Objective Diagnostics in observational research

Martijn Schuemie, Marc Suchard, Patrick Ryan, Yong Chen, George Hripcsak

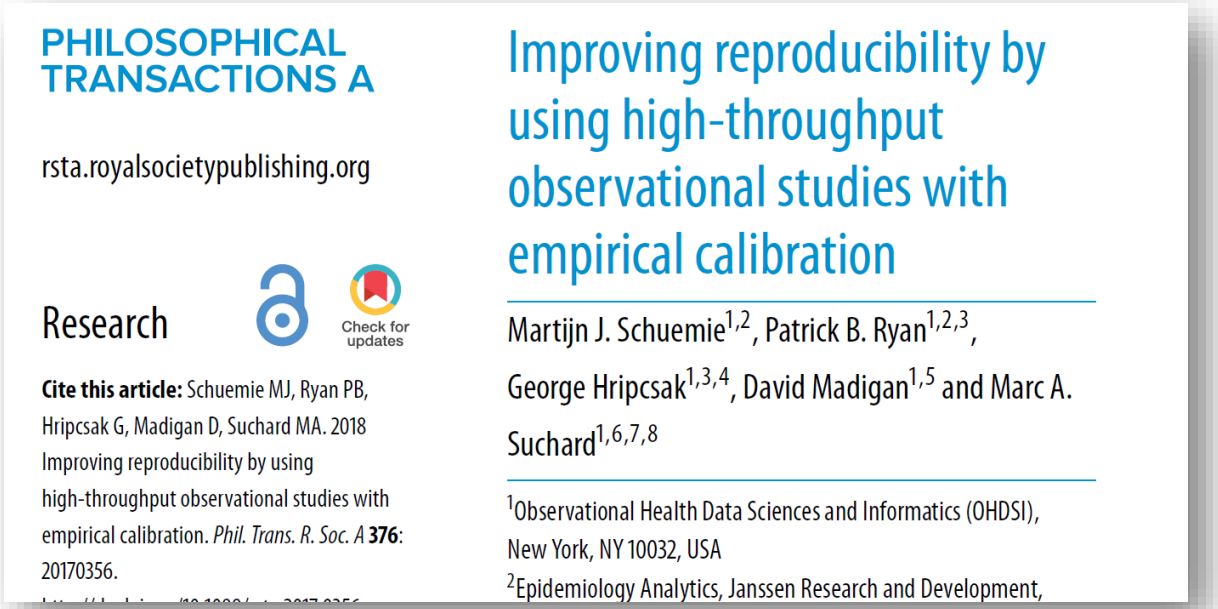
Current state of observational healthcare observational research

Marc A. Suchard, MD, PhD

Department of Biostatistics, University of California, Los Angeles

Long-standing issues

- Lack of trust in real-world evidence to guide clinical practice
- Major issues:
  - Observational study bias
    - e.g. confounding
  - Publication bias
  - P-hacking



PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Check for updates

Cite this article: Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. 2018 Improving reproducibility by using high-throughput observational studies with empirical calibration. *Phil. Trans. R. Soc. A* **376**: 20170356.

Improving reproducibility by using high-throughput observational studies with empirical calibration

Martijn J. Schuemie[1,2], Patrick B. Ryan[1,2,3], George Hripcsak[1,3,4], David Madigan[1,5] and Marc A. Suchard[1,6,7,8]

[1]Observational Health Data Sciences and Informatics (OHDSI), New York, NY 10032, USA

[2]Epidemiology Analytics, Janssen Research and Development,

Credibility of observational studies

ESC European Society of Cardiology

European Heart Journal (2018) **39**, 3417–3438
doi:10.1093/eurheartj/ehy407

CLINICAL REVIEW
Controversies in cardiovascular medicine

Association is not causation: treatment effects cannot be estimated from observational data in heart failure

Christopher J. Rush, Ross T. Campbell, Pardeep S. Jhund, Mark C. Petrie, and John J.V. McMurray*

British Heart Foundation Cardiovascular Research Centre, Institute of Cardiovascular and Medical Sciences, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK

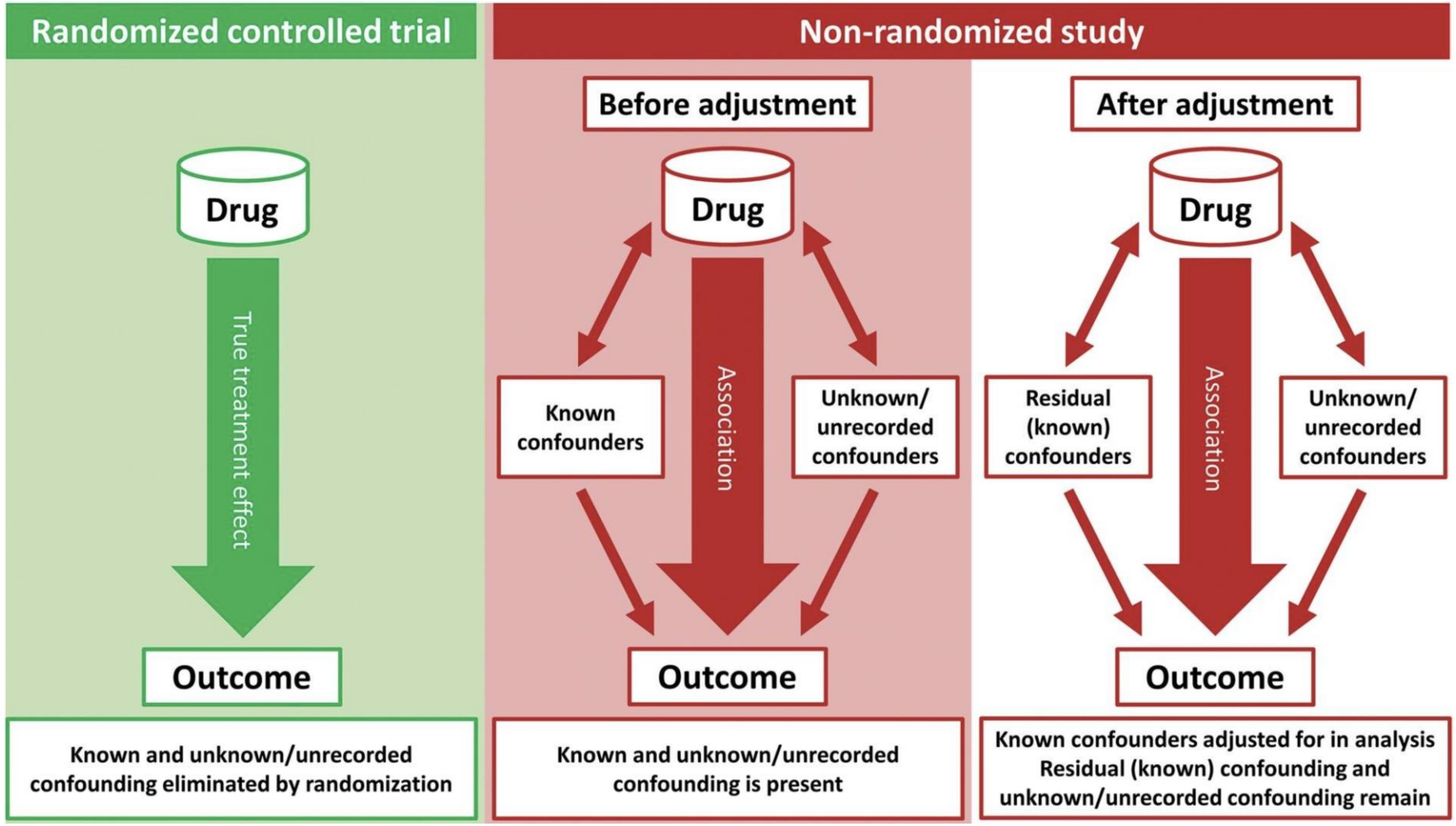
Received 16 January 2018; revised 1 April 2018; editorial decision 22 June 2018; accepted 27 June 2018; online publish-ahead-of-print 1 August 2018

Aims Treatment 'effects' are often inferred from non-randomized and observational studies. These studies have inherent biases and limitations, which may make therapeutic inferences based on their results unreliable. We compared the conflicting findings of these studies to those of prospective randomized controlled trials (RCTs) in relation to pharmacological treatments for heart failure (HF).

Methods and results We searched Medline and Embase to identify studies of the association between non-randomized drug therapy and all-cause mortality in patients with HF until 31 December 2017. The treatments of interest were: angiotensin-

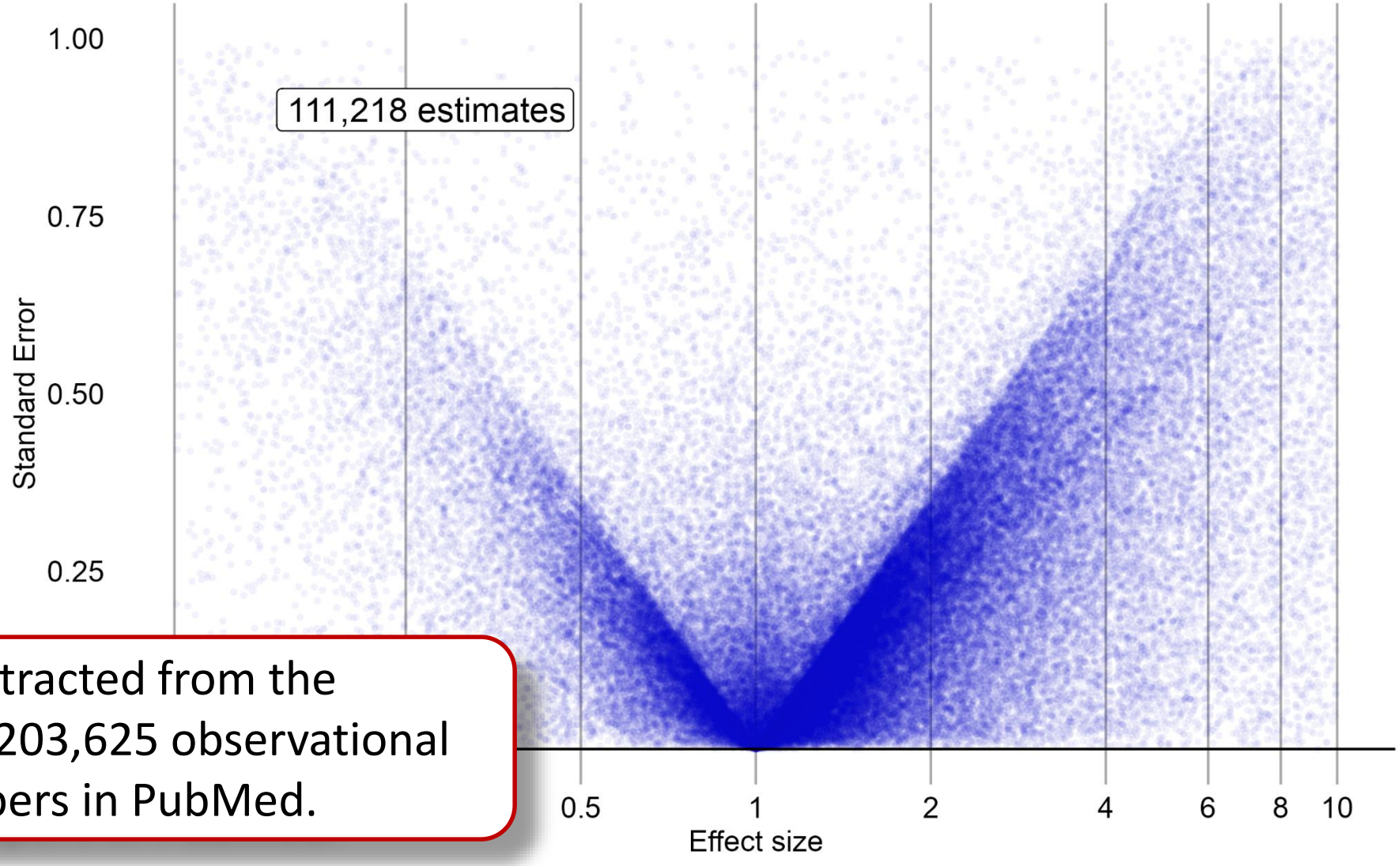
- Mistrust varies across clinical fields
- Particularly disparaged in cardiology

Residual study bias / systematic error



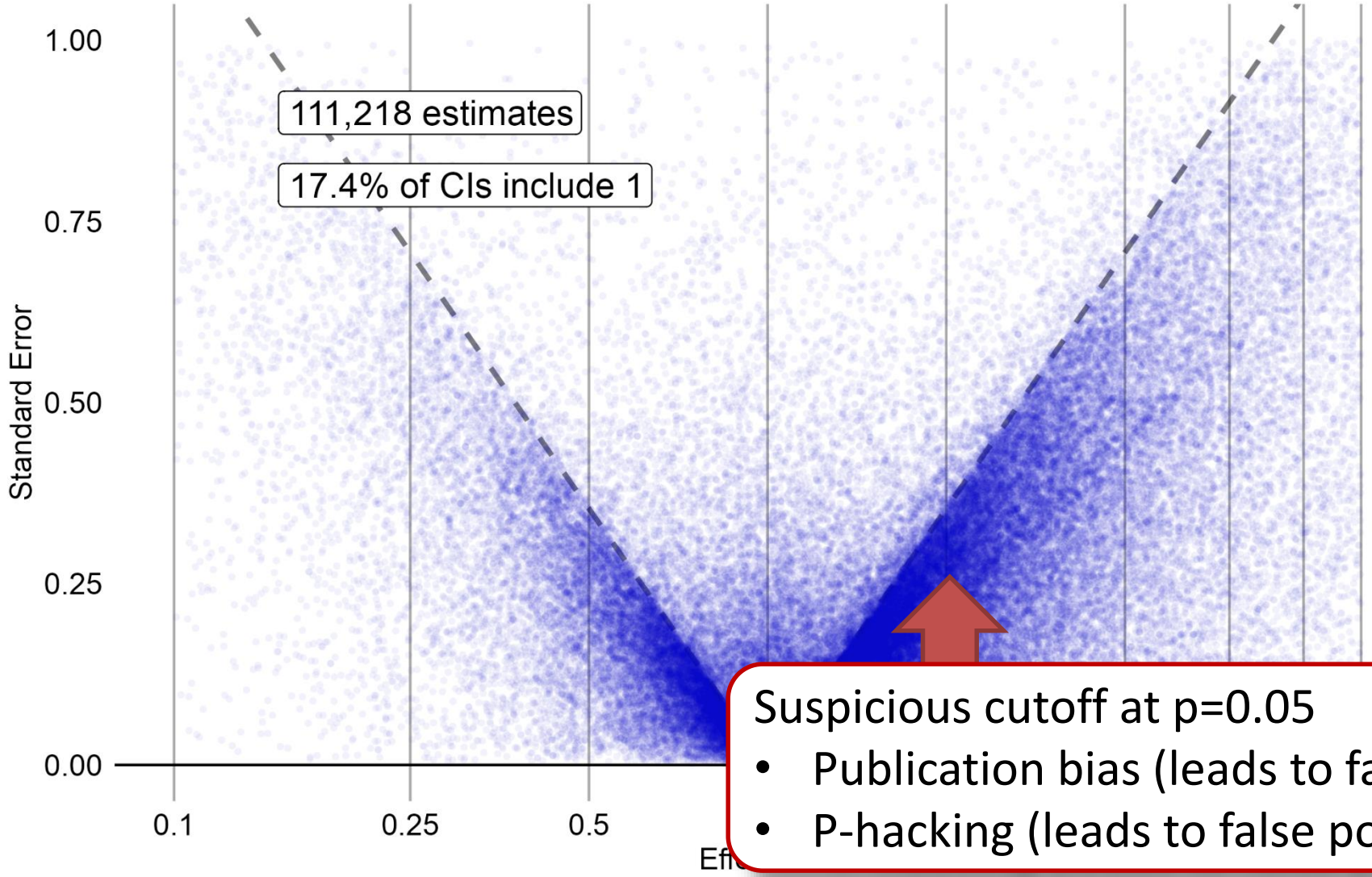
Rush et al., 2018

Published observational study results



Estimates extracted from the abstracts of 203,625 observational research papers in PubMed.

Published observational study results



Suspicious cutoff at p=0.05

- Publication bias (leads to false positives)
- P-hacking (leads to false positives)

LEGEND principles and the LEGEND Hypertension Study

Martijn J. Schuemie PhD

Observational Health Data Analytics

Johnson & Johnson

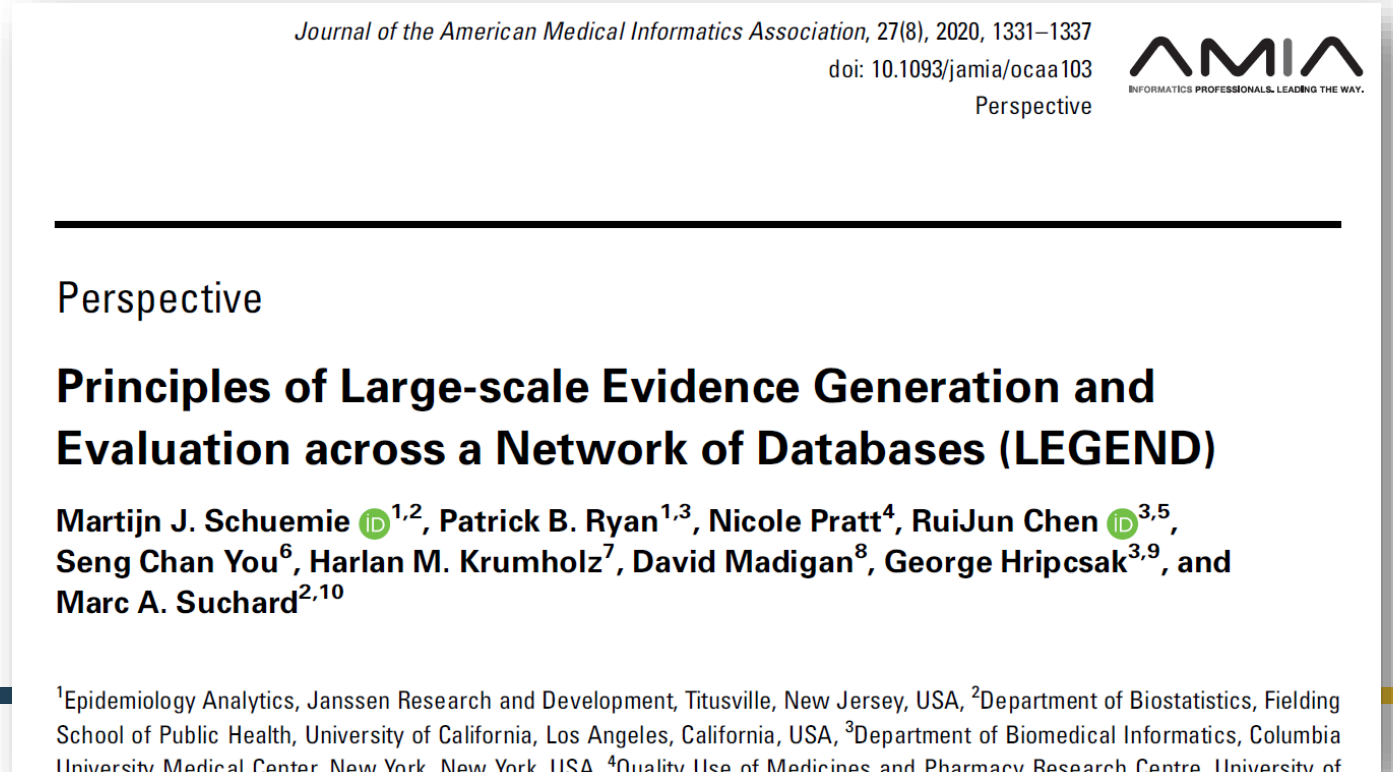
Department of Biostatistics,

University of California, Los Angeles

LEGEND

LARGE-SCALE EVIDENCE GENERATION AND EVALUATION IN A NETWORK OF DATABASES

- OHDSI's LEGEND aims to generate reliable evidence by following a set of principles that address
  - Observational study bias (e.g. confounding)
  - Publication bias
  - P-hacking



Journal of the American Medical Informatics Association, 27(8), 2020, 1331–1337
doi: 10.1093/jamia/ocaa103
Perspective

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Perspective

Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND)

Martijn J. Schuemie[1,2], Patrick B. Ryan[1,3], Nicole Pratt[4], RuiJun Chen[3,5], Seng Chan You[6], Harlan M. Krumholz[7], David Madigan[8], George Hripcsak[3,9], and Marc A. Suchard[2,10]

[1]Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA, [2]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California, USA, [3]Department of Biomedical Informatics, Columbia University Medical Center, New York, New York, USA, [4]Quality Use of Medicines and Pharmacy Research Centre, University of

LEGEND Guiding Principles

1. LEGEND will generate evidence at a **large scale.**
2. **Dissemination** of the evidence will not depend on the estimated effects.
3. LEGEND will generate evidence using a **prespecified analysis design**.
4. LEGEND will generate evidence by consistently applying a **systematic process** across all research questions.
5. LEGEND will generate evidence using **best practices**.
6. LEGEND will include empirical evaluation through the use of **control questions**.
7. LEGEND will generate evidence using **open-source software** that is freely available to all.
8. LEGEND will **not** be used to **evaluate new methods**.
9. LEGEND will generate evidence across a network of **multiple databases**.
10. LEGEND will **maintain data confidentiality**; patient-level data will not be shared between sites in the network.

LEGEND Guiding Principles

1. LEGEND will generate evidence at a **large scale.**
2. **Dissemination** of the evidence will not depend on the estimated effects.
3. LEGEND will generate evidence using a **prespecified analysis design**.
4. LEGEND will generate evidence by consistently applying a **systematic process** across all research questions.
5. LEGEND will generate evidence using **best practices**.
6. LEGEND will include empirical evaluation through the use of **control questions**.
7. LEGEND will generate evidence using **open-source software** that is freely available to all.
8. LEGEND will **not** be used to **evaluate new methods**.
9. LEGEND will generate evidence across a network of **multiple databases**.
10. LEGEND will **maintain data confidentiality**; patient-level data will not be shared between sites in the network.

LEGEND Guiding Principles

1. LEGEND will generate evidence at a **large scale.**
2. **Dissemination** of the evidence will not depend on the estimated effects.
3. LEGEND will generate evidence using a **prespecified analysis design**.
4. LEGEND will generate evidence by consistently applying a **systematic process** across all research questions.
5. LEGEND will generate evidence using **best practices**.
6. LEGEND will include empirical evaluation through the use of **control questions**.
7. LEGEND will generate evidence using **open-source software** that is freely available to all.
8. LEGEND will **not** be used to **evaluate new methods**.
9. LEGEND will generate evidence across a network of **multiple databases**.
10. LEGEND will **maintain data confidentiality**; patient-level data will not be shared between sites in the network.

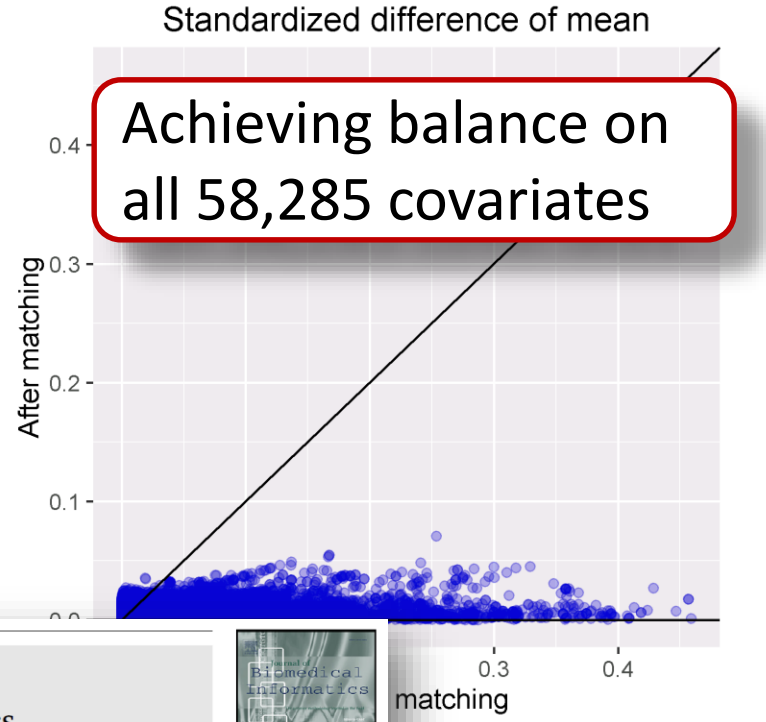
LEGEND Guiding Principles

1. LEGEND will generate evidence at a **large scale.**
2. **Dissemination** of the evidence will not depend on the estimated effects.
3. LEGEND will generate evidence using a **prespecified analysis design**.
4. LEGEND will generate evidence by consistently applying a **systematic process** across all research questions.
5. LEGEND will generate evidence using **best practices**.
6. LEGEND will include empirical evaluation through the use of **control questions**.
7. LEGEND will generate evidence using **open-source software** that is freely available to all.
8. LEGEND will **not** be used to **evaluate new methods**.
9. LEGEND will generate evidence across a network of **multiple databases**.
10. LEGEND will **maintain data confidentiality**; patient-level data will not be shared between sites in the network.

Best practice for addressing confounding

Large-Scale Propensity Scores (LSPS)

- Construct large generic set of covariates
  - 10,000 < n < 100,000
- Use regularized regression to fit propensity model
- Match or stratify on propensity score



International Journal of Epidemiology, 20
doi: 10.1093/ij
Origina

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 Cali
Los Angeles, CA, USA, [2]Epidemiology Department, Janssen Research and Development LLC, Titu
NJ, USA, [3]Department of Biostatistics, UCLA Fielding School of Public Health, University of Cali
Los Angeles, CA, USA and [4]Department of Human Genetics, David Geffen School of Medic
UCLA, University of California, Los Angeles, CA, USA

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Adjusting for indirectly measured confounding using large-scale propensity score

Linying Zhang [a], Yixin Wang [b], Martijn J. Schuemie [c], David M. Blei [d,e], George Hripcsak [a,f,*]

[a] Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W. 168th Street, PH20, New York, 10032, NY, USA
[b] Department of Statistics, University of Michigan, 1085 S University Ave, Ann Arbor, 48109, MI, USA
[c] Janssen Research and Development, 1125 Trenton-Harbourton Road, Titusville, 08560, NJ, USA
[d] Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, 10027, NY, USA
[e] Department of Computer Science, Columbia University, 500 West 120 Street, Room 450 MC0401, New York, 10027, NY, USA
[f] Medical Informatics Services, New York-Presbyterian Hospital, 622 W. 168th Street, PH20, New York, 10032, NY, USA

ARTICLE INFO

Keywords:

ABSTRACT

Confounding remains one of the major challenges to causal inference with observational data. This problem

Measuring residual systematic error

Control questions:

- exposure-outcome pairs with known effect size
- negative (and positive) controls

Empirical calibration:

- Adjust p-value and confidence interval using estimates for controls

Research Article

Statistics in Medicine

Received 12 November 2012, Accepted 3 July 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5925

Interpreting observational studies: why empirical calibration is needed to correct *p*-values

Martijn J. Schuemie,[a,b*†] Patrick B. Ryan,[b,c] William DuMouchel,[b,d] Marc A. Suchard[b,e] and David Madigan[b,f]

Often the literature makes assertions of medical product effects on the basis of '$p < 0.05$'. The underlying

PNAS

COLLOQUIUM PAPER

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California, Los Angeles, CA 90095

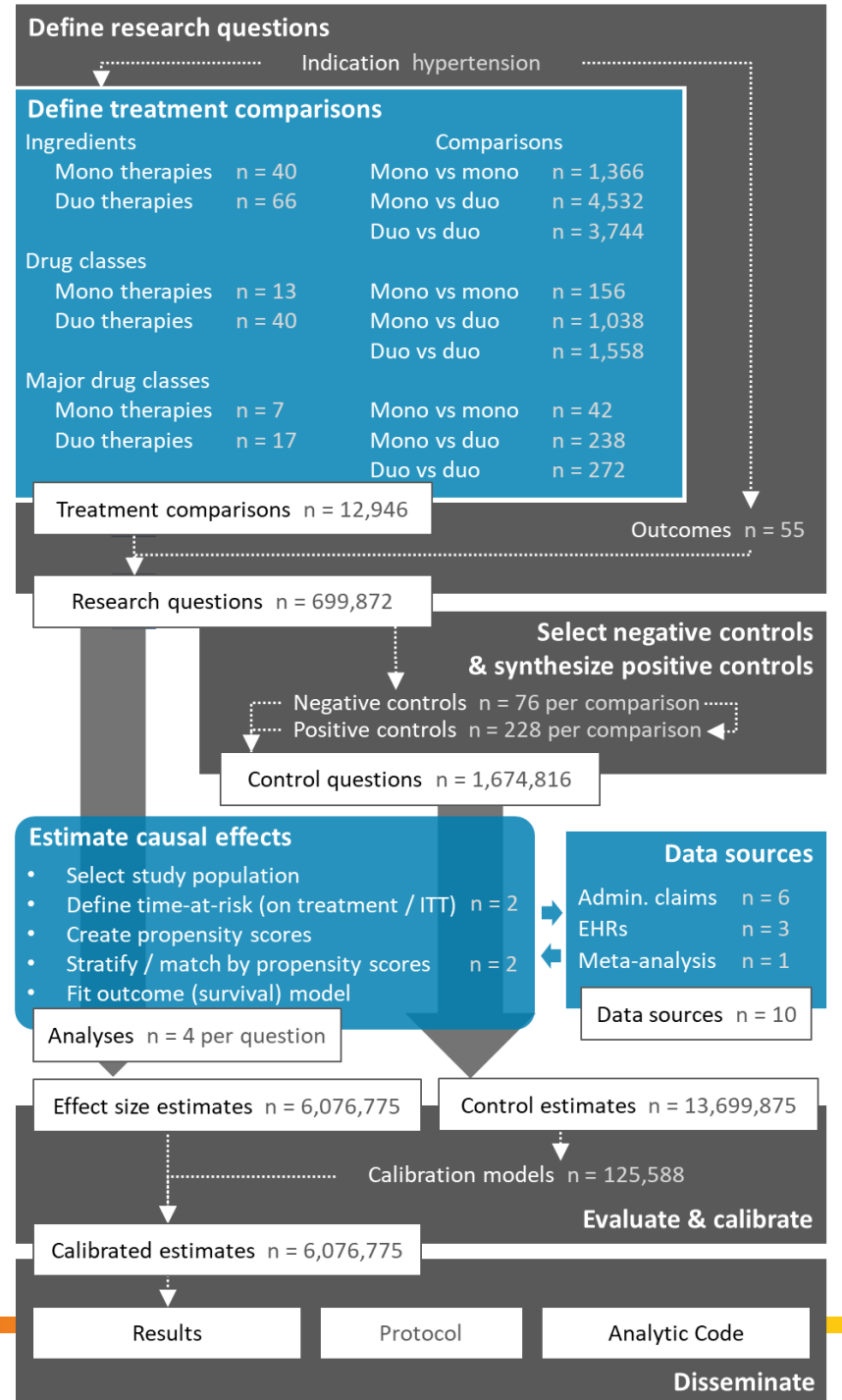
Edited by Victoria Stodden, University of Illinois at Urbana–Champaign, Champaign, IL, and accepted by Editorial Board Member Susan T. Fiske October 26, 2017 (received for review June 15, 2017)

Observational healthcare data, such as electronic health records and administrative claims, offer potential to estimate effects of medical products at scale. Observational studies have often been found to be nonreproducible, however, generating conflicting results even when using the same database to answer the

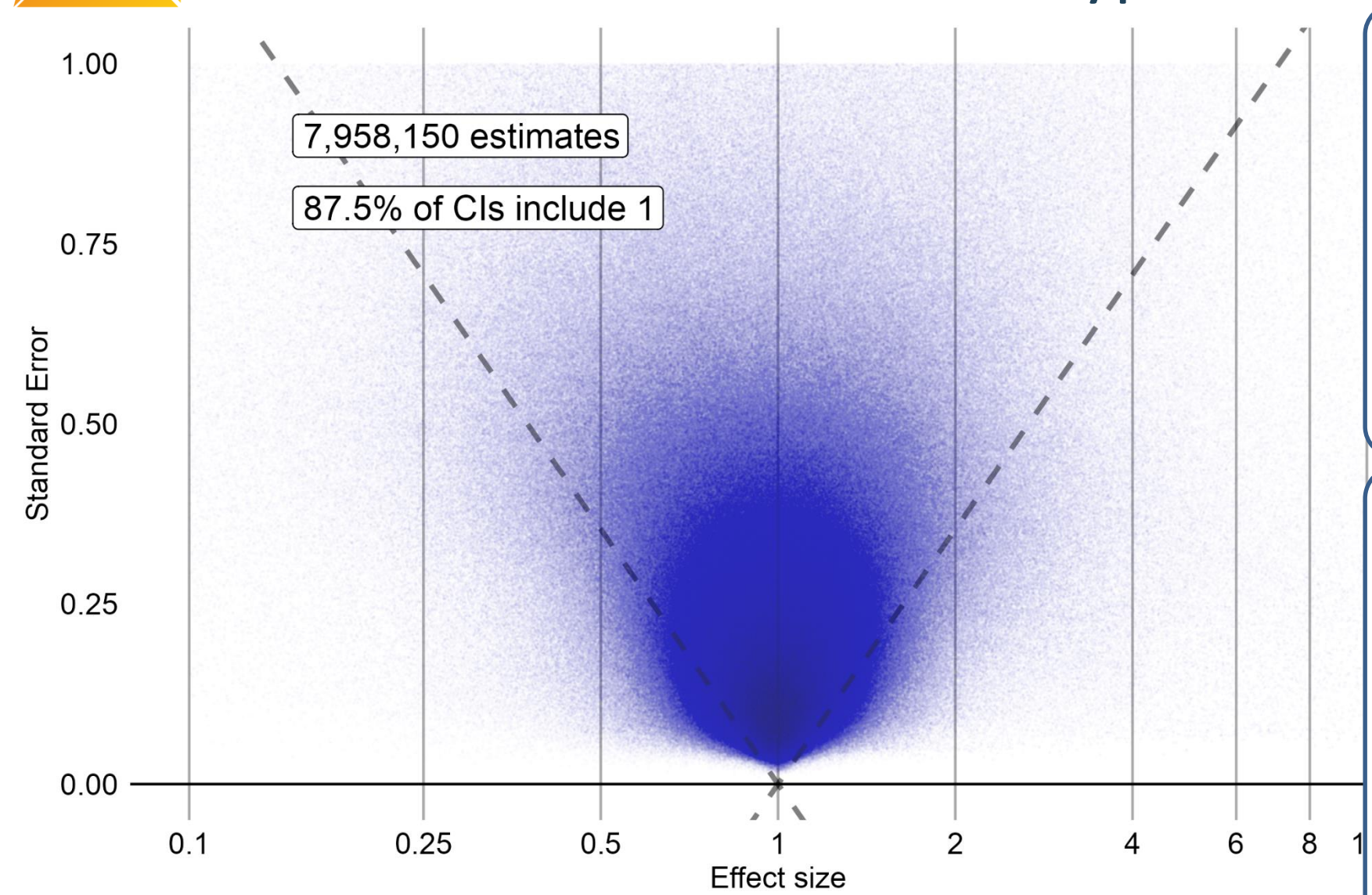
age treatment effect. Systematic error can manifest from multiple sources, including confounding, selection bias, and measurement error. While there is widespread awareness of the potential for systematic error in observational studies and a large body of research that examines how to diagnose and statistically adjust

LEGEND Hypertension Study

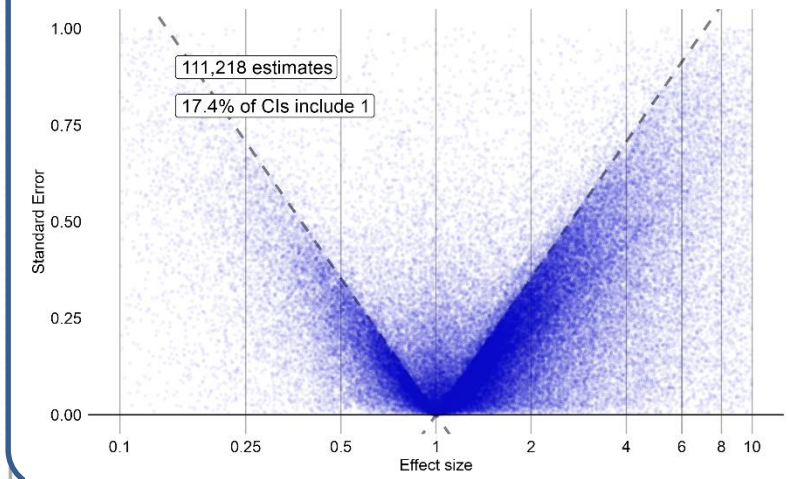
- Comparing all first-line hypertension treatments
  - Mono-therapy
  - Dual-therapy
- Including 55 outcomes of interest
- Including negative control outcomes
- Confounder adjustment using large-scale PS
- Across a global network of 9 databases



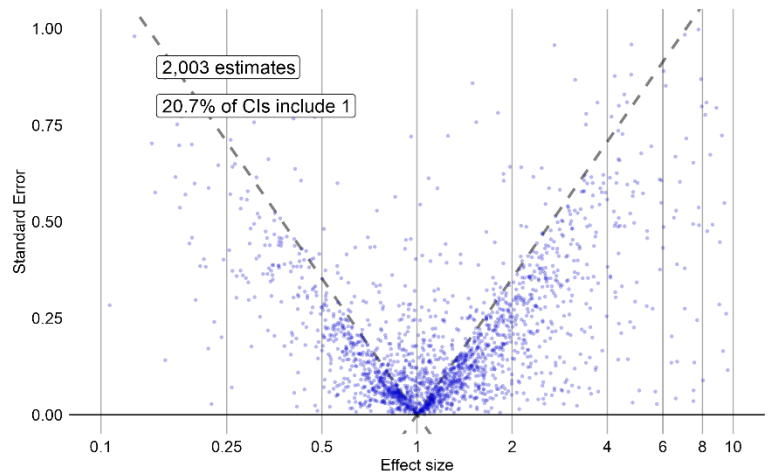
Distribution of estimates from LEGEND Hypertension



All published observational research



All published observational research on hypertension treatments



Several high-impact LEGEND Hypertension papers

THE LANCET

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, la

Marc A Suchard, Martijn J Schuemie, George Hripcsak, Patrick B Ryan

Summary

Background Uncertainty rema mending any primary agent ar enzyme inhibitors, angiotensi calcium channel blockers, in choice.

Methods We developed a con and safety evaluation across r while minimising inherent b cohort design to estimate the

Hypertension

ANTIHYPERTENSIVE TREATMENT

Comparative First-Line Effectiveness and Saf of ACE (Angiotensin-Converting Enzyme) Inhibitors and Angiotensin Receptor Blockers

A Multinational Cohort Study

Hypertension

BETA-BLOCKER THERAPY

Comprehensive Comparative Effectiveness and Safety of First-Line β-Blocker Monotherapy in Hypertensive Patients

JAMA Internal Medicine | Original Investigation

Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension

George Hripcsak, MD, MS; Marc A. Suchard, MD, PhD; Steven Shea, MD; RuiJun Chen, MD; Seng Chan You, MD; Nicole Pratt, PhD; David Madigan, PhD; Harlan M. Krumholz, MD, SM; Patrick B. Ryan, PhD; Martijn J. Schuemie, PhD

Journal of the American Medical Informatics Association, 27(8), 2020, 1268–1277
doi: 10.1093/jamia/ocaa124
Research and Applications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.

Research and Applications

Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie,[1,2] Patrick B Ryan,[1,3] Nicole Pratt,[4] RuiJun Chen,[3,5] Seng Chan You,[6] Harlan M Krumholz,[7] David Madigan,[8] George Hripcsak,[3,9] and Marc A Suchard[2,10]

Diagnostics

- Each LEGEND estimate comes with full diagnostics, e.g.
  - Statistical power
  - Covariate balance
  - Systematic error as observed through negative controls
- In each LEGEND paper we decided whether we were satisfied with the diagnostics for those results
- Can we make the use of diagnostics more systematic?

Objective diagnostics in observational research

Patrick B. Ryan PhD

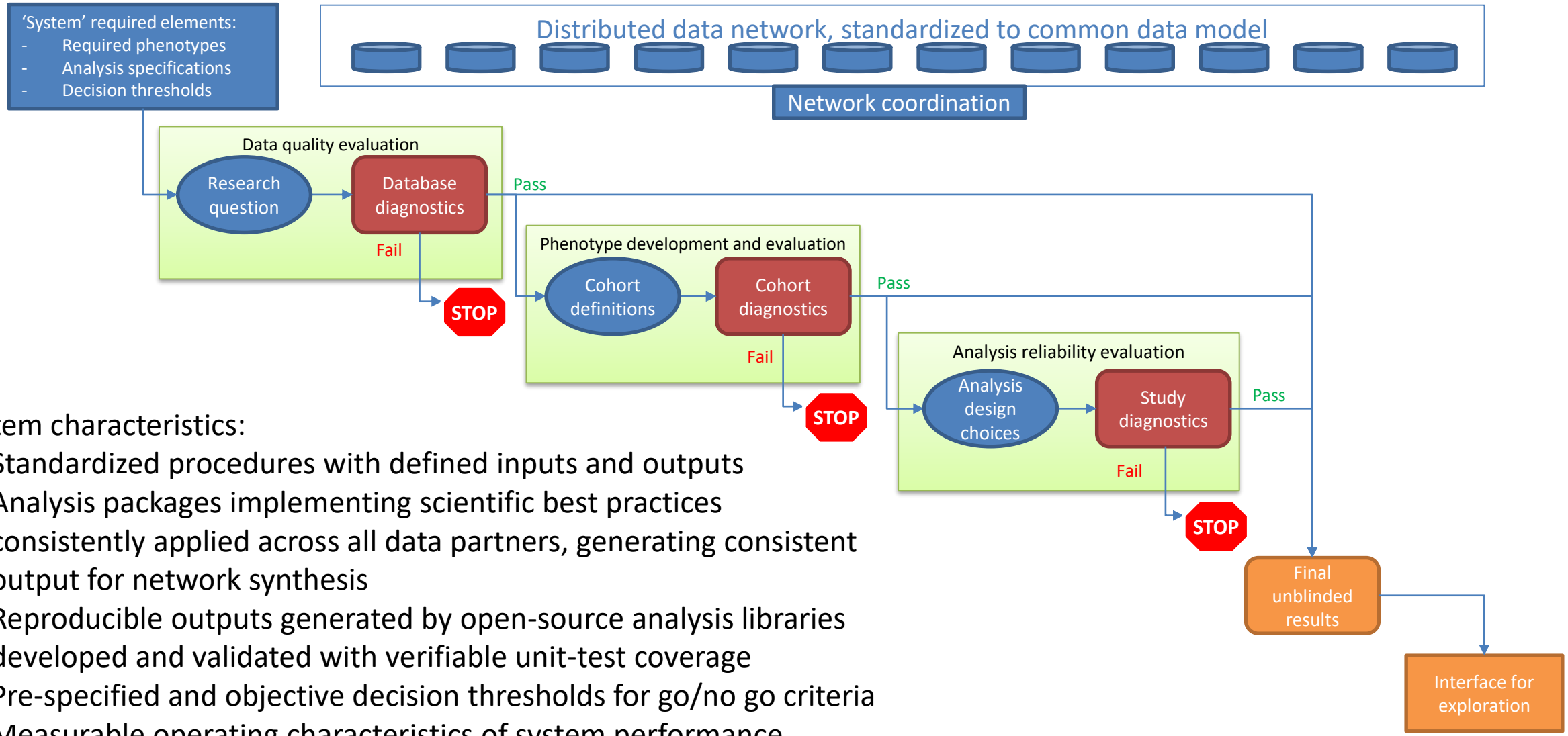
Observational Health Data Analytics

Johnson & Johnson

Department of Biomedical Informatics,

Columbia University Medical Center

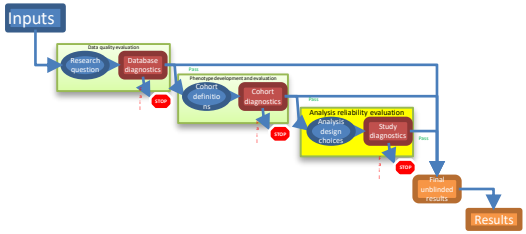
Engineering open science systems that build trust into the real-world evidence generation and dissemination process



System characteristics:

- Standardized procedures with defined inputs and outputs
- Analysis packages implementing scientific best practices consistently applied across all data partners, generating consistent output for network synthesis
- Reproducible outputs generated by open-source analysis libraries developed and validated with verifiable unit-test coverage
- Pre-specified and objective decision thresholds for go/no go criteria
- Measurable operating characteristics of system performance

Study diagnostics

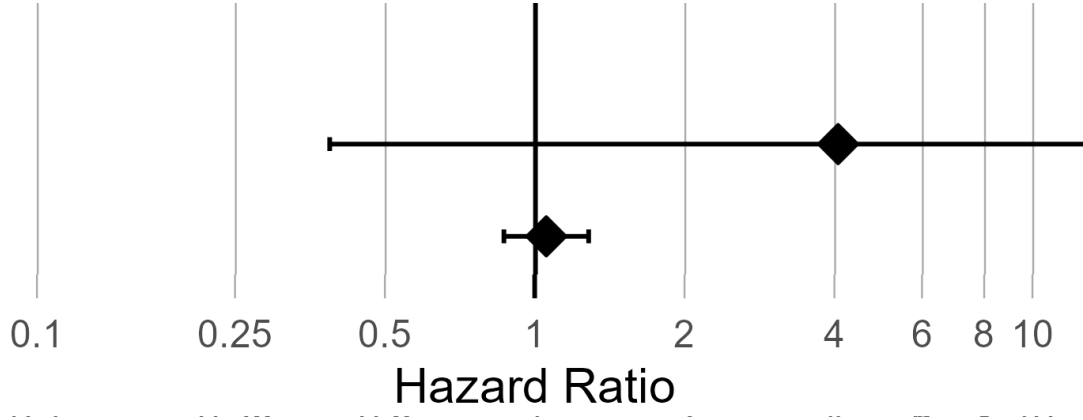


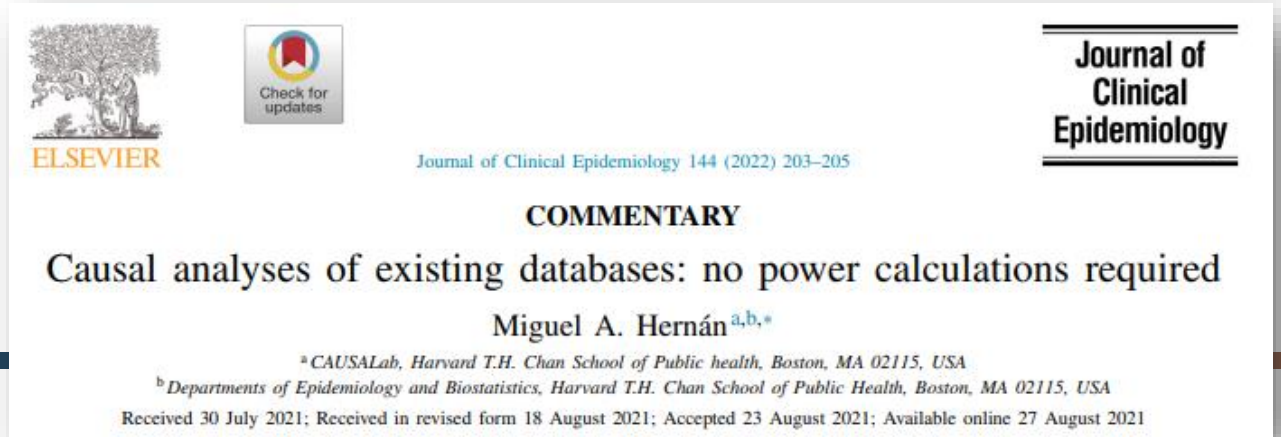
- Challenge: Analyses risk producing misleading estimates due to study design and analytical choices and their application to data.
- Opportunity: Provide objective criteria with pre-specified decision thresholds for evaluating the reliability of analyses with respect to precision, accuracy, and generalizability within each database across a network

Statistical power

- In observational studies, sample size (power) is fixed
  - compute minimum detectable risk (MDRR) given power
- More power is better, but low-power studies can still inform
- However, too little power can confuse

Hazard Ratio (95% CI)
4.06 (0.39 - 42.60)
1.05 (0.86 - 1.28)





ELSEVIER

Check for updates

Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 144 (2022) 203–205

COMMENTARY

Causal analyses of existing databases: no power calculations required

Miguel A. Hernán[a,b,*]

[a] CAUSALab, Harvard T.H. Chan School of Public health, Boston, MA 02115, USA

[b] Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Received 30 July 2021; Received in revised form 18 August 2021; Accepted 23 August 2021; Available online 27 August 2021

Statistical power

Rule: **Minimum Detectable Relative Risk (MDRR) < 10**

Reasoning:

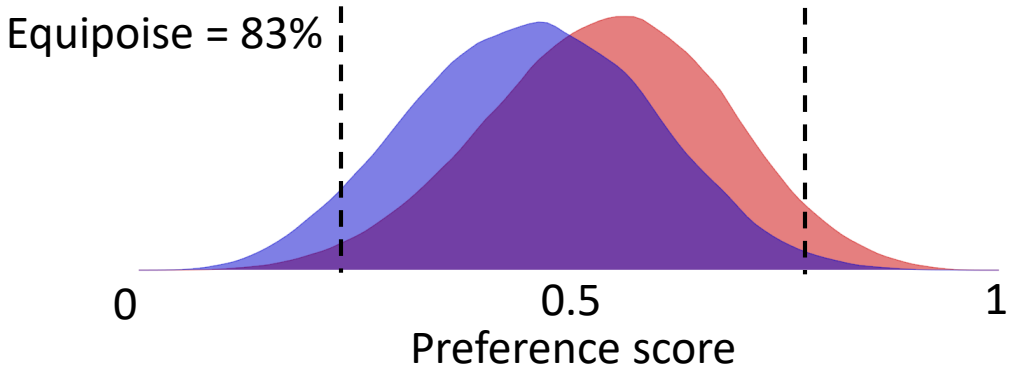
Even low-power estimate (wide CI) could be helpful, but we want to avoid misinterpreting grossly underpowered studies

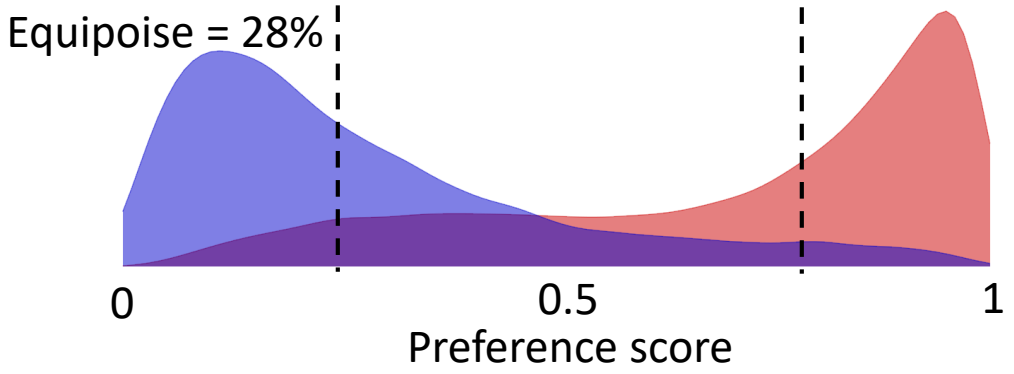
Note:

In LEGEND Hypertension, we required exposure cohorts > 2,500 subjects, so already eliminated most underpowered estimates.

Equipoise

- Randomized clinical trials: each subject has same probability of having each intervention
  - Randomization allows for assumption of exchangeability
- Non-interventional studies: observed (non-random) treatment choices
  - Preference = probability of patient choosing target vs. comparator treatment, given baseline features
- Potential **pre-adjustment design diagnostic**: what proportion of the population has preference close to 0.5 (0.3-0.7)?





Equipoise

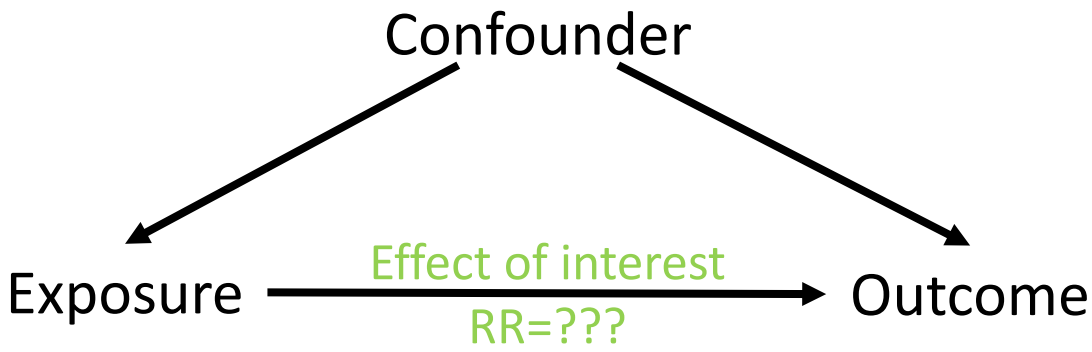
Rule: **Equipoise > 0.5**

(Equipoise is percent of population that has 0.3 < preference score < 0.7)

Reasoning:

If equipoise is low, the populations are too incomparable, and we probably shouldn't trust our ability to make them comparable.

Covariate balance



- Confounding variables associated with both exposure and outcome can bias effect estimates if not properly addressed
- Various strategies (e.g. PS matching) can reduce confounding by balancing confounder prevalence in target and comparator cohort
- Potential **post-adjustment analytic diagnostic**: are observed baseline characteristics sufficiently similar between **target** and **comparator** cohorts?

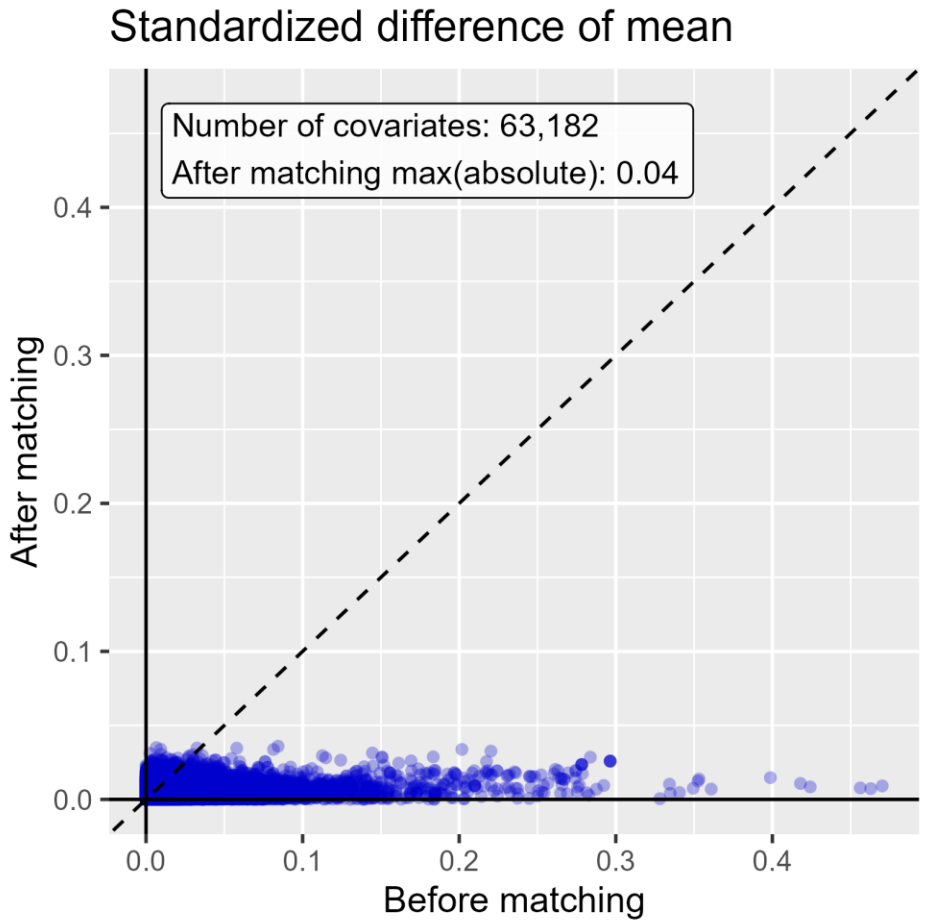
Covariate balance

Rule: **Max standardized difference of mean (SDM) < 0.1**

(no covariate may have a SDM >= 0.1 after PS adjustment)

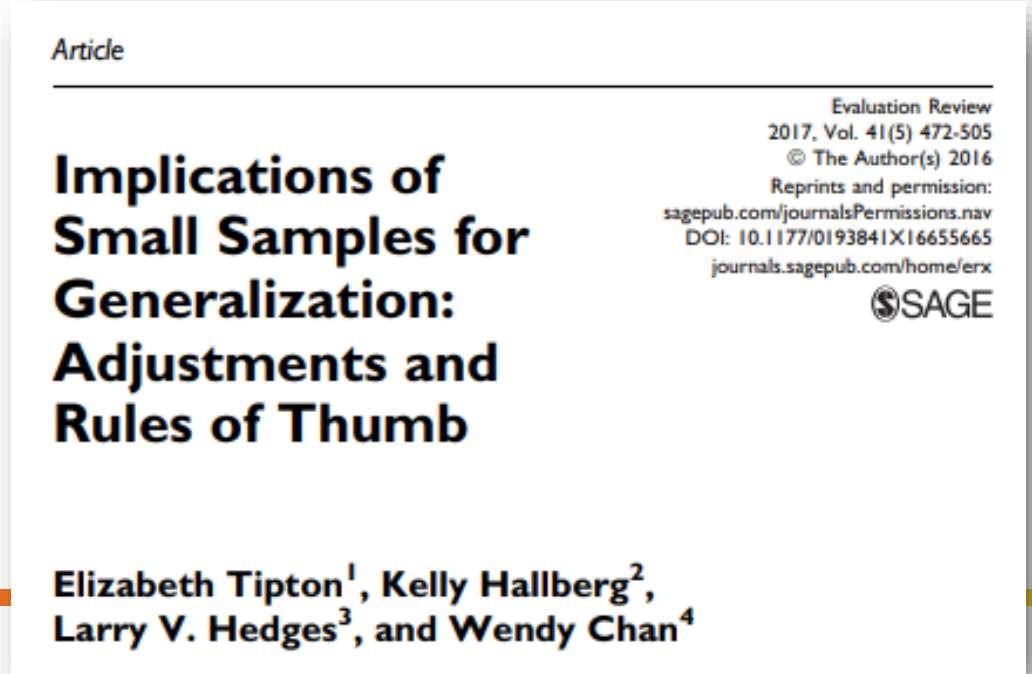
Reasoning:

If covariates are unbalanced there may be confounding.



Generalizability

- Generalizability is the extent to which a study result can be applied to a target population of interest
- Strategies employed to reduce confounding (e.g. PS matching) can shift the composition of the analytic cohort away from the original target
- Potential **post-adjustment analytic diagnostic**: are observed baseline characteristics sufficiently similar between the **pre-adjustment target** and **post-adjustment analytic cohorts**?



Article

Evaluation Review
2017, Vol. 41(5) 472-505
© The Author(s) 2016
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0193841X16655665
journals.sagepub.com/home/erx
SAGE

Implications of Small Samples for Generalization: Adjustments and Rules of Thumb

Elizabeth Tipton[1], Kelly Hallberg[2], Larry V. Hedges[3], and Wendy Chan[4]

Generalizability

Rule: **Max SDM between *analytic* cohort and *target* cohort < 0.25**

(target cohort: the cohort we started with (those exposed))

(analytic cohort: the cohort after all adjustments)

Reasoning:

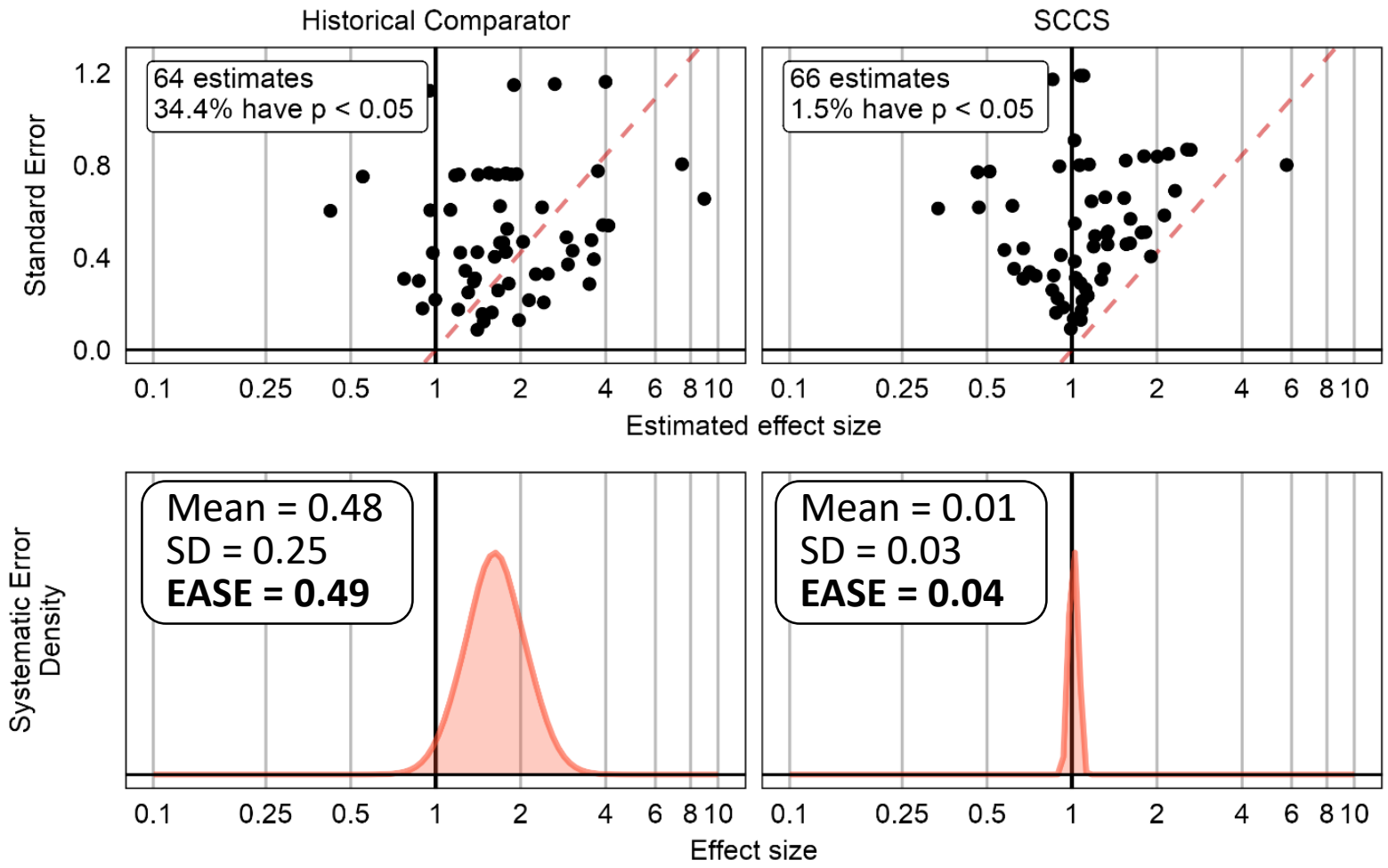
Estimate may not generalize to our target population if differences are too great.

Residual systematic error

- Design and analysis choices aim to produce unbiased estimates, but residual systematic error can still exist due to model misspecification inherent to analysis or data
- Bias – expected value of systematic error – can be estimated using negative control experiments in which estimates can be compared with known truth
- Potential **post-adjustment analytic diagnostic**: is the residual bias observed from negative controls small enough to accept that calibrated effect estimates can be trusted as unbiased?

Residual systematic error: EASE statistic

- Estimates for negative controls outcomes can be used to fit a (normal) systematic error distribution
- The Expected Absolute Systematic Error (EASE) summarizes this distribution
- EASE = 0 means all variation in negative control estimates can be explained by random error alone.



Residual systematic error

Rule: **Expected Absolute Systematic Error (EASE) < 0.25**

(EASE is the expected abs(log(estimated RR) – log(true RR)), based on negative control estimates)

Reasoning:

Even though we can and should empirically calibrate to account for residual error, readers may not trust results if calibration shifts the estimates too much.

Note:

Our evaluation uses calibrated estimates, which already incorporates the systematic error observed for the original set of negative controls.

Concluding thoughts

- Diagnostics can provide evidence to build trust in the results of our studies, but…
  - Post-hoc interpretation allows for investigator bias
  - Current decision thresholds are based on asserted expert opinions and arbitrary rules of thumb

- How can we develop empirical evidence to set objective decision thresholds and allow pre-specification of diagnostics to increase trust and improve the reliability of our studies?

Evaluation of objective diagnostics

Yong Chen PhD

Department of Biostatistics, Epidemiology and Informatics (DBEI), the Perelman School of Medicine, University of Pennsylvania

Evaluation of objective diagnostics

Diagnostic with cutoffs	Fraction failing	Objective function: EASE on new negative controls
Power (MDRR < 10)	?	?
Equipoise (>50%)	?	?
Covariate balance (max SDM < 0.10)	?	?
Generalizability (max SDM < 0.25)	?	?
All diagnostics	?	?

LEGEND estimates where no effect is expected

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, Patrick B Ryan

- Product labels tend to be inclusive for adverse reactions: High sensitivity
- Conservative approach:

  For the list of outcomes in LEGEND

  When comparing two drugs, if neither target nor comparator drug has the outcome on the label, and no other drug in the same classes have the outcome on the label

  Then both drugs likely don't cause the outcome, and the hazard ratio is likely to be "close to 1".

LEGEND estimates where no effect is expected

- ACE inhibitors like lisinopril have angioedema on their label.
- Calcium channel blockers and ARBs are not believed to have this side effect, but still list in 'Postmarketing experience'.
- None of the direct vasodilators and alpha-1 blockers have angioedema on their label.

5 WARNINGS AND PRECAUTIONS

5.1 Fetal Toxicity

Lisinopril can cause fetal harm when administered to a
renin-angiotensin system during the second and third t
function and increases fetal and neonatal morbidity an
associated with fetal lung hypoplasia and skeletal defo
include skull hypoplasia, anuria, hypotension, renal failu
discontinue lisinopril as soon as possible *[see Use in sp*

5.2 Angioedema and Anaphylactoid Reactions

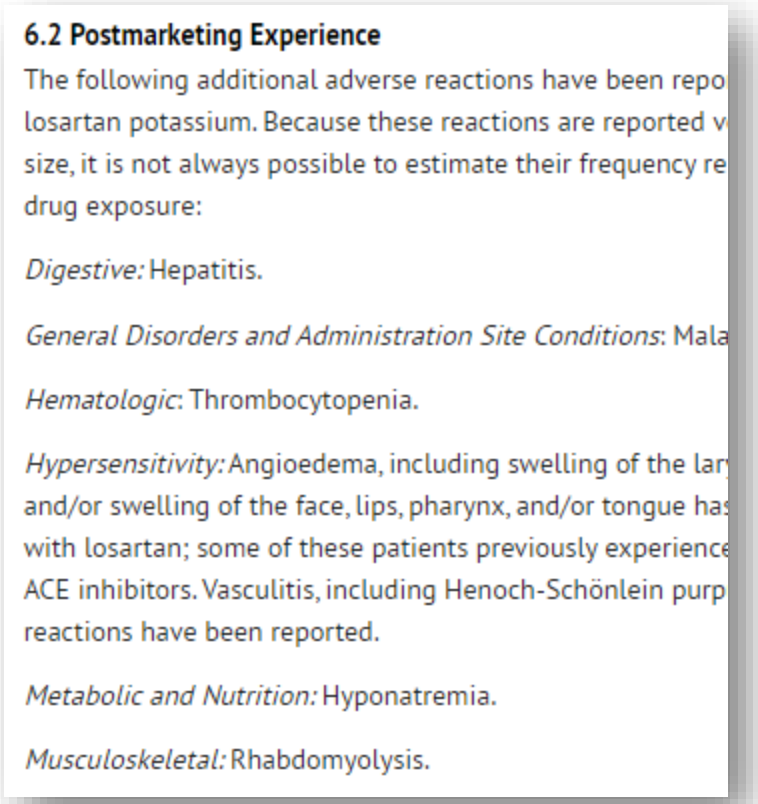
Patients taking concomitant mTOR inhibitor (e.g. temsir
neprilysin inhibitor may be at increased risk for angioed

Hydralazine (vasodilator) vs **prazosin** (alpha blocker) for **angioedema** is **likely null** (i.e., hazard ratio being "close to 1").

LEGEND estimates where no effect is expected

- ARBs like losartan have **rhabdomyolysis** listed as an adverse event
- Beta-blockers and loop diuretics do not

metoprolol (beta-blocker) vs **furosemide** (loop diuretic) for **rhabdomyolysis** is **likely null**



6.2 Postmarketing Experience

The following additional adverse reactions have been repo
losartan potassium. Because these reactions are reported v
size, it is not always possible to estimate their frequency re
drug exposure:

Digestive: Hepatitis.

General Disorders and Administration Site Conditions: Mala

Hematologic: Thrombocytopenia.

Hypersensitivity: Angioedema, including swelling of the lar
and/or swelling of the face, lips, pharynx, and/or tongue has
with losartan; some of these patients previously experience
ACE inhibitors. Vasculitis, including Henoch-Schönlein purp
reactions have been reported.

Metabolic and Nutrition: Hyponatremia.

Musculoskeletal: Rhabdomyolysis.

LEGEND estimates where no effect is expected

- 9,752 target-comparator-outcomes are likely null (2 x 4,876)
- A new set of (imperfect) negative controls (null may not be true)
- Difference with negative controls used in LEGEND:
  - Will use these across all analyses to evaluate overall distribution
  - Outcomes being more similar to the outcomes of interest: better exchangeability?
  - Using full outcome phenotypes instead of ‘occurrence of concept’
- Expected Absolute Systematic Error (EASE) on new set of negative controls: 0.38

Evaluating one diagnostic at a time

Diagnostic	Fraction failing	EASE on new negative controls*
Power (MDRR < 10)	5%	0.38
Equipoise (>50%)	71%	0.02
Covariate balance (max SDM < 0.10)	56%	0.28
Generalizability (max SDM < 0.25)	57%	0.37
Systematic error (EASE < 0.25)	16%	0.35

* Without filtering by any diagnostic, EASE = 0.38

Evaluating all diagnostic together

Diagnostic	Fraction failing	EASE on new negative controls*
Power (MDRR < 10)	5%	0.38
Equipoise (>50%)	71%	0.02
Covariate balance (max SDM < 0.10)	56%	0.28
Generalizability (max SDM < 0.25)	57%	0.37
Systematic error (EASE < 0.25)	16%	0.35
All diagnostics	88%	0.00

* Without filtering by any diagnostic, EASE = 0.38

Can we do better?

- Current thresholds are arbitrary
- Can we do any better?

Rules as an optimization problem

- We have an ‘objective’ optimization criterion:
  - Maximize number of remaining estimates
  - Under constraint of low residual bias as measured on new negative controls (EASE< 0.05)
- What set of thresholds is optimal?

‘Optimal’ thresholds

Diagnostic	Literature-derived threshold	Data-driven threshold
Statistical power (MDRR)	10	-
Equipoise	0.50	0.50
Covariate balance (SDM)	0.10	0.50
Generalizability (SDM)	0.25	-
Systematic error (EASE)	0.25	-
Fraction failing	88%	71%
EASE	0.00	0.02

Some conclusions

Marc A. Suchard, MD, PhD

Department of Biostatistics, University of California, Los Angeles

Wrapping up the evaluation of diagnostics

- Some of our study diagnostics can help improve the ***reliability*** of the evidence, as measured as systematic error.
- Other diagnostics have different goals, such as improved ***interpretability*** and ***generalizability***.
- Up to now, diagnostics rules were arbitrary.
- Our empirical evaluation provides evidence for choices of thresholds, under various constraints.

Strong advice: Pre-specify diagnostic rules

- Post-hoc interpretation of diagnostics allows for investigator bias (p-hacking).
- Diagnostics rules should be pre-specified, for example in the protocol.

SAMPLE SIZE AND STUDY POWER

Within each data source, we will execute all comparisons with ≥1000 eligible patients per arm. Blinded to effect estimates, investigators and stakeholders will evaluate extensive study diagnostics for each comparison to assess reliability and generalisability, and only report risk estimates that pass.[25 35] These diagnostics will include:

1. Minimum detectable risk ratio as a typical proxy for power.
2. Preference score distributions to evaluate empirical equipoise10 and population generalisability.
3. Extensive patient characteristics to evaluate cohort balance before and after PS adjustment.
4. Negative control calibration plots to assess residual bias.
5. Kaplan-Meier plots to examine HR proportionality assumptions.

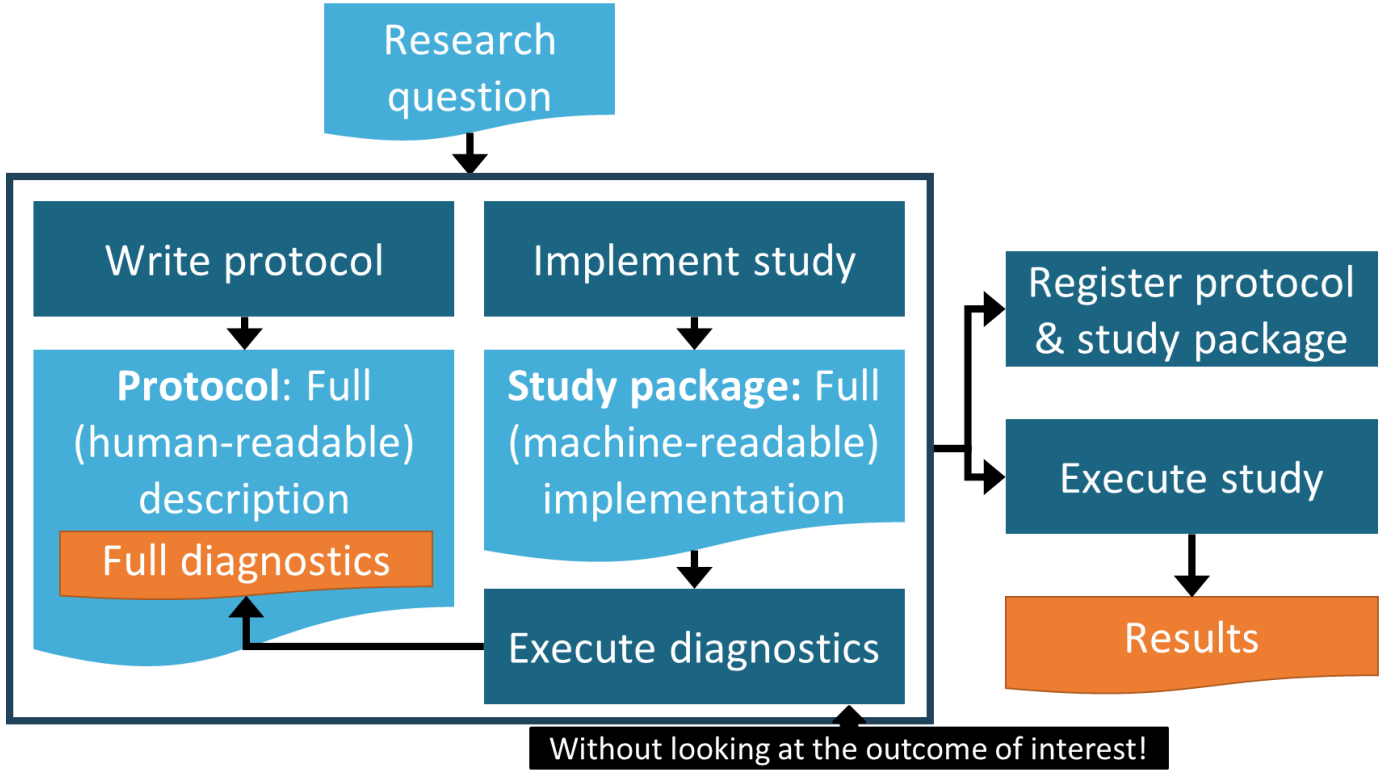
We will define cohorts to stand in empirical equipoise if the majority of patients carry preference scores between 0.3 and 0.7 and to achieve balance if all after-adjustment characteristics return absolute standardised mean differences <0.1.[118]

Khera, et al, *BMJ Open*, 2022

Avoid investigator bias when interpreting diagnostics

- Diagnostics need to be evaluated prior to looking at the study results – 2 approaches:

1. Protocol can contain **diagnostics results**, or
2. Protocol can contain **prespecified diagnostics rules** (so long as they are not modified *post-hoc*)



Pre-specification of a systematic approach