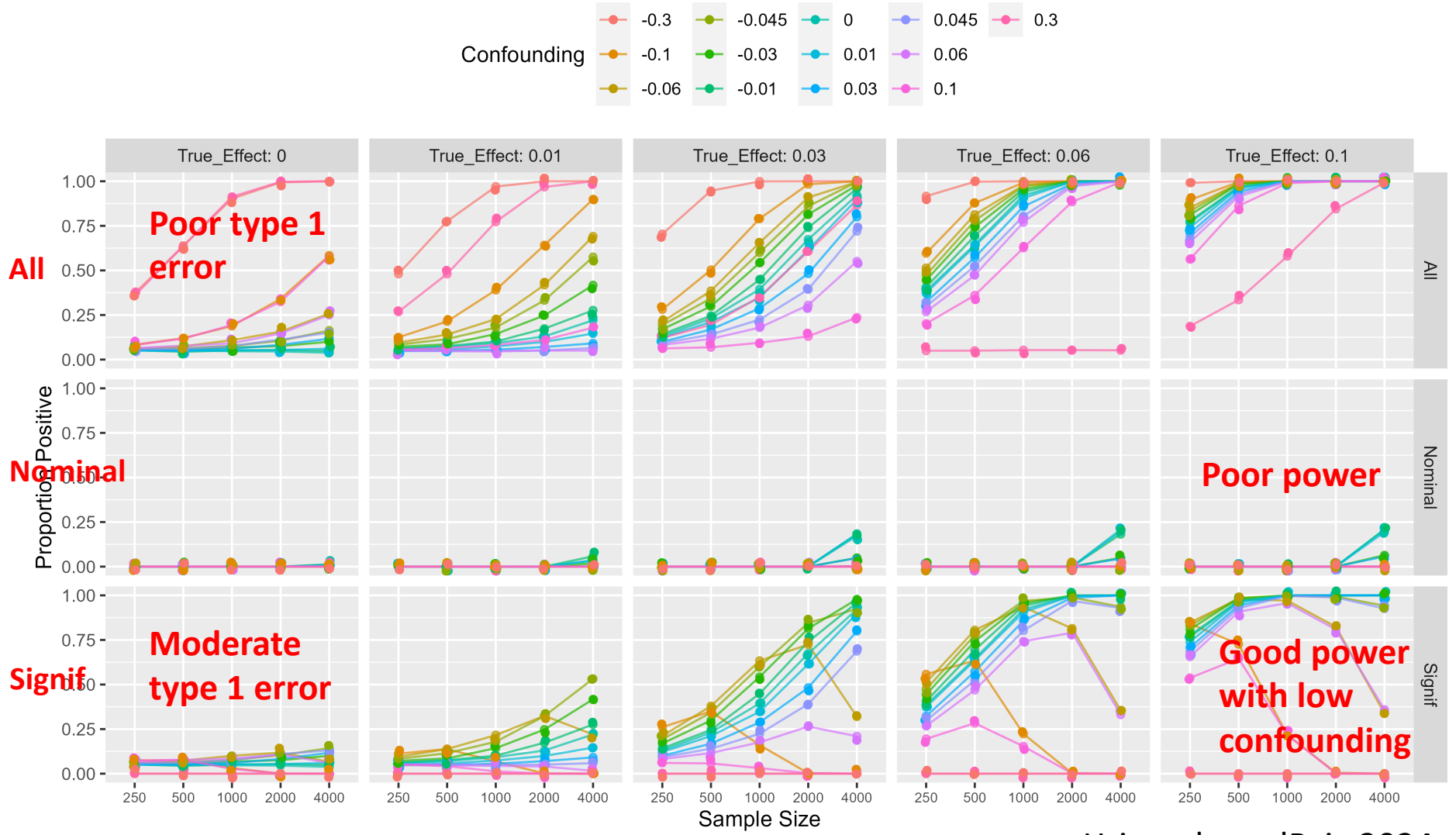
Hripcsak medRxiv 2024

Rule performance at the database level on simulation



Hripcsak medRxiv 2024

Rule performance at the database level on simulation



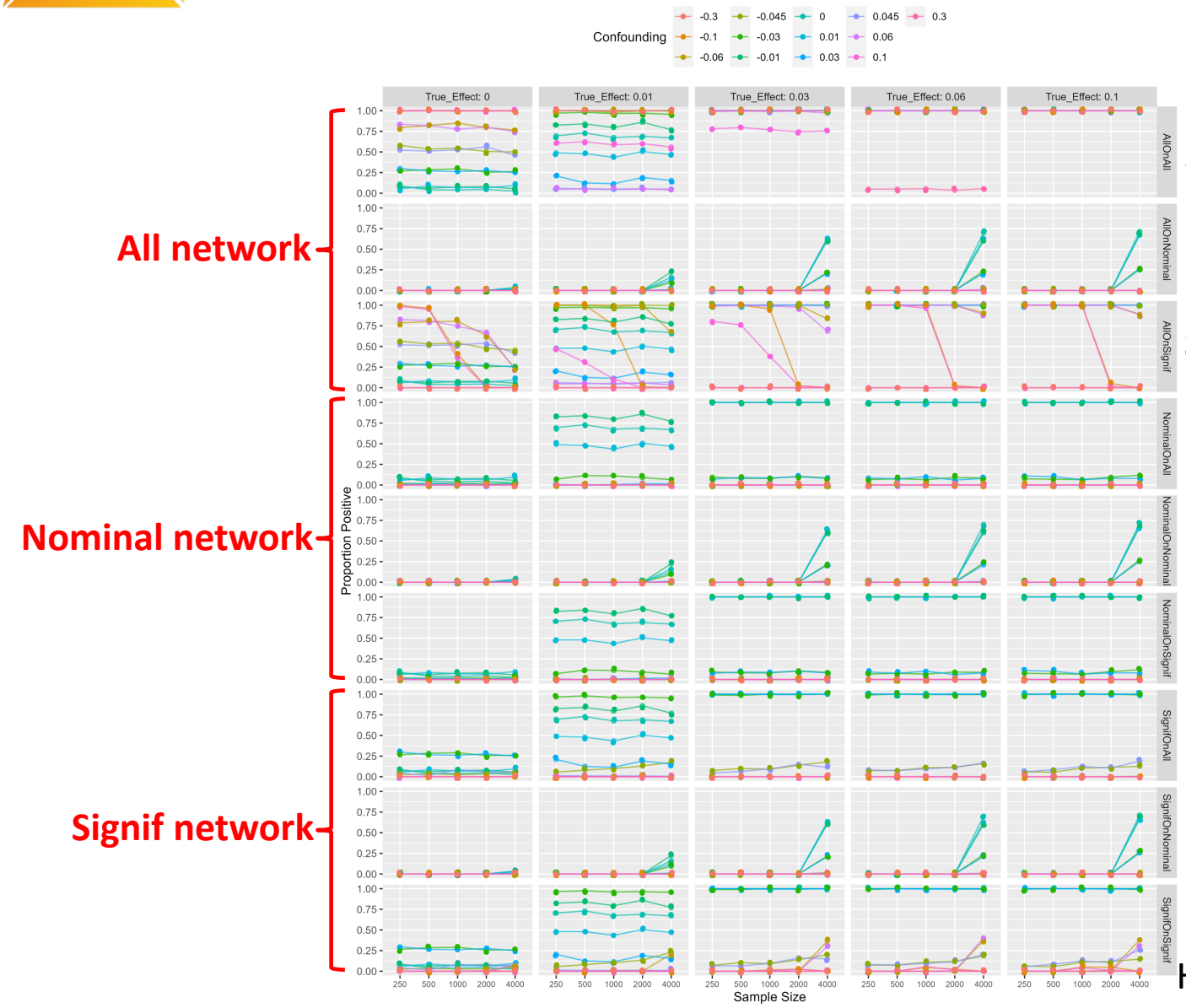
Hripcsak medRxiv 2024

Rule performance at the database level on simulation

- All (no diagnostic) – unacceptable type 1 error (near 1)
- Nominal (SMD>0.1) – unacceptable power (near 0)
- Signif (SMD statistically significant >0.1) – moderate type 1 error (0 to 0.2) and good power (near ideal)

Hripcsak medRxiv 2024

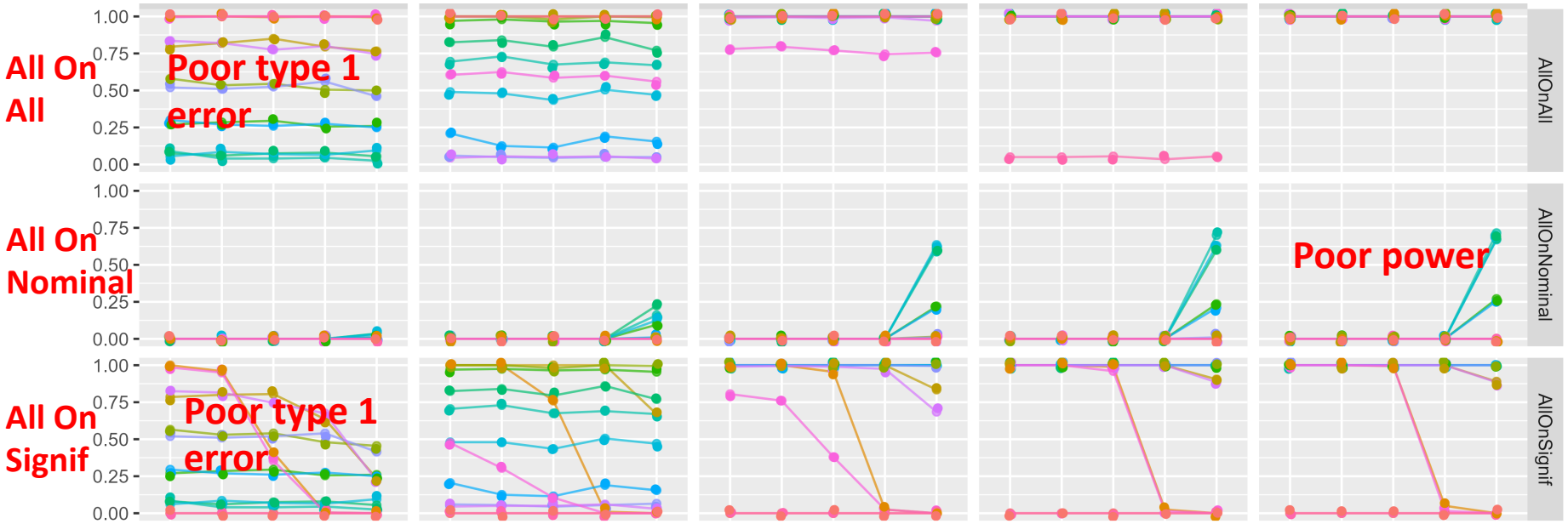
Rule performance at the network level on simulation



Hripcsak medRxiv 2024

Rule performance at the network level on simulation

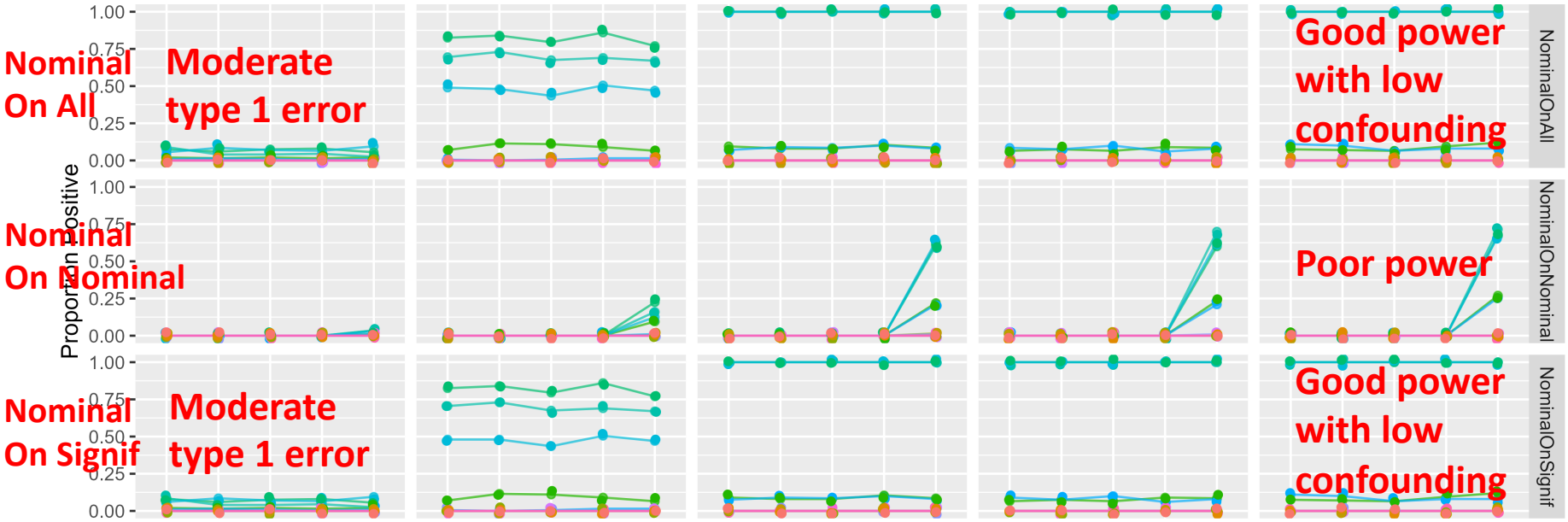
- All network = no network diagnostic
  - Three rows fail
  - Note: Signif just at database level fails
    - Network improves precision of effect estimate but not of SMD



Hripcsak medRxiv 2024

Rule performance at the network level on simulation

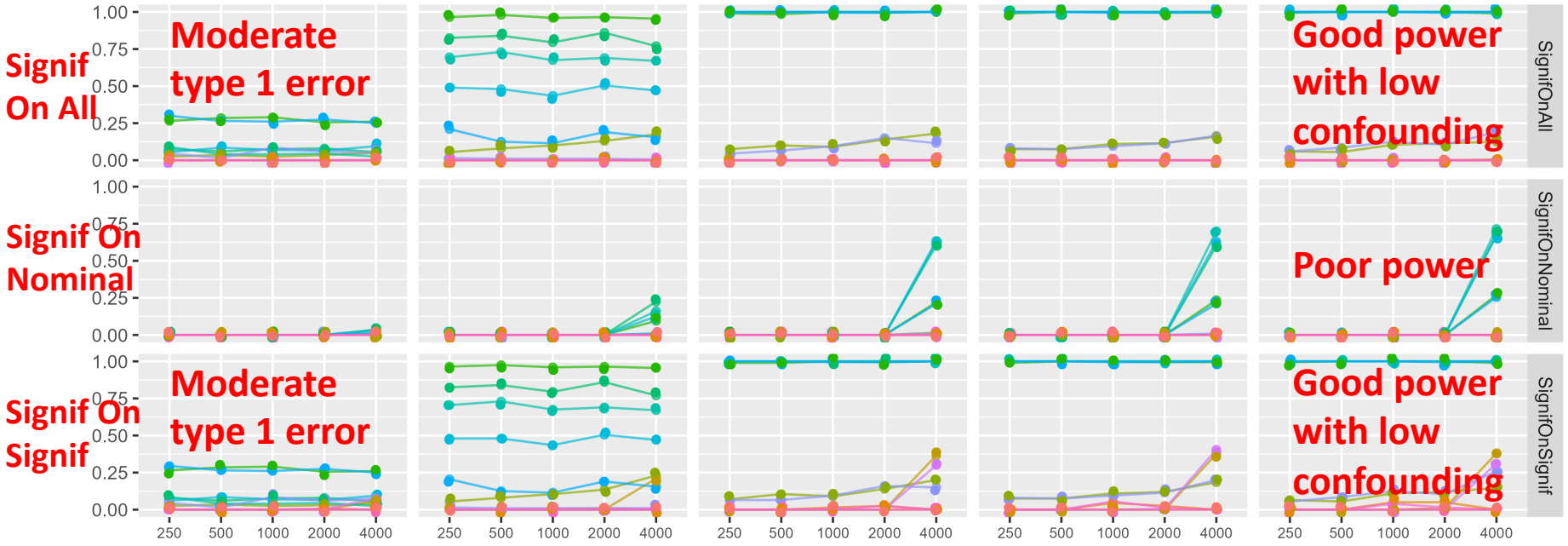
- Nominal at network level
  - Nominal-On-All, Nominal-On-Signif good here
  - Meta-analysis has enough power to avoid failing by chance



Hripcsak medRxiv 2024

Rule performance at the network level on simulation

- Signif at network level
  - Signif-On-All, Signif-On-Signif good here
  - But higher type 1 error



Hripcsak medRxiv 2024

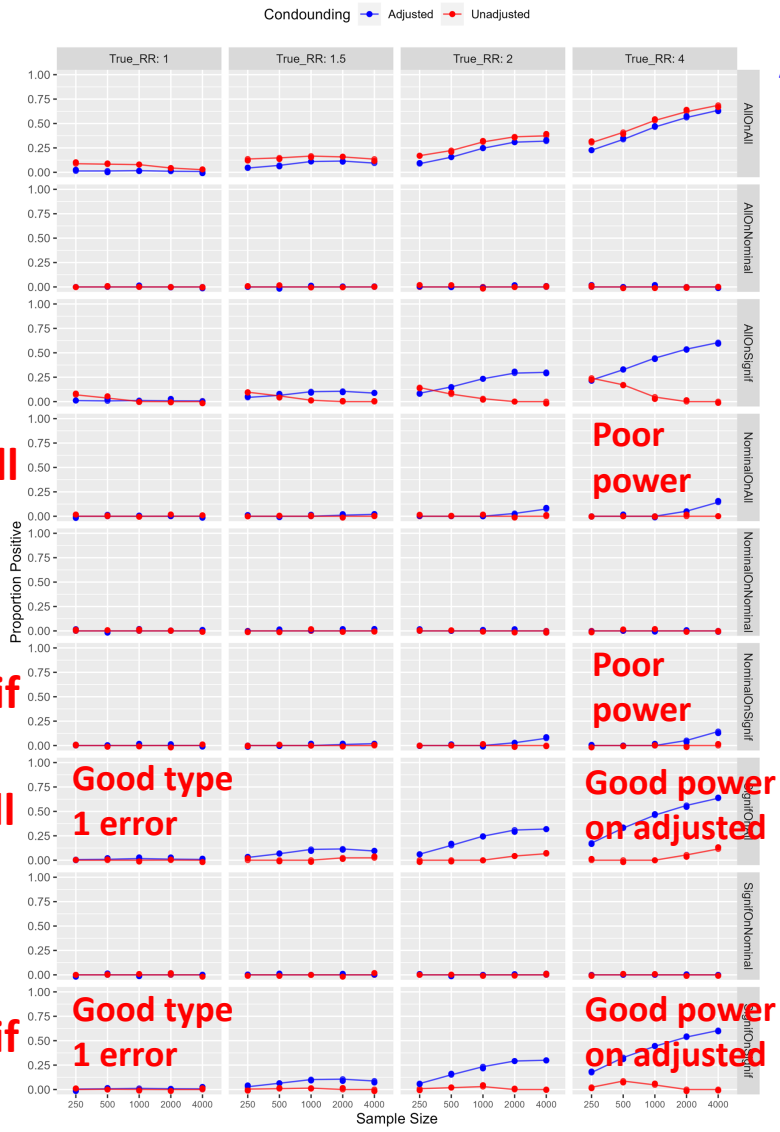
Rule performance at the network level on simulation

- These seem to work with moderate excess type 1 error but good power
  - Nominal-On-All
  - Nominal-On-Signif
  - Signif-On-All
  - Signif-On-Signif

Real-world data

- Reused data from OHDSI LEGEND hypertension and type 1 diabetes studies
  - [Suchard Lancet 2019, Khera BMJ Open 2022]
  - Four treatment comparisons
    - lisinopril vs hydrochlorothiazide, lisinopril vs metoprolol, sitagliptin vs liraglutide, sitagliptin vs glimepiride
  - 110 real negative controls (hazard ratio 1)
  - Corresponding synthetic positive controls (HR 1.5, 2, 4)
    - L1-regularized Poisson regression model
- Data and analysis
  - Three sources: Merative Medicare, Merative Medicaid, Optum EHR
  - 20,000 cases divided among “databases” with 250 to 4000 cases
  - 98,681 covariates, built a large-scale propensity model
  - Several analytic methods: unadjusted (crude) versus adjusted
  - Cox proportionate hazards model on matched or stratified sample or crude sample

Rule performance at the network level on real-world data



Unadjusted (more confounded)

Adjusted (less confounded)

NominalOnAll — **Poor power**

NominalOnSignif — **Poor power**

SignifOnAll — **Good type 1 error**; **Good power on adjusted**

SignifOnSignif — **Good type 1 error**; **Good power on adjusted**

Hripcsak medRxiv 2024

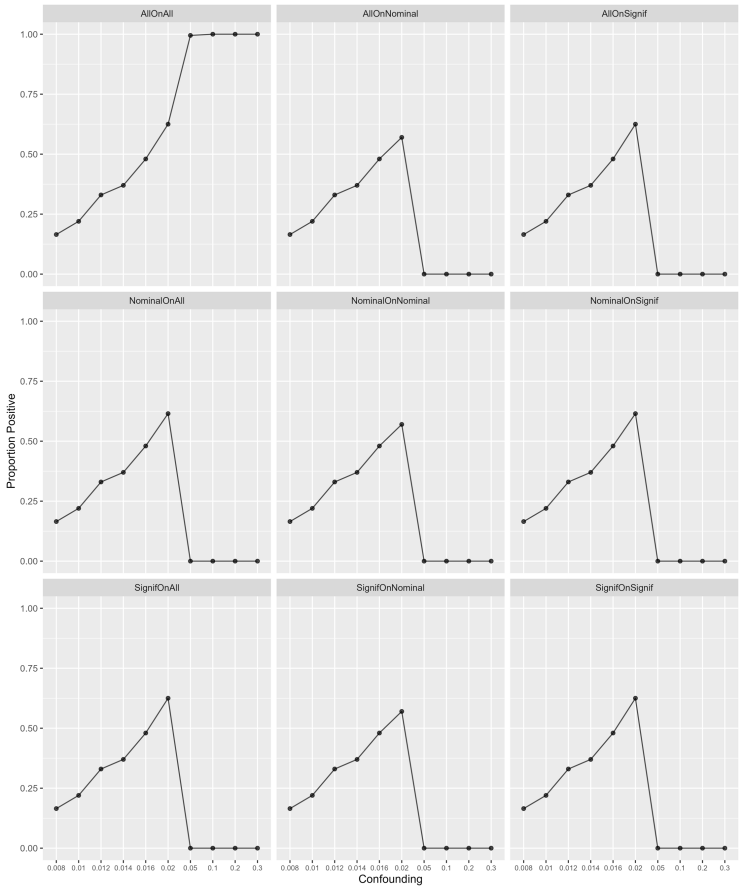
Rule performance at the network level on real-world data

- On real-world data
  - Nominal-On-All and Nominal-On-Signif have near zero power
    - And these were studies with good equipoise and proven good balance on the adjusted ones using larger sample sizes
  - Signif-On-All and Signif-On-Signif have good type 1 error and as good power as no diagnostics
- On 4 hypotheses, 3 data sources, 5 analytic methods, Signif-On-(All,Signif) worked

Hripcsak medRxiv 2024

Shouldn’t type 1 error be 0.05?

- Given a threshold on SMD, it is possible to create a bad-case simulation scenario
  - Typical study with 20,000 cases and 20 covariates under no true effect but with confounding, all 9 rules get type 1 error over 0.5
- We purposely found the weak points using our simulation
  - Could do Bayesian analysis
  - Probability of getting these parameters under reasonable priors is low (thus RWD result)

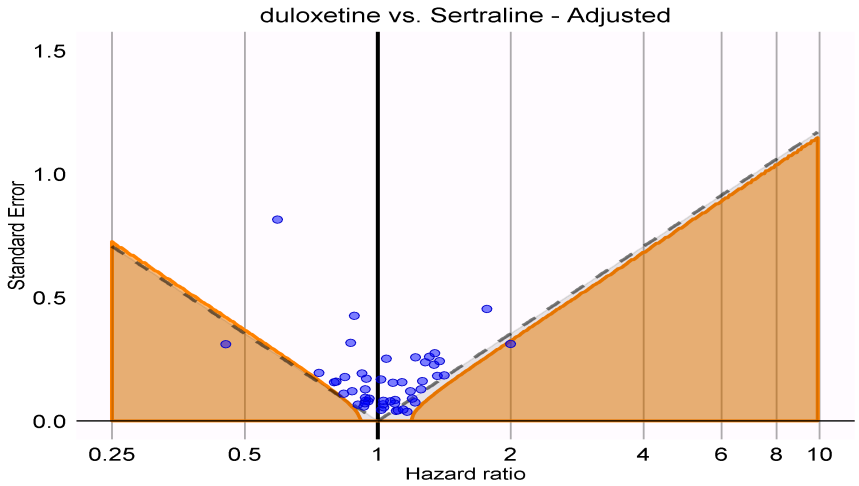


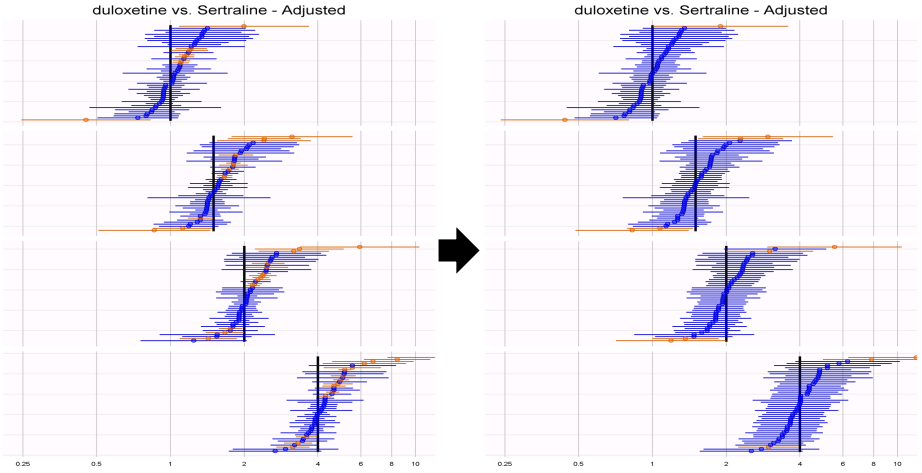
Hripcsak medRxiv 2024

Can correct for type 1 error

Confidence interval calibration using negative controls: residual bias

- Address residual confounding using hypotheses you know the answer for
  - 50 to 100 controls
- If too many are positive, then systematic error is operative
- Calibrate to keep the type 1 error at 0.05

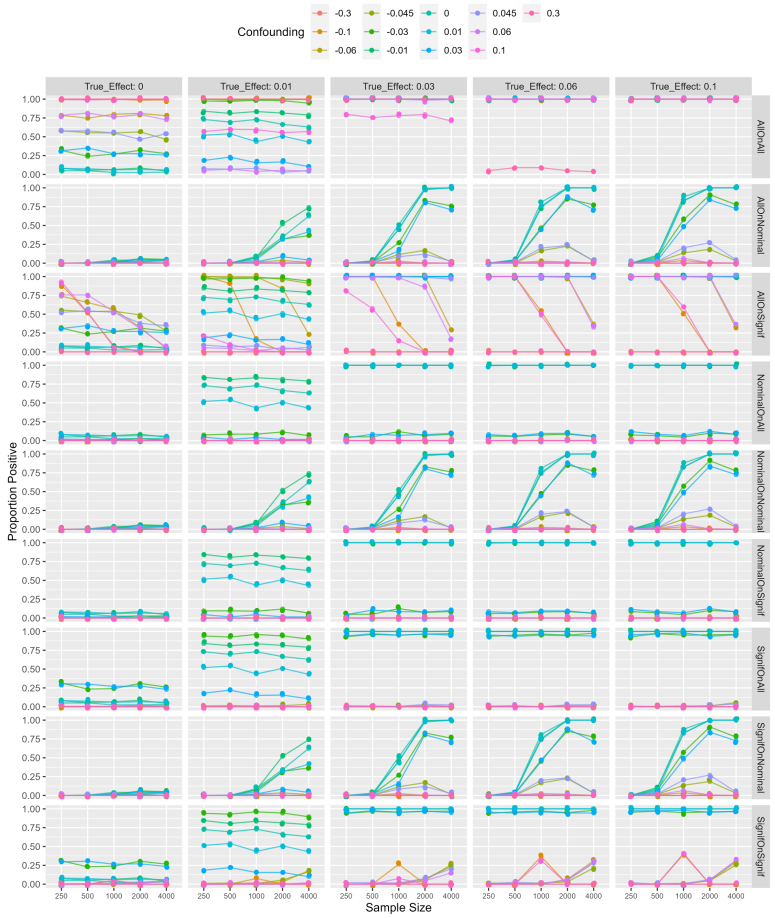




Schuemie PNAS 2018

Same results for 20 covariates

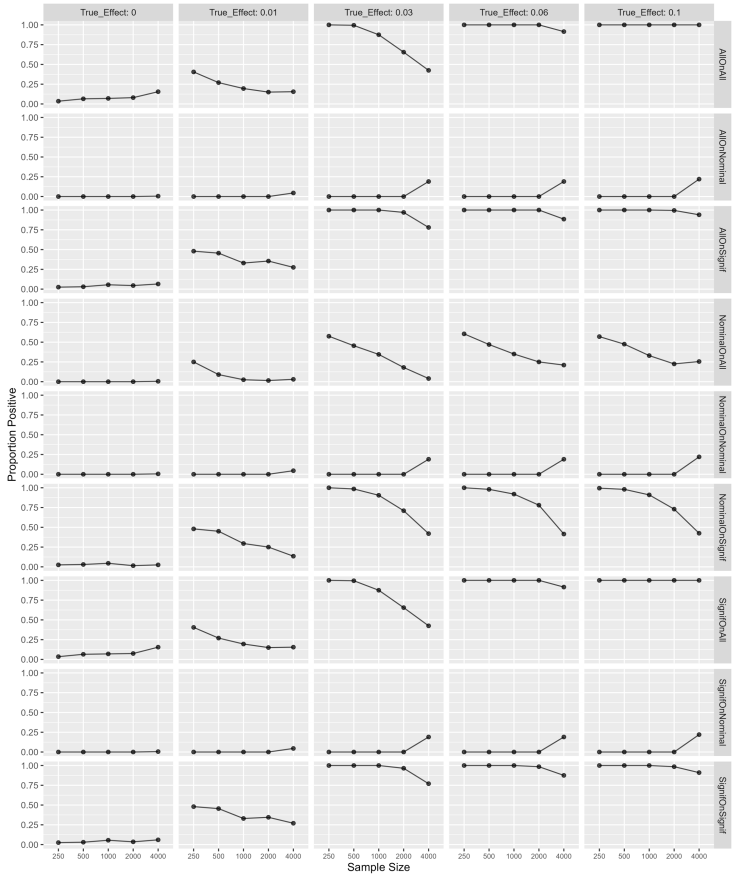
- Curve shifted to the left, but same pattern and tradeoff for type 1 error versus power



Hripcsak medRxiv 2024

What if confounding is heterogeneous?

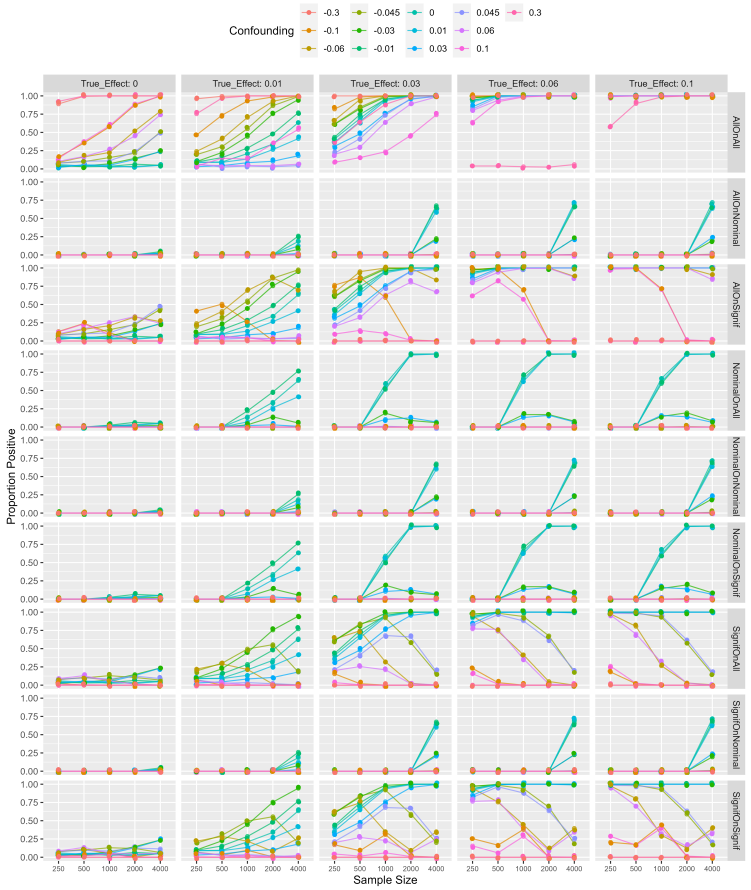
- The effective rules still work
  - Signif-On-Signif has a little more power and a little less type 1 error than Signif-On-All



Hripcsak medRxiv 2024

What if only 5 databases

- Nominal at network level (which appeared otherwise to have potential in simulations) loses all power on smaller databases
  - Meta-analysis of the SMDs no longer gain enough precision to avoid chance rejection
- Thus even simulation favors Signif-On-Signif



Hripcsak medRxiv 2024

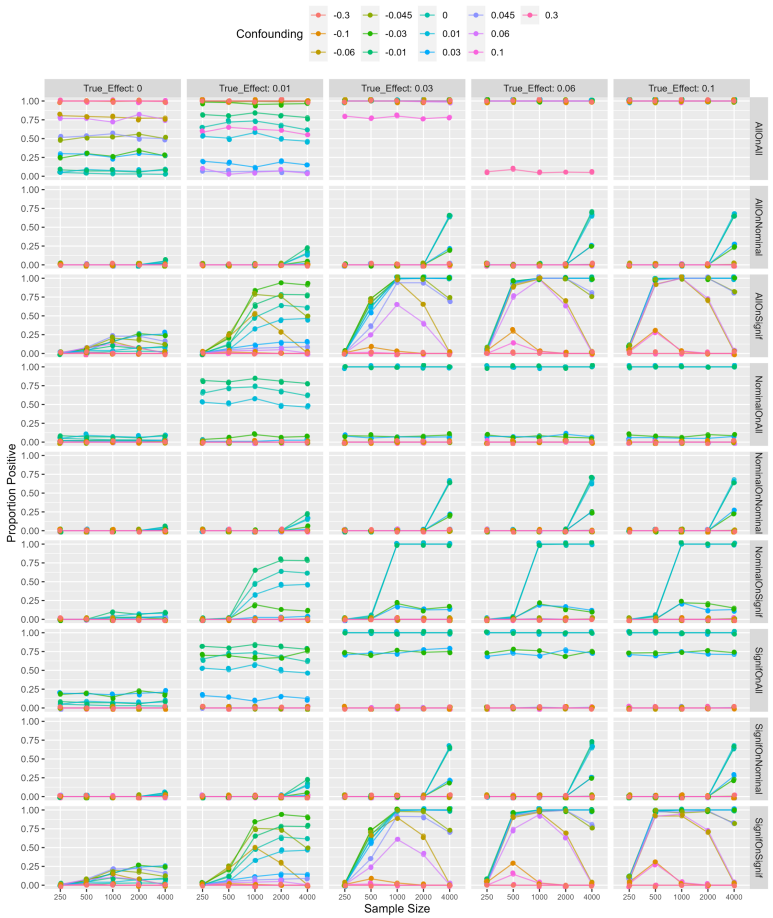
Sensitivity to prevalence

- If drop prevalence of positive covariate (10%) or outcome (1%), get same results

Hripcsak medRxiv 2024

Is Bonferroni correction needed?

- Eliminating the Bonferroni correction does not improve the type 1 error rate but does drop power to 0 at the smallest sample sizes



Hripcsak medRxiv 2024

Doesn't increasing # covariates hide confounding?

- Bonferroni correction for many covariates effectively raises the SMD threshold; doesn't that unfairly allow more confounding?
- If we have actual knowledge that there is no confounding, then follow that
  - (never happens)
- Otherwise, assume confounders distributed in the covariates
  - Probability 0.001 of covariate being imbalanced
  - Sample size 4000; 10,000 covariates; reject 0.62 of studies
  - With 60,000 covariates, rose to 1.0
- Bonferroni does **not** overwhelm imbalance detection

Hripcsak medRxiv 2024

Can you produce a good PS model in such small databases?

- Yes
  - Using same data sources and hypotheses
  - Worked well ≥1000, usually >250, sometimes 125
  - [Schuemie OHDSI 2023]

Haven’t we already decided that a statistical test for imbalance is bad?

- Mostly arguing against test for >0, not >0.1
- [Imai J Royal Stat Soc A 2008]
  - Statistical test depends on sample size but imbalance does not
    - But the impact of imbalance does depend on sample size
  - P-value thresholds are arbitrary
    - 0.1 is also arbitrary
    - It is an empirical question that can be studied with RWD
  - The target of analysis is the sample, not an underlying population
    - Insofar as we are distinguishing chance from systematic imbalance, we do care about an underlying population
    - Can see the method as a heuristic

Hripcsak medRxiv 2024

Shouldn’t we correct for imbalance?

- We sometimes correct for imbalance that is clearly chance (RCTs)
  - You can still correct for imbalance
  - We are deciding whether to reject a study, not whether to correct once it passes

Conclusions

- Small cohorts result in rejection for chance imbalance (SMD>0.1) and zero power
- As sample size falls, effect CIs lengthen, rendering small confounding less important
  - Using a statistical test for sufficient imbalance raises the threshold where a given degree of confounding is tolerable
- Our results comparing no diagnostic (old), nominal threshold (old), statistical test (new)
  - Statistical test maintains the best type-1-error to power balance across the simulations and RWD

Hripcsak medRxiv 2024

Conclusions

- Meta-analysis of network studies may produce a more precise effect estimate
  - Therefore you also need a more precise diagnostic for imbalance, else systematic bias will predominate
  - Our results show that meta-analysis of SMDs and a statistical test produce the best type-1-error to power balance

Hripcsak medRxiv 2024

Conclusions

- The statistical test for imbalance makes it feasible to check thousands of covariates
  - Regardless of how many confounders are adjusted for, the data set includes information about imbalance and the effect of potential confounding
  - Not checking for imbalance on all covariates is a head-in-the-sand approach
  - Imbalanced variables should be justified as known or proven instruments

Hripcsak medRxiv 2024

Conclusions

- Can produce a simulation with type 1 error greater than alpha
  - But this is true across sample sizes, number of covariates, and diagnostics
  - Therefore aimed for best balance of type 1 error and power

Hripcsak medRxiv 2024

Recommendations

- For PS-adjusted cohort studies, check for imbalance of covariates
- Check for imbalance (SMD) statistically significantly greater than 0.1 (or other pre-specified threshold) in any covariate after Bonferroni correction
- Network studies require meta-analysis of each covariate and checking for statistically significant imbalance (at database and network level)
- Check all available covariates, not just the ones adjusted for

Hripcsak medRxiv 2024

Team and funding

George Hripcsak, MD, Columbia University

Linying Zhang, PhD , Washington University in St. Louis

Kelly Li, University of California, Los Angeles

Marc A. Suchard, Md, PhD, University of California, Los Angeles

Patrick B. Ryan, PhD, Johnson & Johnson, Columbia University

Martijn J. Schuemie, PhD, Johnson & Johnson

This work is partially supported through US National Institutes of Health grants (T15 LM007079, R01 LM006910, and R01 HL169954).