Reliability in Observational Research: Assessing Covariate Imbalance in Small Studies

George Hripcsak, MD, MS

Biomedical Informatics, Columbia University

COLUMBIA | COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

Large-scale propensity score (LSPS)

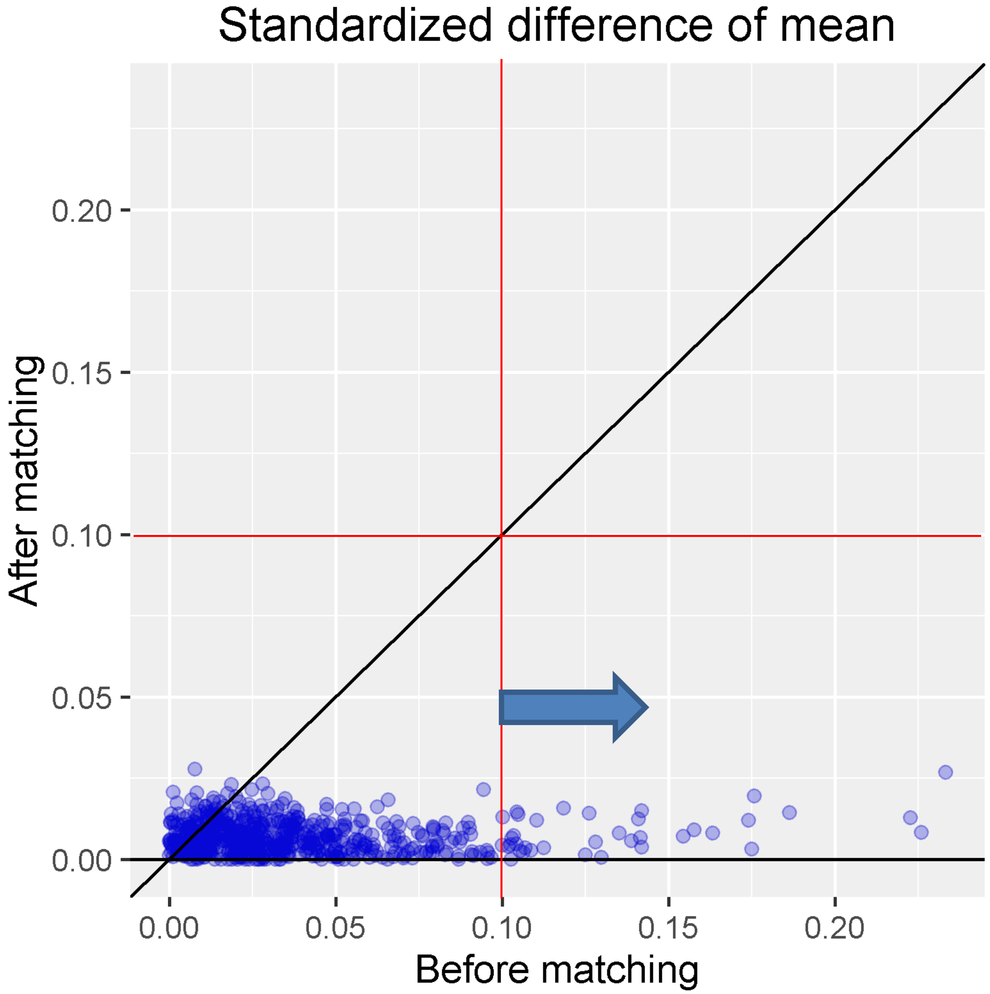
- A **systematic** approach to propensity adjustment
- Use a large set of covariates (10,000 < n < 100,000)
- But don't want to balance *everything*
  - Mediators – pre-treatment
  - Simple colliders – pre-treatment
  - Instruments – diagnostics, domain knowledge
  - M-bias – correlation with underlying 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

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)

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

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

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

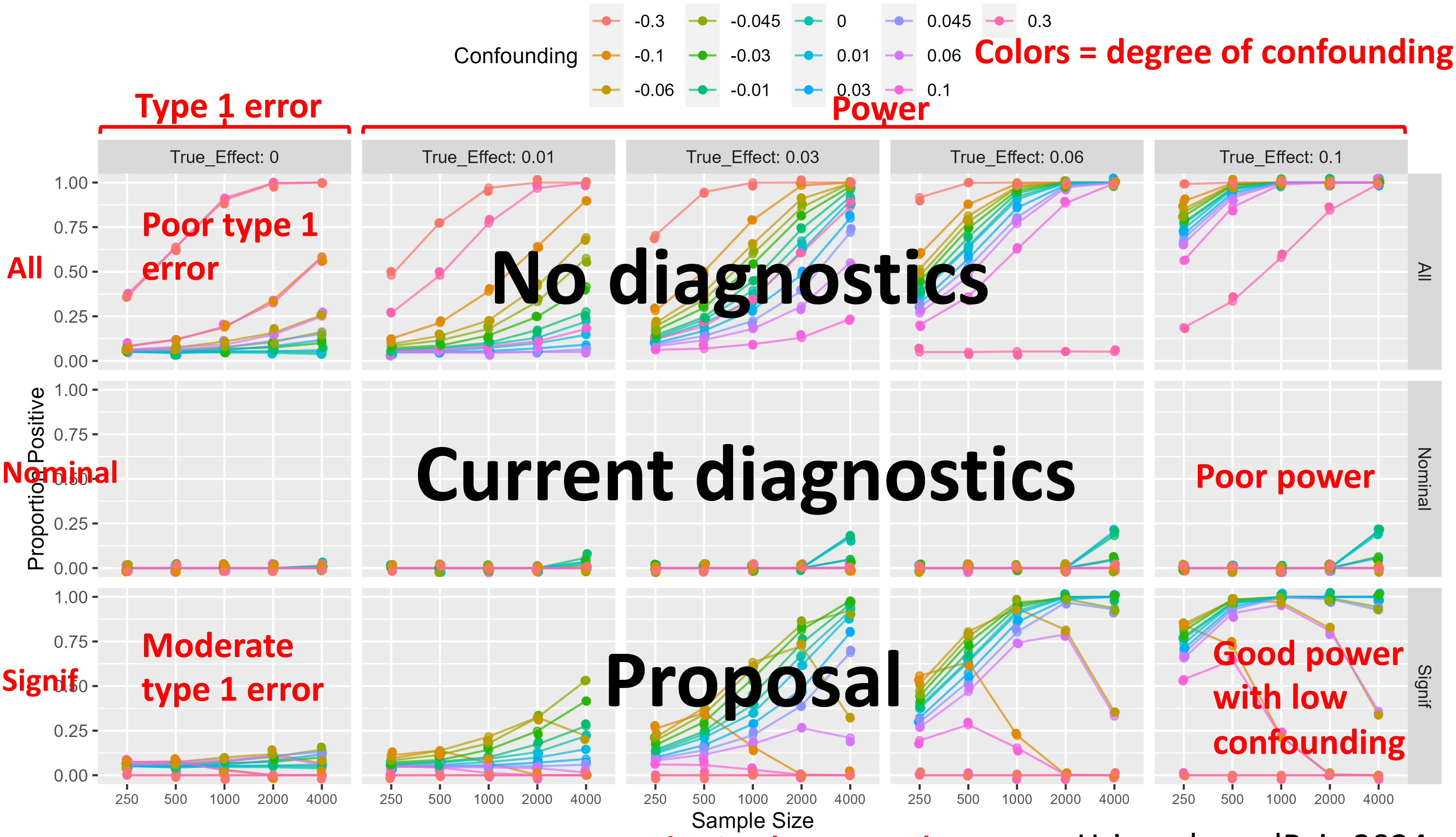
Hripcsak medRxiv 2024

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



Colors = degree of confounding

Type 1 error

Power

All

Nominal

Signif

Poor type 1 error

No diagnostics

Current diagnostics

Poor power

Moderate type 1 error

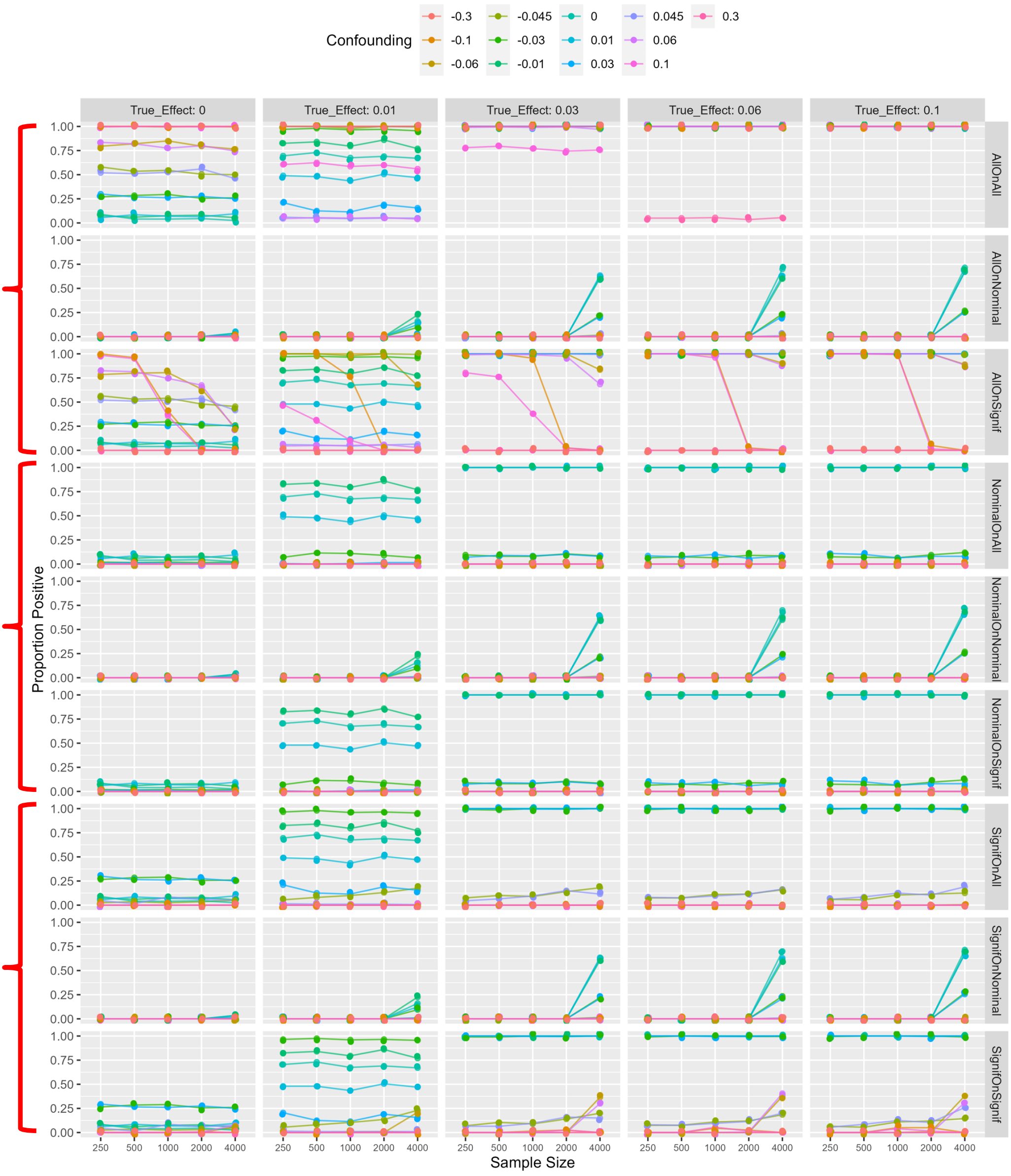
Proposal

Good power with low confounding

Sample size (250-4000)

Hripcsak medRxiv 2024

Rule performance at the network level on simulation



All network

Nominal network

Signif network

All database

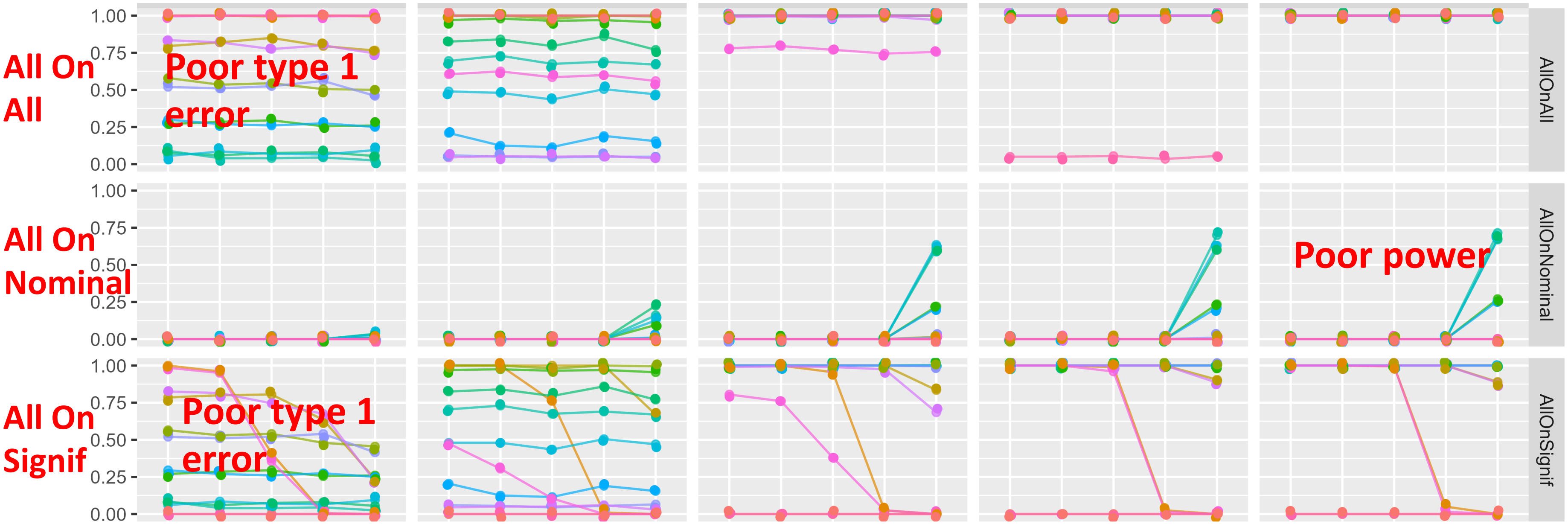
Nominal database

Signif database

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

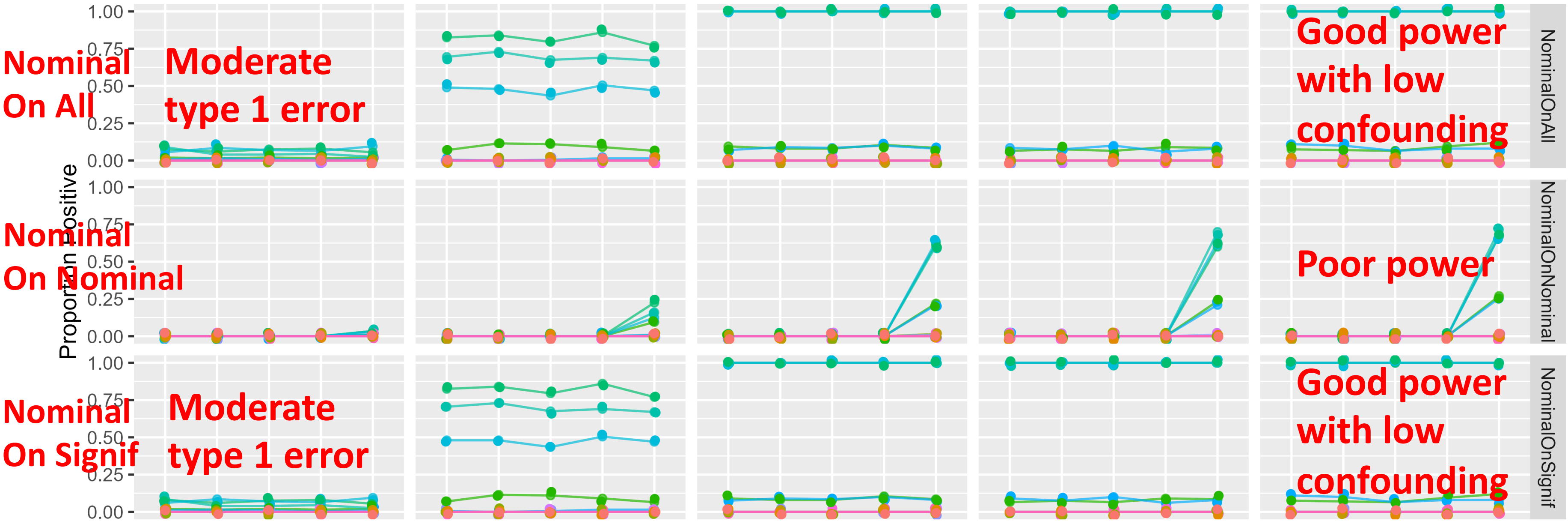


Cannot ignore balance at the network level

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



Nominal On All — Moderate type 1 error — Good power with low confounding

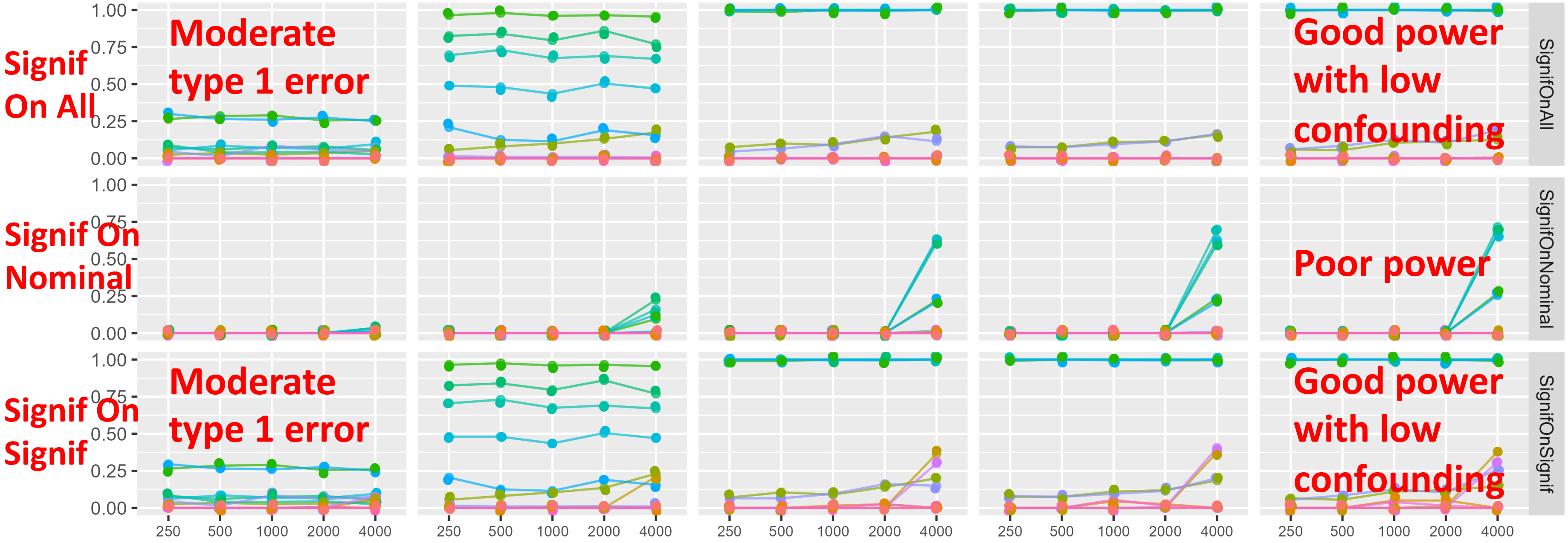
Nominal On Nominal — Poor power

Nominal On Signif — Moderate type 1 error — Good power with low confounding

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



Signif On All

Moderate type 1 error

Good power with low confounding

Signif On Nominal

Poor power

Signif On Signif

Moderate type 1 error

Good power with low confounding

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

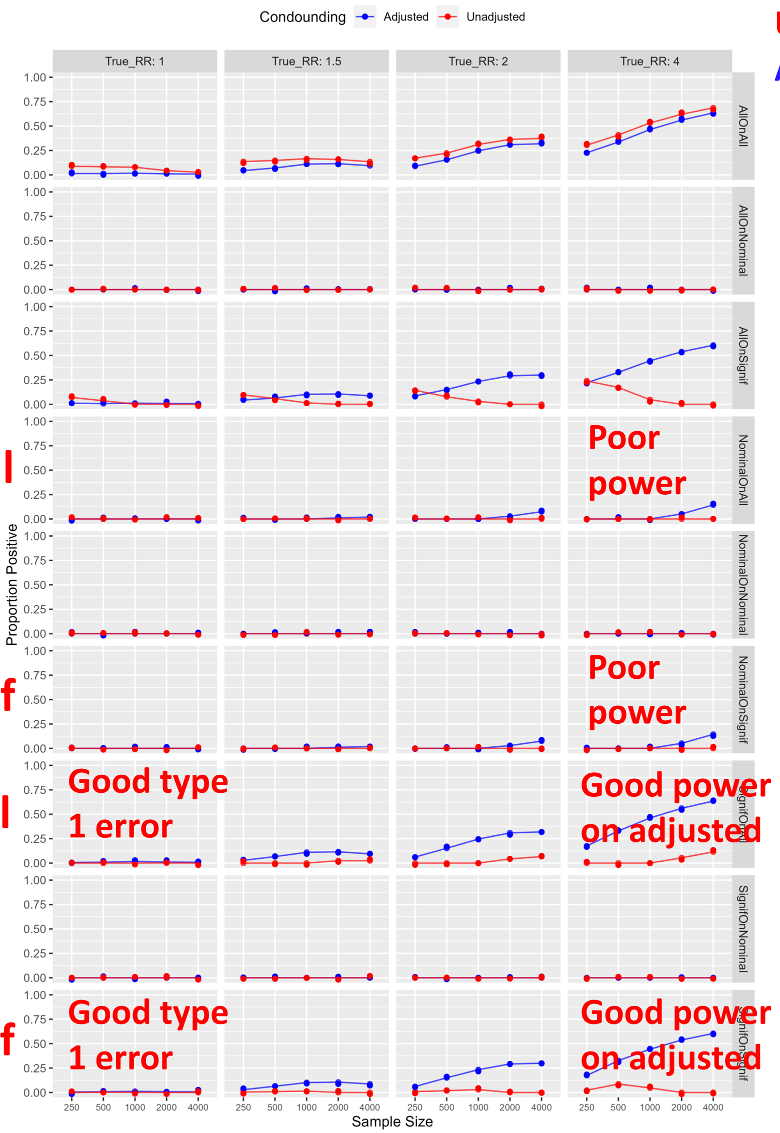
Real-world data

- 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

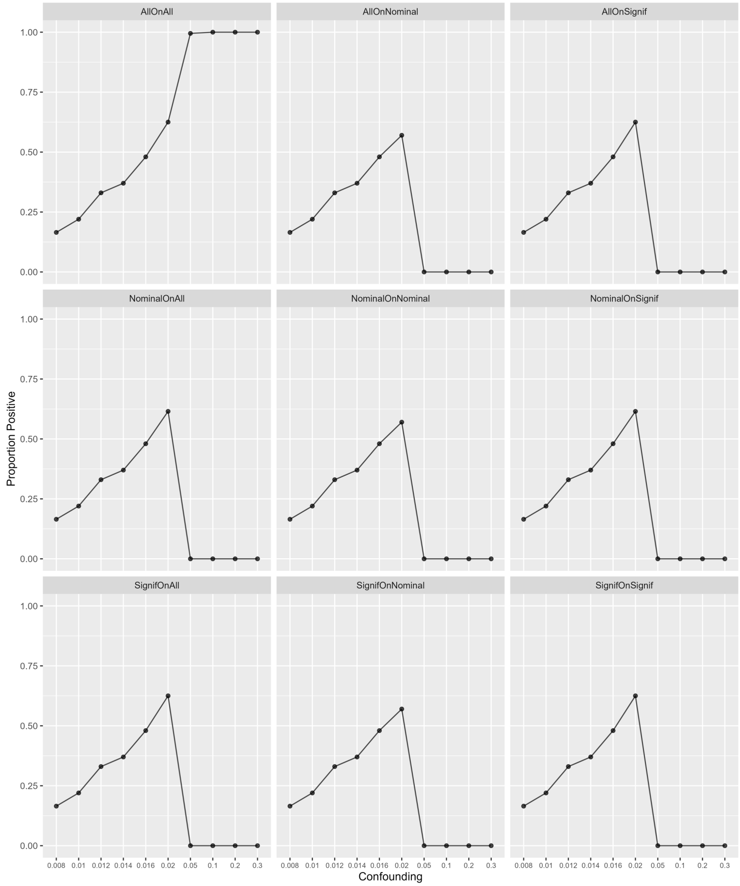
Good power on adjusted

Type 1 error < 0.05 and power as good as no diagnostics

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)

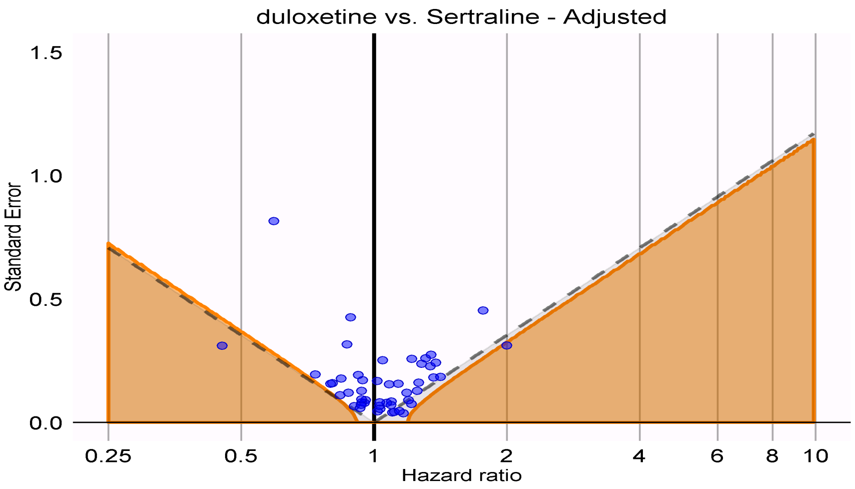


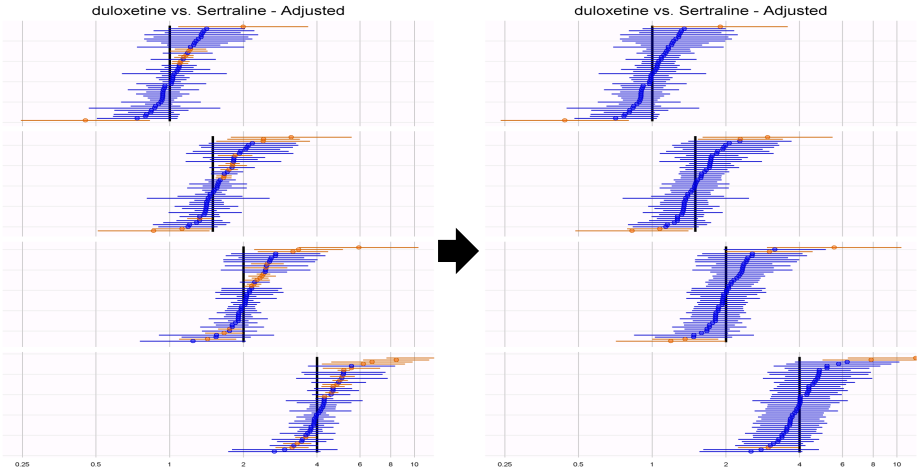
Hripcsak medRxiv 2024

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

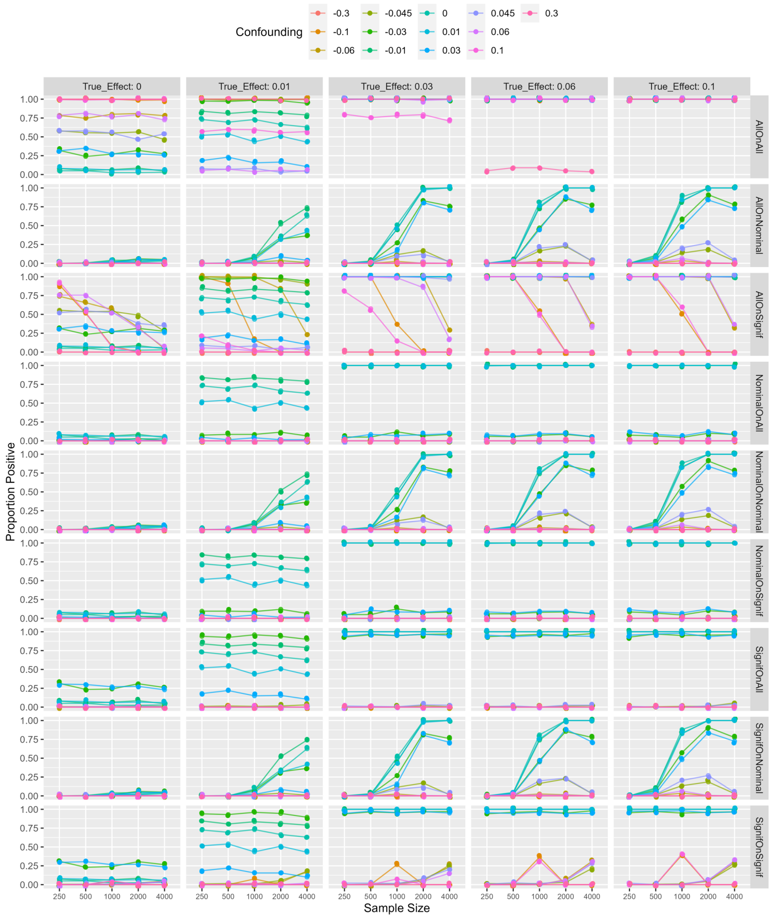




Schuemie PNAS 2018

Same results for 20 covariates

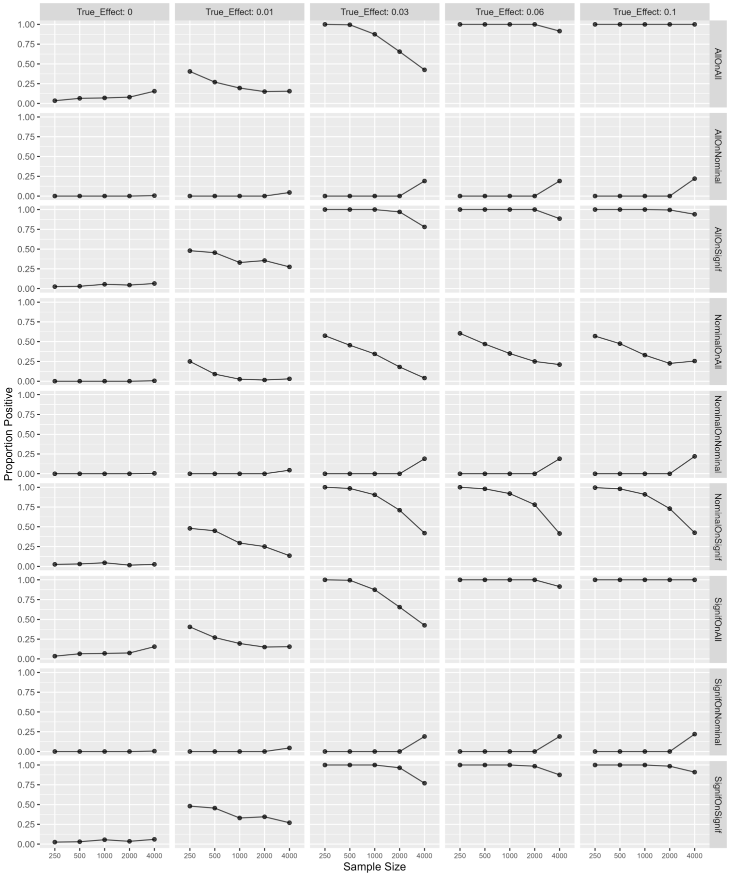
- Curve shifted to the left, but same pattern and tradeoff for type 1 error versus power



Hripcsak medRxiv 2024

What if confounding is heterogeneous?

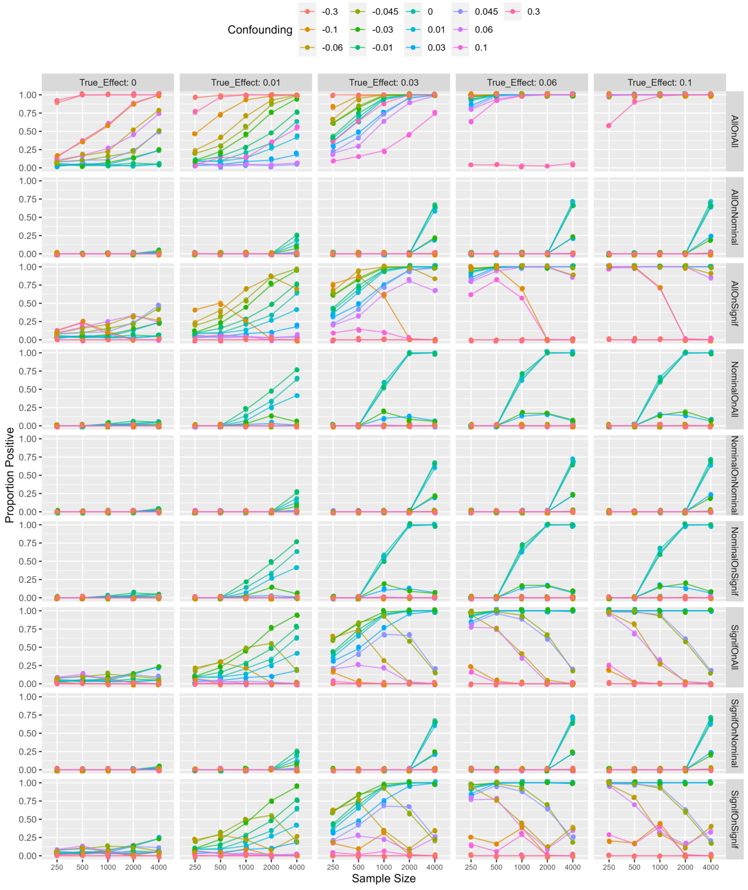
- The effective rules still work
  - Signif-On-Signif has a little more power and a little less type 1 error than Signif-On-All



Hripcsak medRxiv 2024

What if only 5 databases

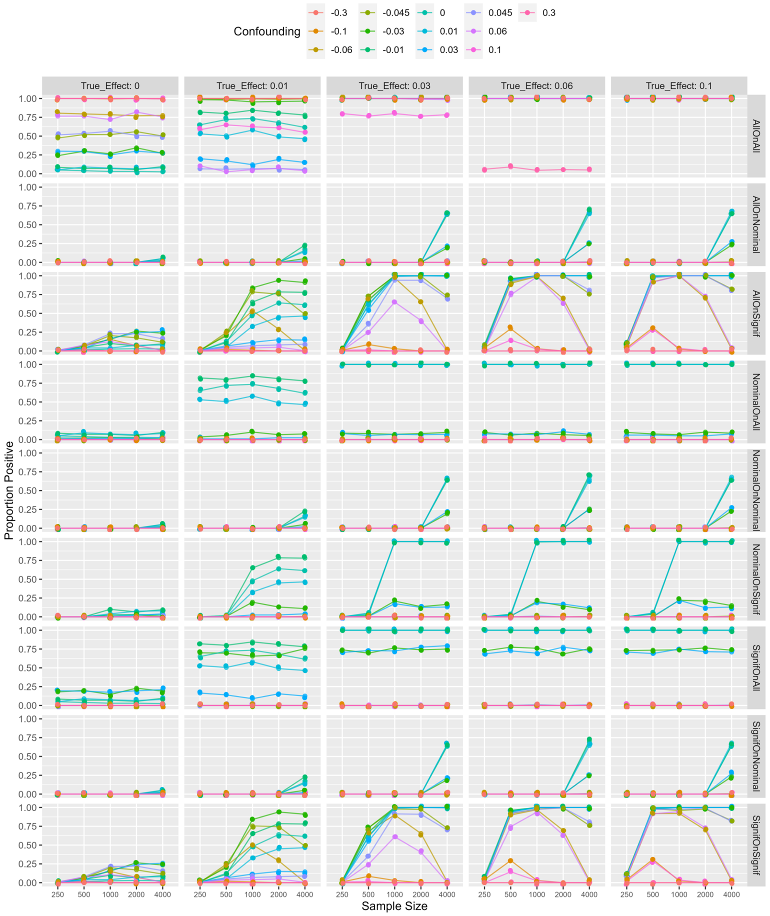
- Nominal at network level (which appeared otherwise to have potential in simulations) loses all power on smaller databases
  - – Meta-analysis of the SMDs no longer gain enough precision to avoid chance rejection
- Thus even simulation favors Signif-On-Signif



Hripcsak medRxiv 2024

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]

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

Must do meta-analysis of diagnostics

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

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
Yong Chen, PhD, University of Pennsylvania

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

Hripcsak G, Zhang L, Li K, Suchard MA, Ryan PB, Schuemie MJ. Assessing Covariate Balance with Small Sample Sizes. medRxiv 2024 Apr 24:2024.04.23.24306230. doi: 10.1101/2024.04.23.24306230. PMC11071580