Objective Diagnostics:
A pathway to provably reliable evidence

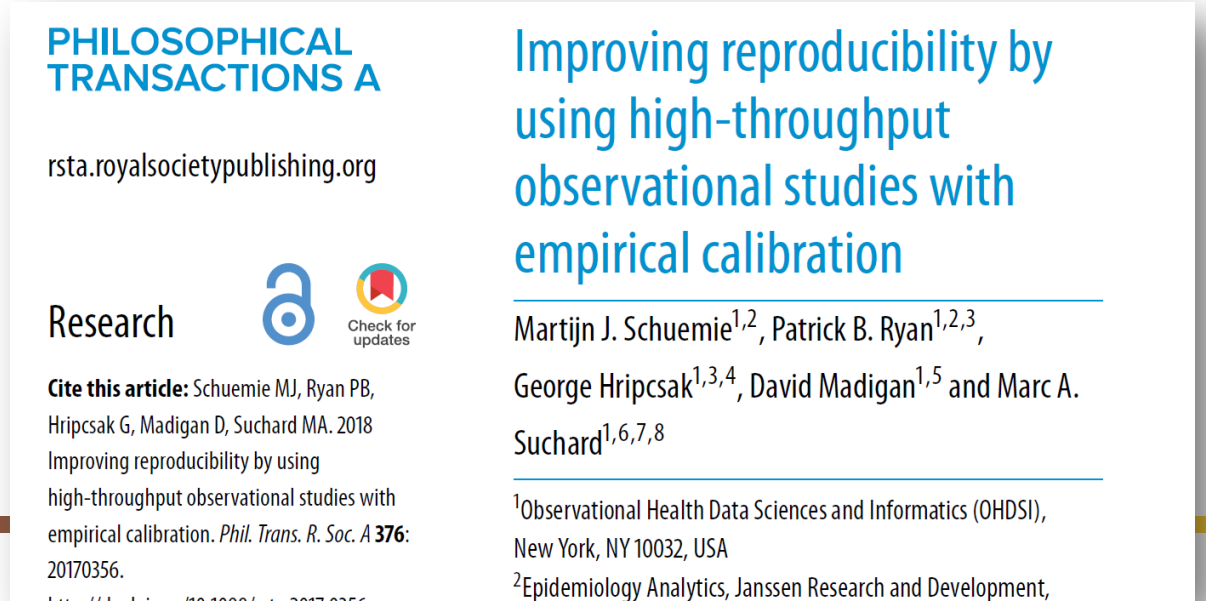
Martijn Schuemie
Observational Health Data Analytics, Johnson & Johnson
Biostatistics, UCLA

Objective Diagnostics:
How to avoid generating unreliable evidence

Martijn Schuemie
Observational Health Data Analytics, Johnson & Johnson
Biostatistics, UCLA

Issues with observational research

- Reproducibility crisis
- Lack of trust in real-world evidence
- Major issues:
  - Observational study bias (e.g. confounding)
  - Publication bias
  - P-hacking



PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Check for updates

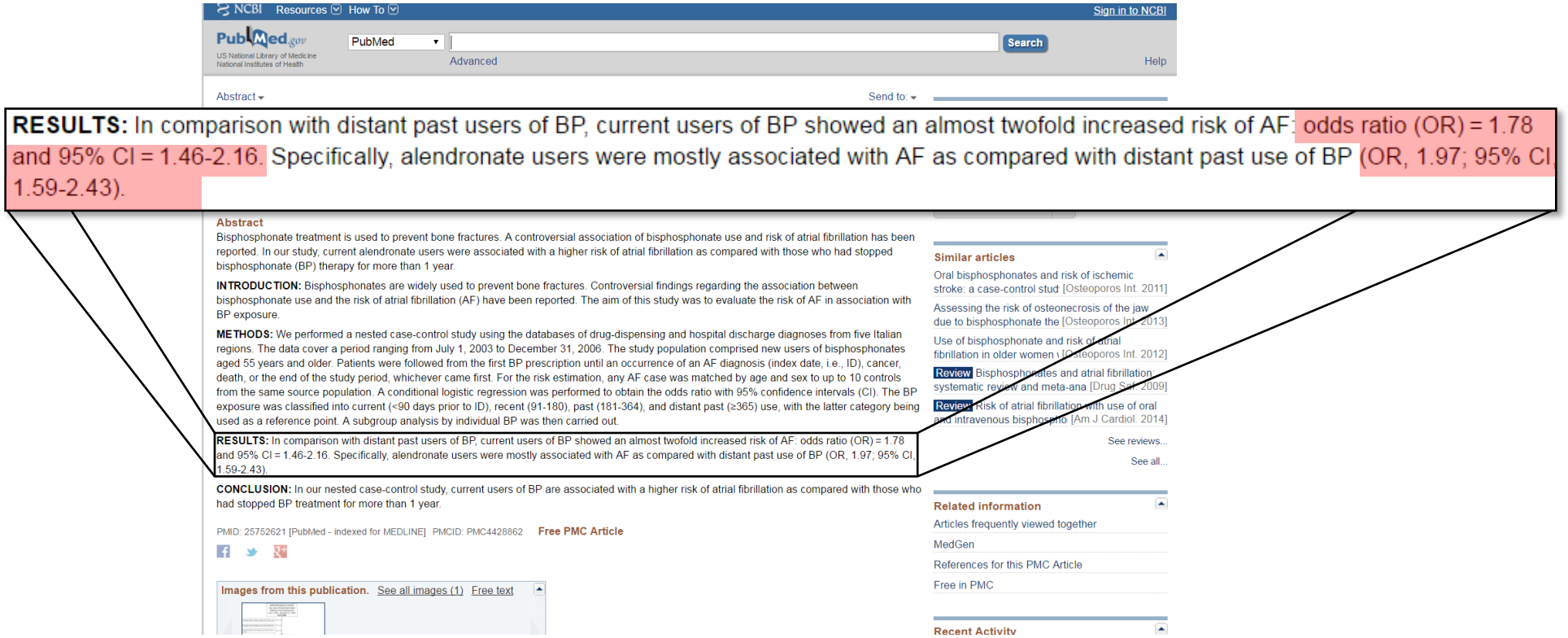
Cite this article: Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. 2018 Improving reproducibility by using high-throughput observational studies with empirical calibration. *Phil. Trans. R. Soc. A* **376**: 20170356.

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

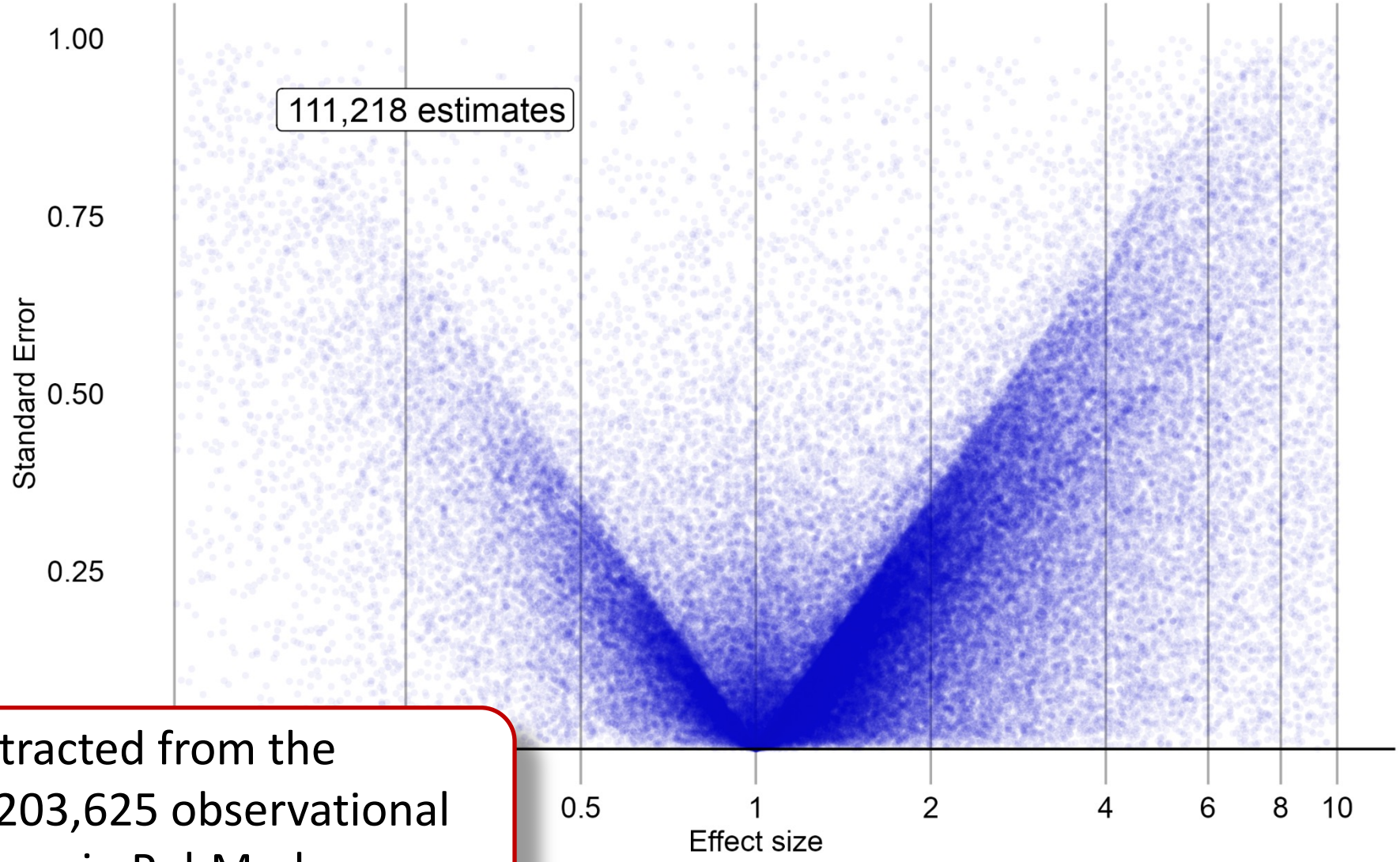
Martijn J. Schuemie[1,2], Patrick B. Ryan[1,2,3], George Hripcsak[1,3,4], David Madigan[1,5] and Marc A. Suchard[1,6,7,8]

[1]Observational Health Data Sciences and Informatics (OHDSI), New York, NY 10032, USA
[2]Epidemiology Analytics, Janssen Research and Development,

Automated extraction of effect sizes from literature

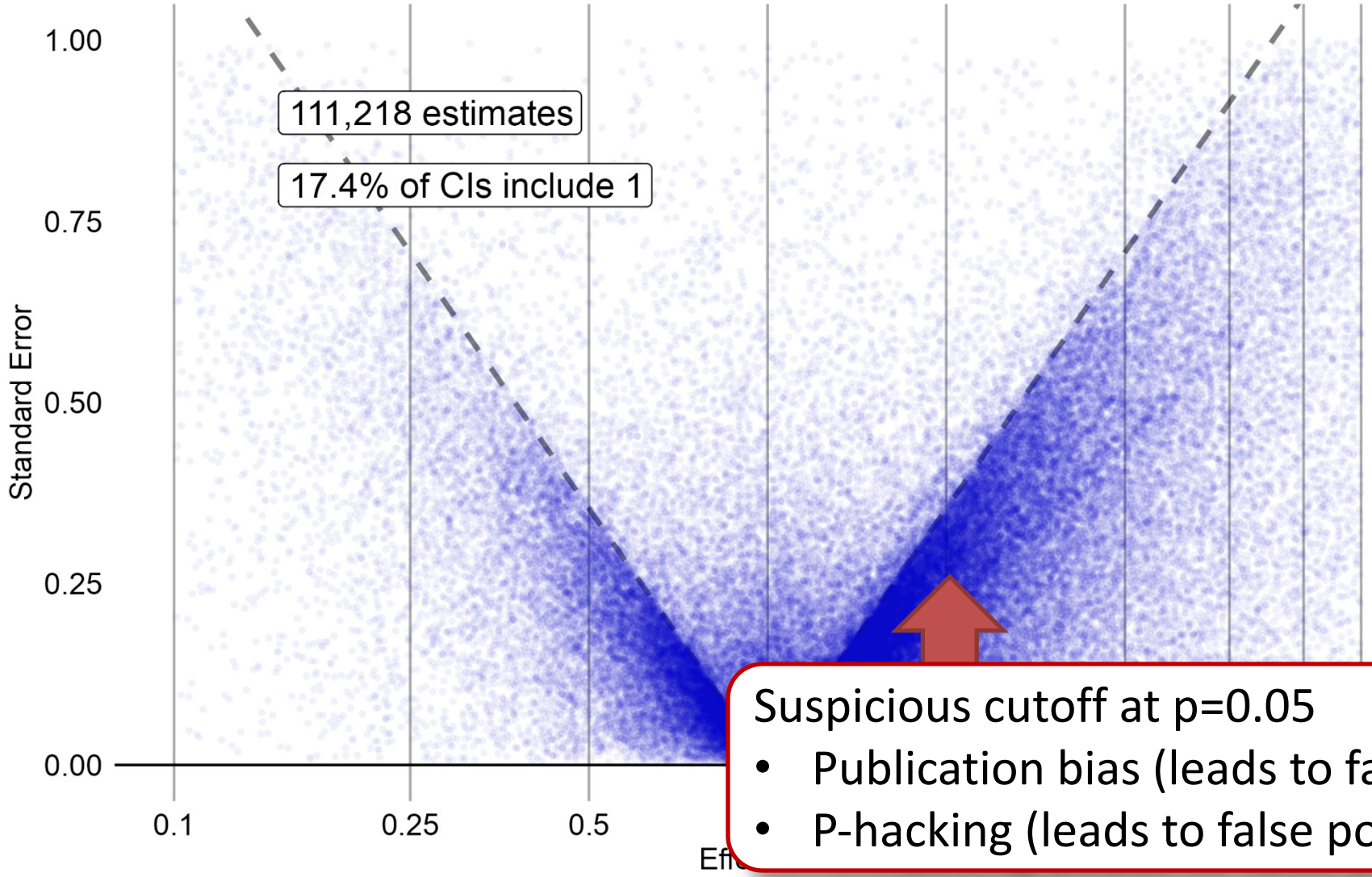


Published observational study results



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

Published observational study results



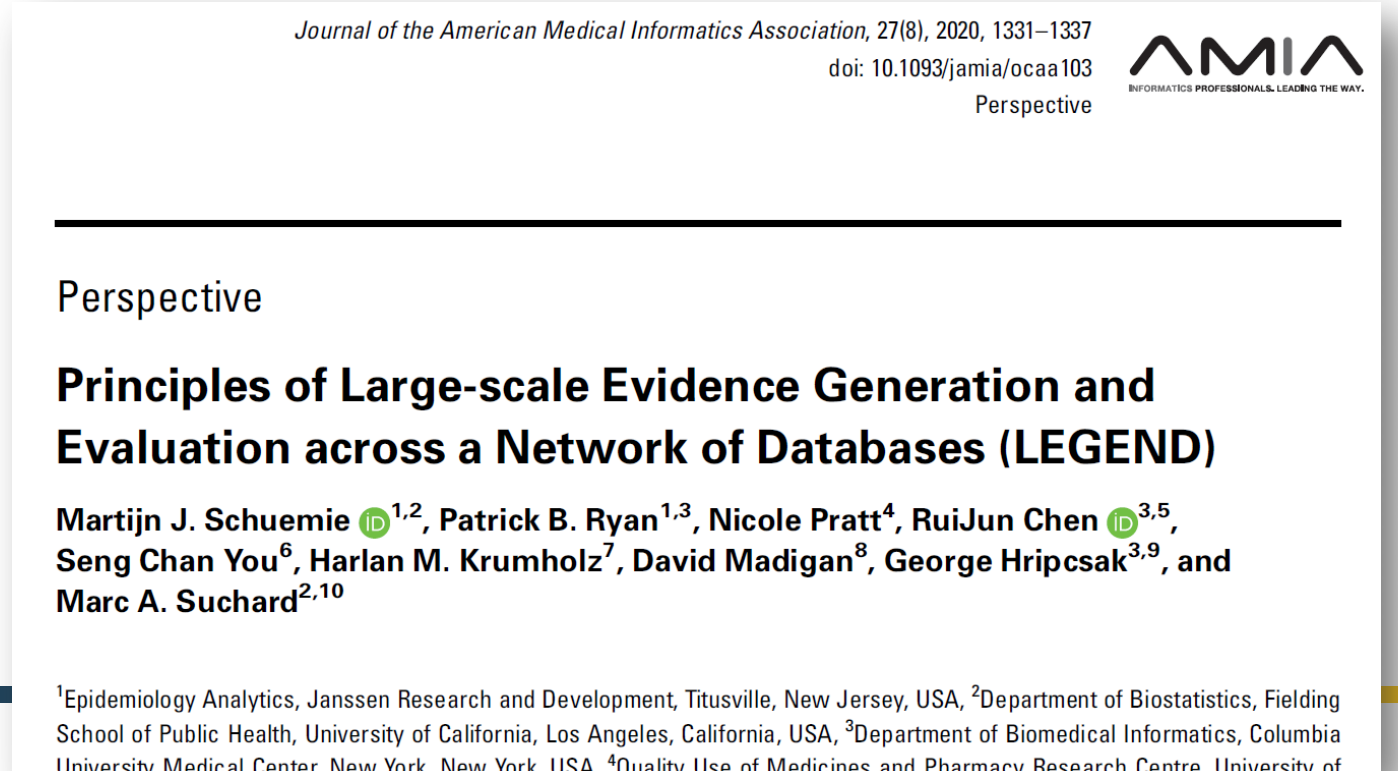
Suspicious cutoff at p=0.05

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

LEGEND

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.

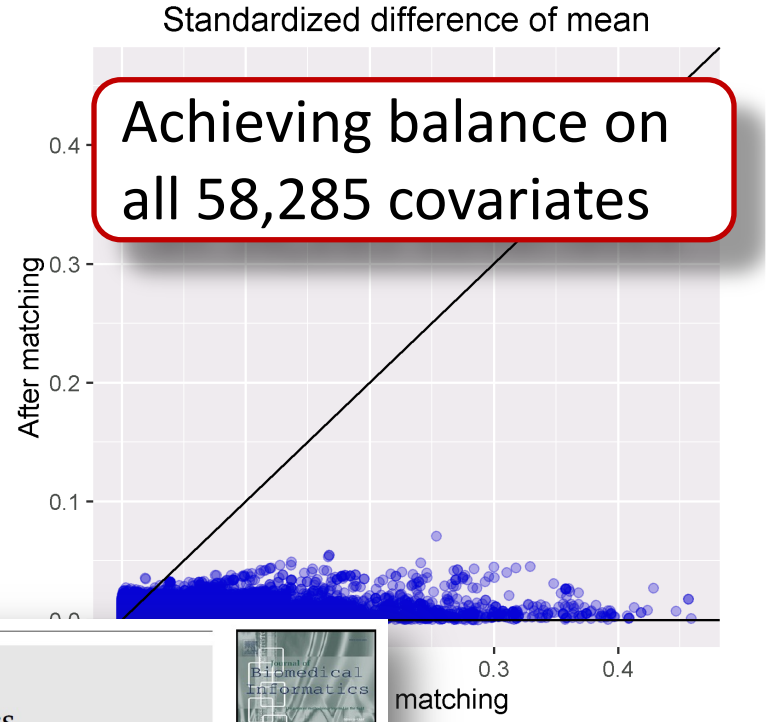
LEGEND Guiding Principles

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

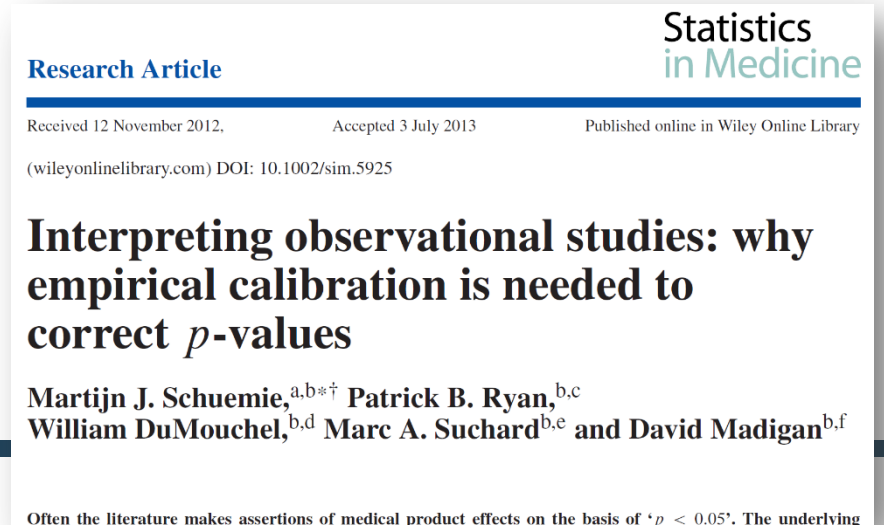
Measuring residual systematic error

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



Statistics in Medicine

Research Article

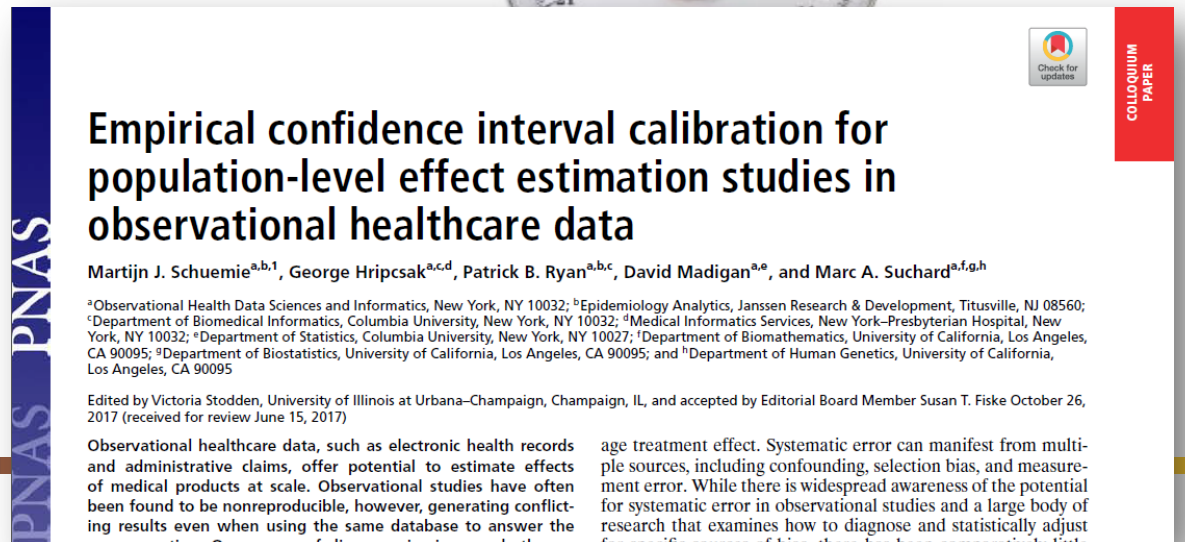
Received 12 November 2012, Accepted 3 July 2013 Published online in Wiley Online Library

(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

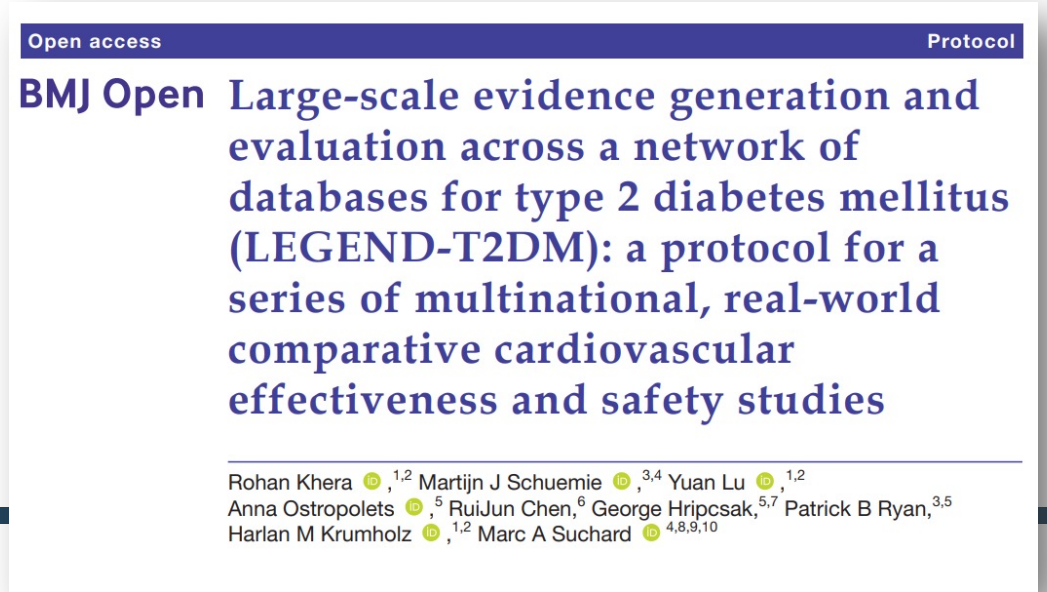
Edited by Victoria Stodden, University of Illinois at Urbana–Champaign, Champaign, IL, and accepted by Editorial Board Member Susan T. Fiske October 26, 2017 (received for review June 15, 2017)

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 Studies

- LEGEND principles initially tested in depression
- LEGEND Hypertension study has completed
- SCYLLA study also followed LEGEND principles
- Next LEGEND study is currently underway, estimating effects of diabetes treatments


Open access Protocol

BMJ Open

Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies

Rohan Khera ,[1,2] Martijn J Schuemie ,[3,4] Yuan Lu ,[1,2] Anna Ostropolets ,[5] RuiJun Chen,[6] George Hripcsak,[5,7] Patrick B Ryan,[3,5] Harlan M Krumholz ,[1,2] Marc A Suchard [4,8,9,10]

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

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.

JAMA Internal Medicine | Original Investigation

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

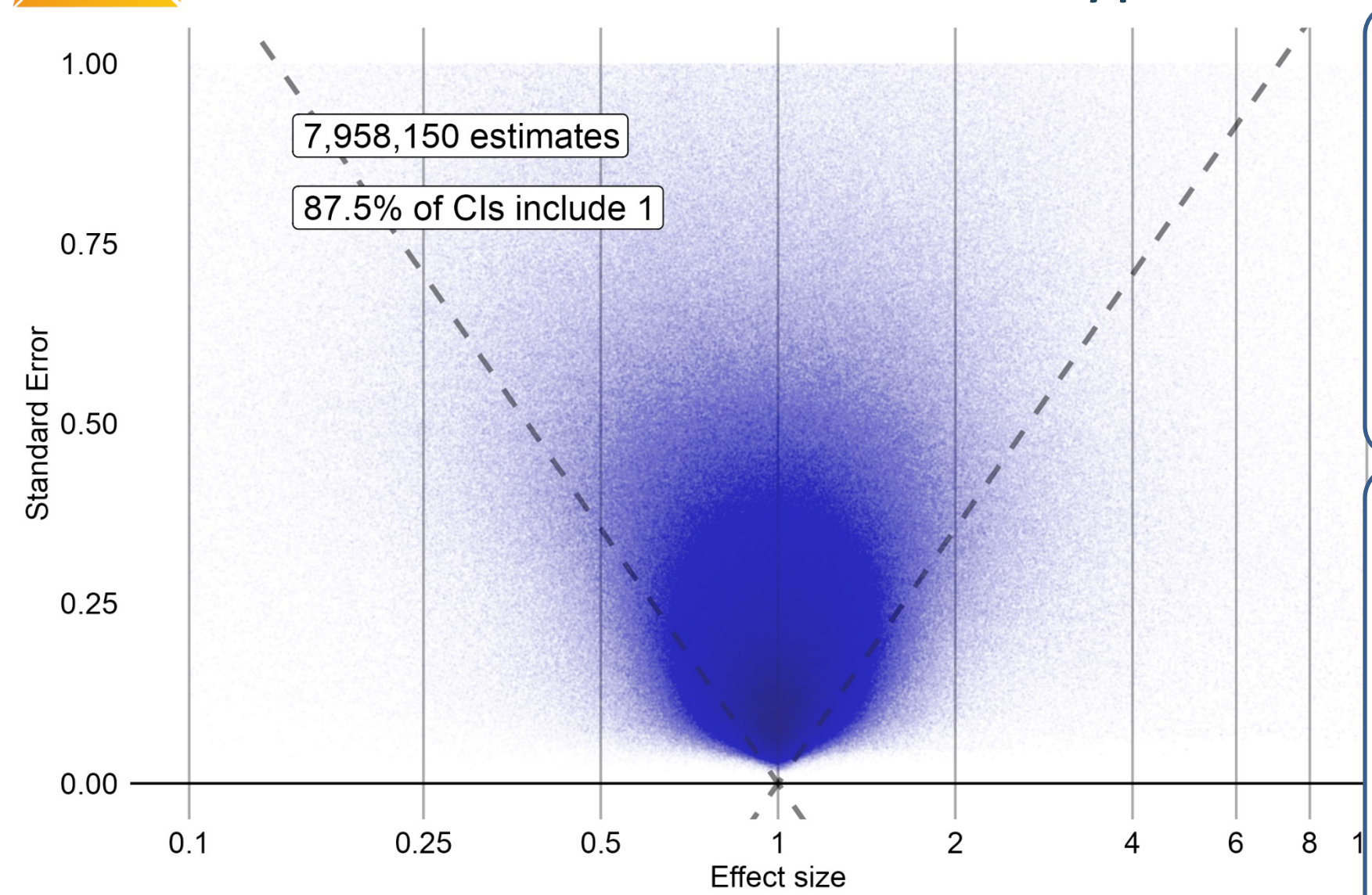
George Hripcsak, MD, MS; Marc A. Suchard, MD, PhD; Steven Shea, MD; RuiJun Chen, MD; Seng Chan You, MD; Nicole Pratt, PhD; David Madigan, PhD; Harlan M. Krumholz, MD, SM; Patrick B. Ryan, PhD; Martijn J. Schuemie, PhD

Research and Applications

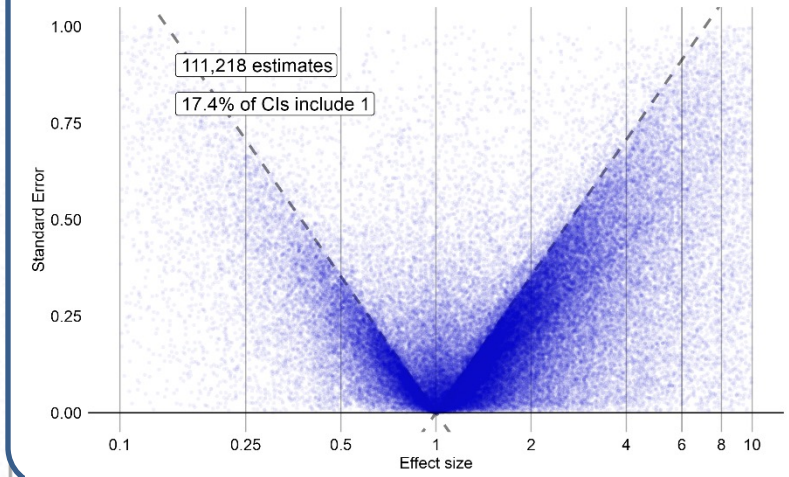
Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

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]

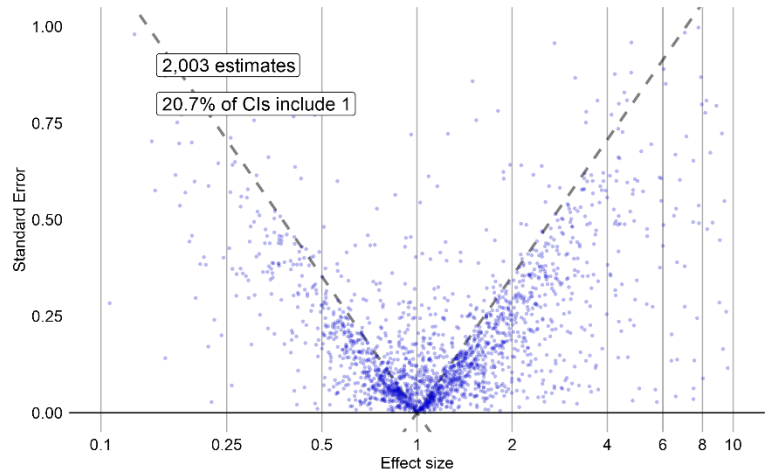
Distribution of estimates from LEGEND Hypertension



All published observational research



All published observational research on hypertension treatments

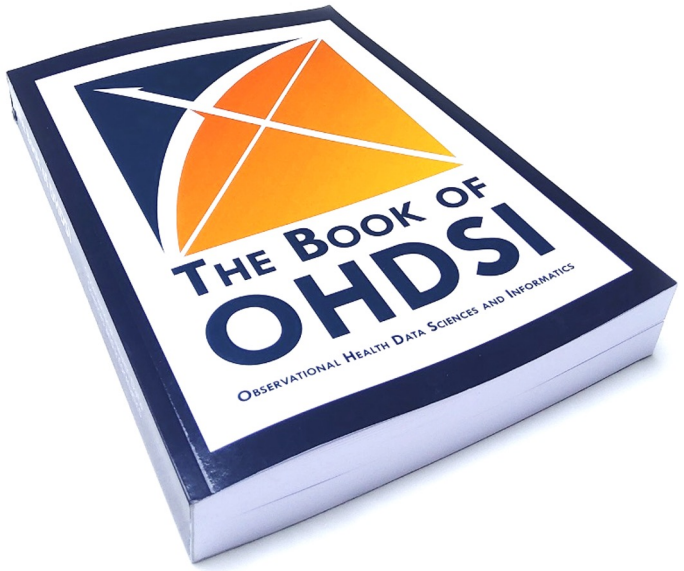


Diagnostics

- Each LEGEND estimate comes with full diagnostics, e.g.
  - Statistical power
  - Covariate balance
  - Systematic error as observed through negative controls
- Book of OHDSI chapter 19:

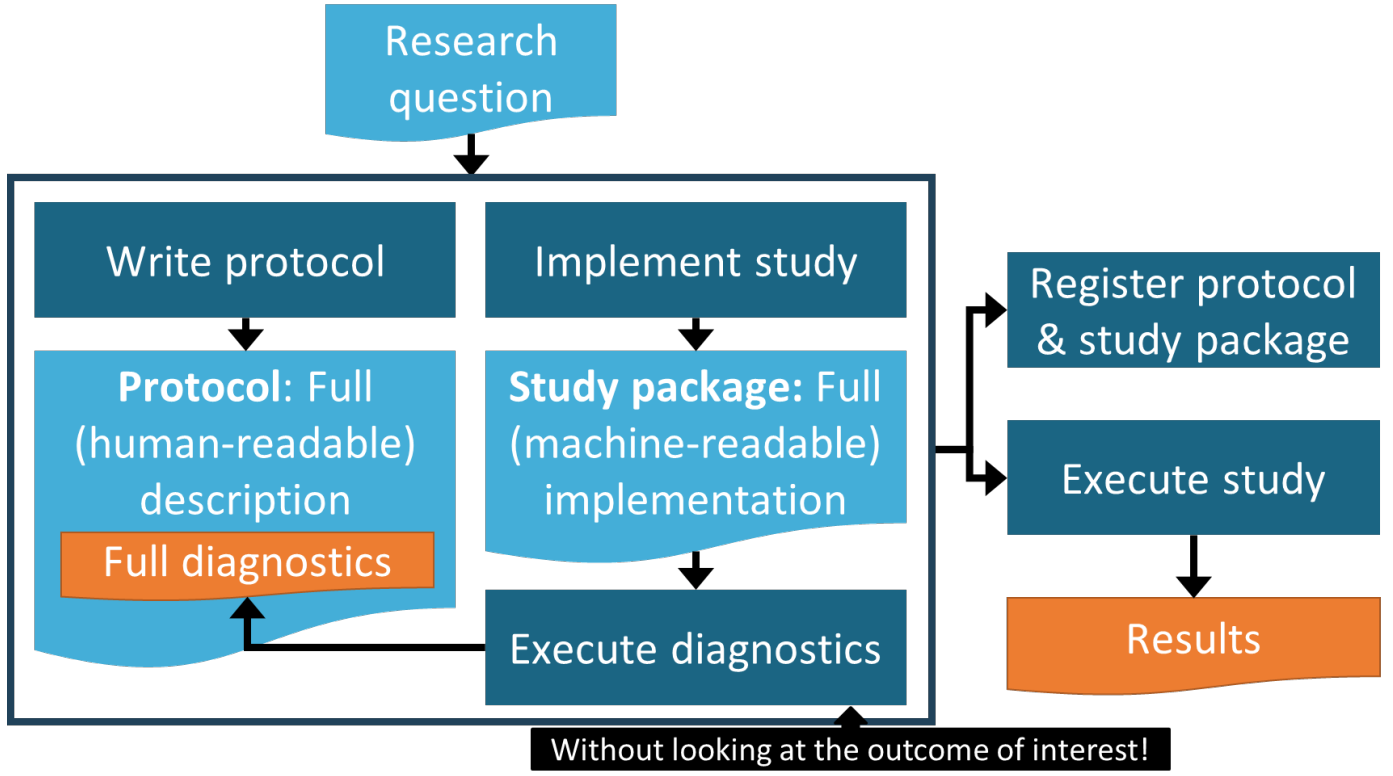
  "On the basis of these diagnostics results, a decision can then be made whether or not to move forward with executing the final outcome model."

How to make this decision?



Interpreting diagnostics

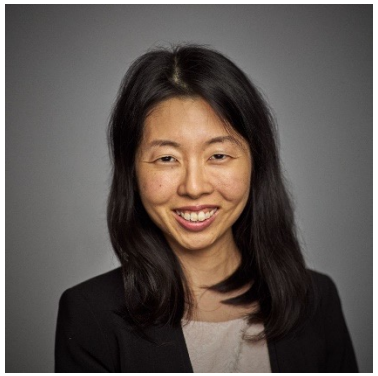
- Diagnostics inform us on whether we can trust the results of a study
- I've argued you shouldn't unblind results unless you pass diagnostics
- LEGEND itself did not blind
- For LEGEND papers the researchers checked the diagnostics



SCYLLA project

SARS-**C**ov-2 **L**arge-scale **L**ongitudinal **A**nalyses on the comparative safety and effectiveness of treatments under evaluation for COVID-19 across an international observational data network

- LEGEND-like study into the safety and effectiveness of drugs proposed to treat COVID-19.
  - 650 treatment comparisons
  - 31 outcomes of interest
  - Several different analyses



SCYLLA's approach to diagnostics

- Pre-defined rules for
  - Equipoise
  - Covariate balance
  - Statistical power
- Only unblind results that met all diagnostics
- Only a small fraction (1.5% - 8.9%) met all diagnostics

Challenges with study diagnostics

- Interpretation of diagnostics is currently subjective.
- Failing a diagnostic should mean you stop the study.
- That is a big ask when
  - You've invested a lot of time and energy in the study
  - You've already looked at the result
- Failing a diagnostic is currently not publishable
  - No credit for your hard work
  - Others who do not evaluate diagnostic will publish potentially unreliable evidence anyway

Diagnostics are hard!

Are they worth it?

Objective diagnostics are!

Patrick Ryan

Johnson & Johnson

Columbia University Irving Medical Center

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Who We Are | OHDSI Updates & News | Standards | Software Tools | OHDSI Studies | Book of OHDSI | Resources | New To OHDSI?

OHDSI Community Calls | Events & Past Collaborations | Workgroups | EHDEN Academy | This Week In OHDSI | Our Journey (PDF)

Publications | Support & Sponsorship | 2022 OHDSI Symposium | 2022 APAC Symposium | Newsletters | Follow OHDSI on Social

OHDSI Community Calls

Everybody is invited to the weekly OHDSI community call, which takes place each Tuesday at 11 am ET. These calls are meant to inform and engage our community through a variety of call formats, including community presentations, working group updates, breakout sessions, focus topics, newcomer-focused sessions, and more. The upcoming schedule is available to the right.

Use this link to get to the weekly meeting.

Videos and slides from previous 2022 calls will be posted below. All presentations from 2021 community calls can be found here. Both videos and slides from community calls prior to 2021 remain available.

— Jan. 11: What Can We Accomplish Together in 2022 (Patrick Ryan)

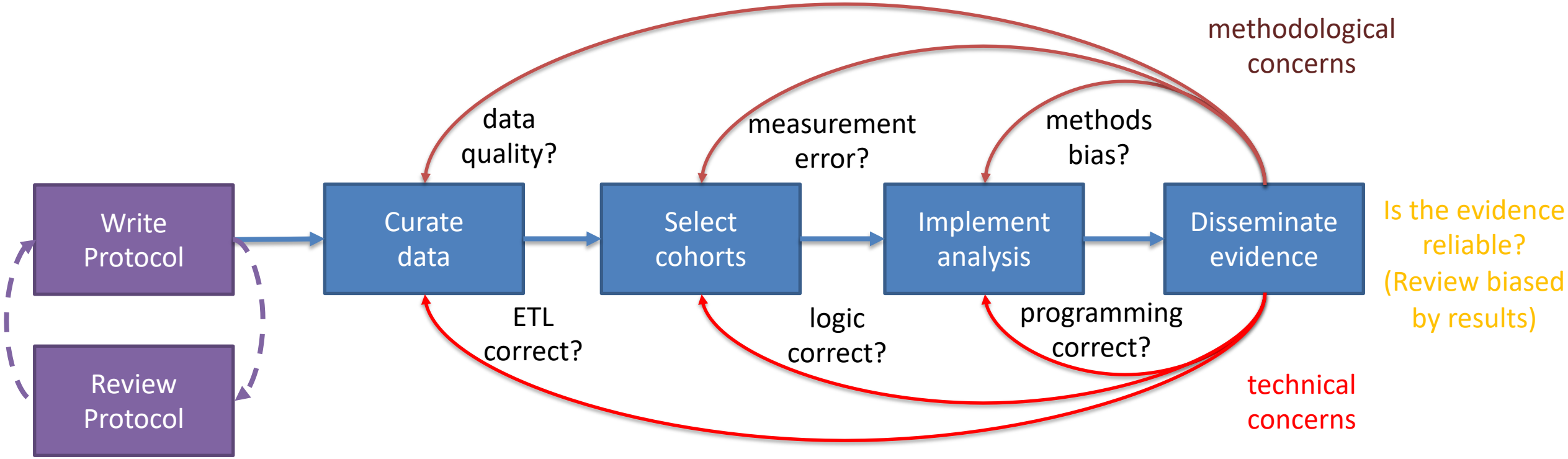
Patrick Ryan led the first OHDSI Community Call of 2022 with a presentation about what OHDSI can accomplish together this year. While the community listed and voted upon several objectives, Patrick discussed his hope to develop a system to generate evidence that characterizes disease and treatment utilization, estimates the effects of medical interventions, and predicts outcomes of patients within a network of observational health databases.

Other aspects of the discussion included a look at OHDSI workgroups and how they can continue moving forward, 2021 achievements, and more. Both video of the presentation and Patrick's slides are available below.

Video

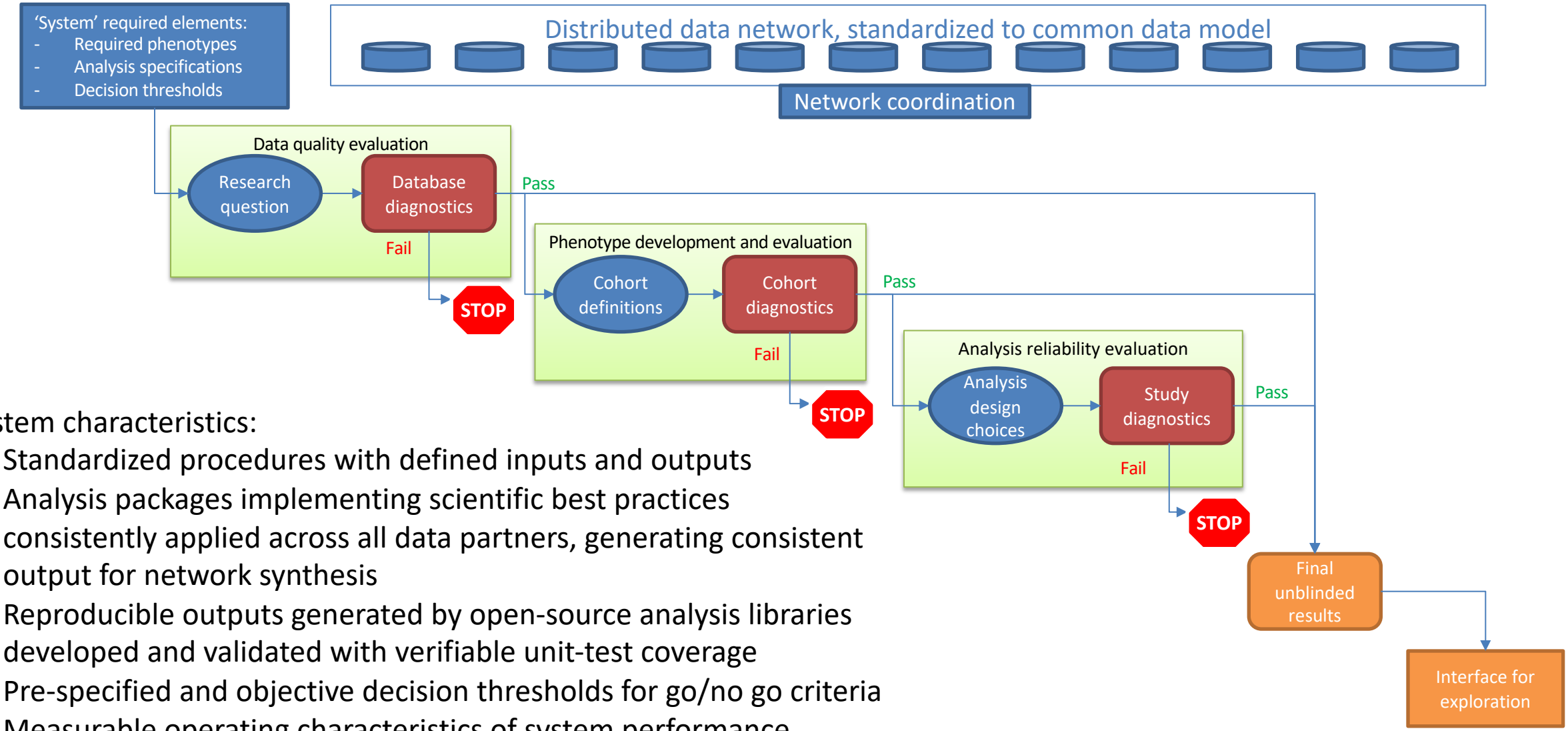
Current status quo in observational research makes it challenging to build trust in evidence

Does the study provide an unbiased effect estimate?
Are the findings generalizable to the population of interest?



Can the study be fully reproduced?
Does the analysis actually do what the protocol said it would do?

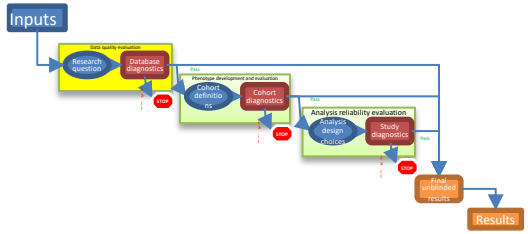
Engineering open science systems that build trust into the real-world evidence generation and dissemination process



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

Database diagnostics



- Challenge: Database selection is often subjective and opportunistic, based on pre-conceived notions of data acceptability
- Opportunity: Provide objective criteria with pre-specified decision thresholds for identifying candidate databases across a network that may be eligible for contributing to an analysis, without requiring direct data access
- Approach: Using only aggregated summary statistics from each data partner (via ACHILLES), assess data fitness-for-use in terms of patient demographics, domain coverage, longitudinality, and capture of target/comparator/outcome

github.com/OHDSI/DbDiagnostics

Search or jump to... / Pull requests Issues Marketplace Explore

OHDSI / DbDiagnostics Public

Edit Pins

Code Issues 1 Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 1 tag Go to file Add file Code

clairblacketer Correct output format from wide to long 5e7945c 20 hours ago 45 commits

R	Correct output format from wide to long	20 hours ago
extras	completed code review	13 days ago
inst	Correct output format from wide to long	20 hours ago
man	Correct output format from wide to long	20 hours ago
.Rbuildignore	Add executeDbDiagnostics function	15 days ago
.gitignore	Cleaning up files	5 months ago
DESCRIPTION	Use writeCsv function from CohortGenerator	7 days ago
DbDiagnostics.Rproj	Fix name of .Rproj file	8 days ago
NAMESPACE	completed code review	13 days ago
README.md	Update README.md	14 days ago

README.md

DbDiagnostics README

The `executeDbProfile` function in this package relies on the Achilles and DataQualityDashboard packages to run a subset of characterization and data quality analyses. This subset is referred to as the database profile. This profile will be used to determine if a database has the necessary elements required to run a study.

It works by connecting to a database through a connectionDetails object created by the DatabaseConnector package. It will then check to see if Achilles results are already present. If so, it will export those results. If not, it will run the required Achilles analyses and then export. Then, it will run a set of DataQualityDashboard checks and export those results as well.

Once the results are generated they are then loaded to a separate results schema. The `executeDbDiagnostics` function will take in a list of analysis settings to compare against the dbProfile results to determine if a database is

About

Package to profile a database and execute data diagnostics based on individual analysis settings

Readme
0 stars
11 watching
0 forks

Releases 1

DbDiagnostics v0.1 Latest
on May 5

Packages

No packages published
Publish your first package

Contributors 3

clairblacketer
msuchard Marc Suchard
fdefalco Frank DeFalco

Languages

R 84.7% Roff 15.3%

Database diagnostics in action

		Study size estimate		Database diagnostic criteria											
Analysis	Database	minimum expected persons (T+C)	maximum possible persons (T+C)	maxSampleSize	Age Range	Gender	Race	Ethnicity	Calendar Time	Lo				Concepts	Total diagnostic fails
				>1,000	18-100	M/F	All	All	2007-2019	>365					
lisinopril v benazepril for acute myocardial infarction				3	0	0	0	0	0	1					11
	IBM_CCAE-20220801	712,284	1,905,789	0	0	0	0	0	0	0	0	0	0	0	0
	OPTUM_Extended_DOD-20220805	423,273	1,437,973	0	0	0	0	0	0	0	0	0	0	0	0
	IBM_MDCR-20220729	280,303	635,033	0	0	0	0	0	0	0	0	0	0	0	0
	PharMetrics-20220515	262,184	975,941	0	0	0	0	0	0	0	0	0	0	0	0
	Optum_EHR-20220730	203,539	992,424	0	0	0	0	0	0	0	0	0	0	0	0
	IBM_MDCD-20220802	44,413	163,702	0	0	0	0	0	0	0	0	0	0	0	0
	AMBULATORY_EMR-20220530	496,469	1,561,268	0	0	0	0	0	0	0	0	1	0	0	1
	PREMIER-20220606	5,786	911,416	0	0	0	0	0	0	1	0	0	0	0	1
	German_DA-20220120	831	6,560	0							0	1	0	0	1
	France_DA-20220120	139	1,086	0							0	1	0	0	1
	JMDC-20220801	68	469	1							0	0	0	0	1
	CPRD-20220606	-	-	1							0	1	0	1	3
	LPDAU-20220121		-	1							0	1	0	1	3

6 databases pass all database diagnostics and are estimated to have adequate sample to move forward to cohort diagnostics

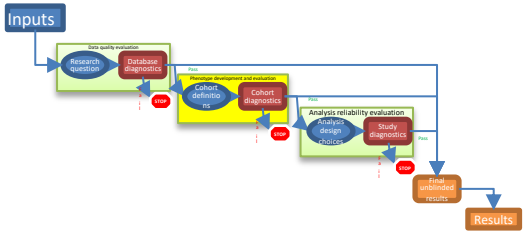
Premier fails diagnostics because it doesn't have longitudinal follow-up

JMDC fails diagnostics it is not expected to have sufficient exposures

5 databases fail diagnostics because they don't capture inpatient visits

CPRD and Iqvia Australia fail diagnostics because they don't have records for the comparator drug

Phenotype diagnostics



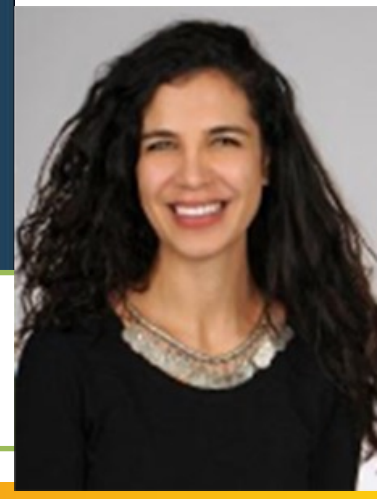
- Challenge: Phenotype algorithms to identify exposures and outcomes are subject to measurement error which can cause misspecification bias in analyses
- Opportunity: Provide objective criteria with pre-specified decision thresholds for evaluating the adequacy of candidate phenotype algorithms within each database across a network
- Approach: Develop a standardized process for developing phenotype algorithms and estimating all dimensions of measurement error (sensitivity, specificity, positive predictive value, index date misspecification) to determine the extent to which the magnitude of error will bias study results

Encouraging progress in 2022 on phenotype development and evaluation…

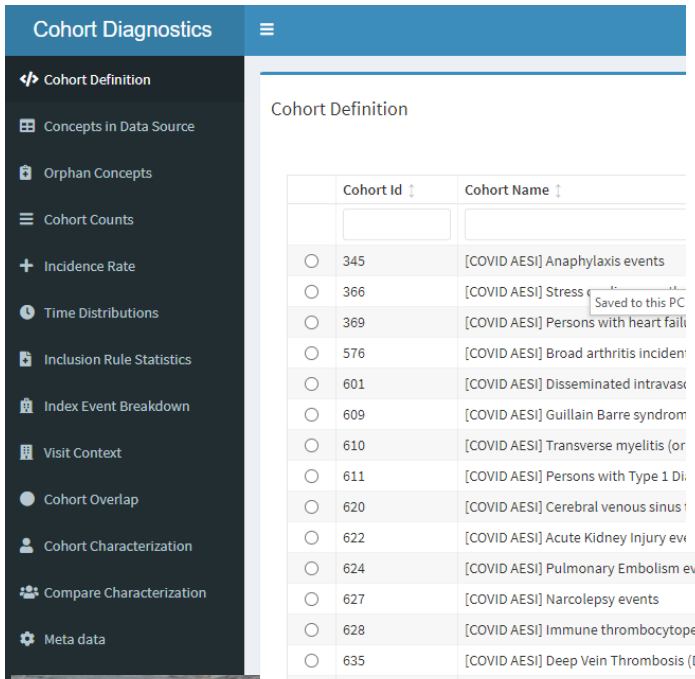
Phenotype Phebruar

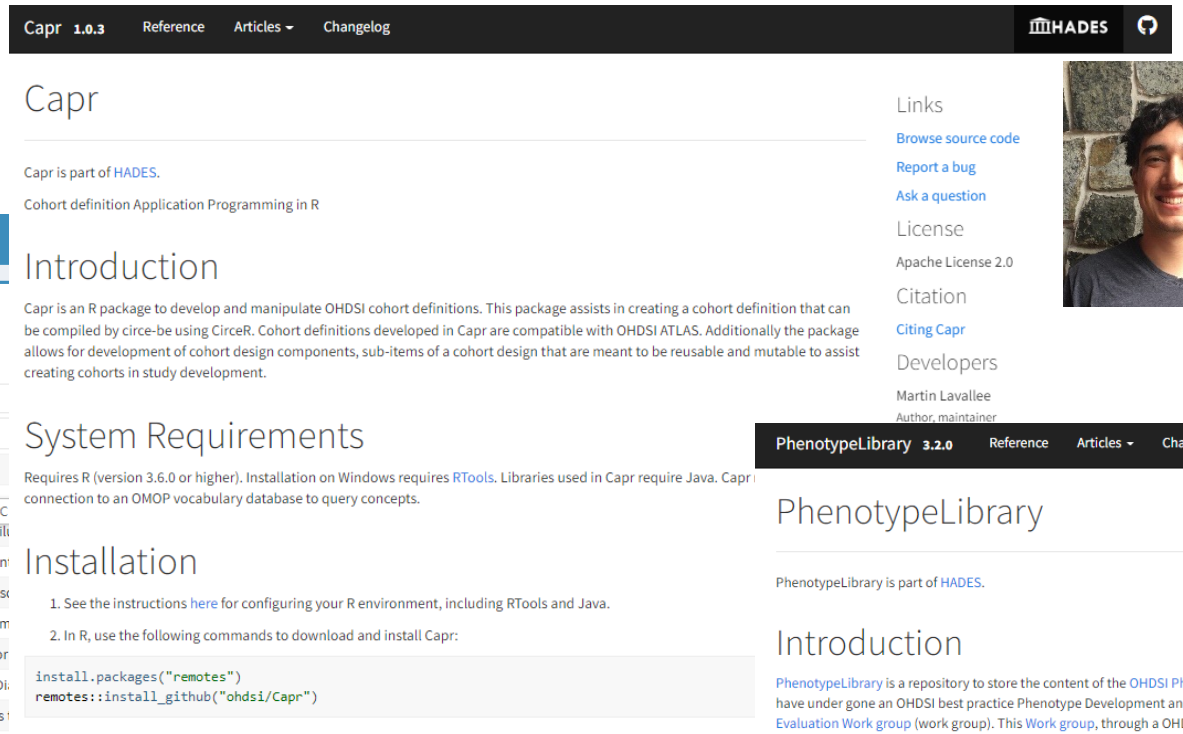
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
		1 Type 2 Diabetes Mellitus (Patrick Ryan)	2 Type 1 Diabetes Mellitus (Ryan)	3 Atrial Fibrillation (Ryan)	4 Multiple Myeloma (Ryan)	5 Alzheimer's Disease (Ryan)
6 Hemorrhagic Events (Ryan)	7 Neutropenia (Ryan)	8 Kidney Stones (Ryan)	9 Delirium (Azza Shoaibi)	10 Systemic Lupus Erythematosus (Joel Swerdel)	11 Suicidal Thoughts (Shoaibi)	12 Parkinson's Disease (Allan Wu)
13 Attention Deficit Hyperactivity Disorder (Ryan)	14 Hypertension (Gowtham Rao)	15 Acute Myocardial Infarction (Rao)	16 Heart Failure (Rao)	17 Cardiomyopathy (Rao)	18 Multiple Sclerosis (Shoaibi)	19 Triple Negative Breast Cancer (Adam Black)
20 Pulmonary Hypertension (Evan Minty)	21 Prostate Cancer (Asieh Golozar)	22 HIV (Rupa Makadia)	23 Hidradenitis Suppurativa (Jill Hardin)	24 Anaphylaxis (Erica Voss)	25 Depression (Tiffany Callahan, Juan Banda)	26 Non-Small Cell Lung Cancer (Golozar)
27 Drug-Induced Liver Injury (Anna Ostropolets)	28 Severe Visual Impairment & Blindness (Claudia Pulgarin)	Bonus Acute Kidney Injury (Marcela Rivera, David Vizcaya)				

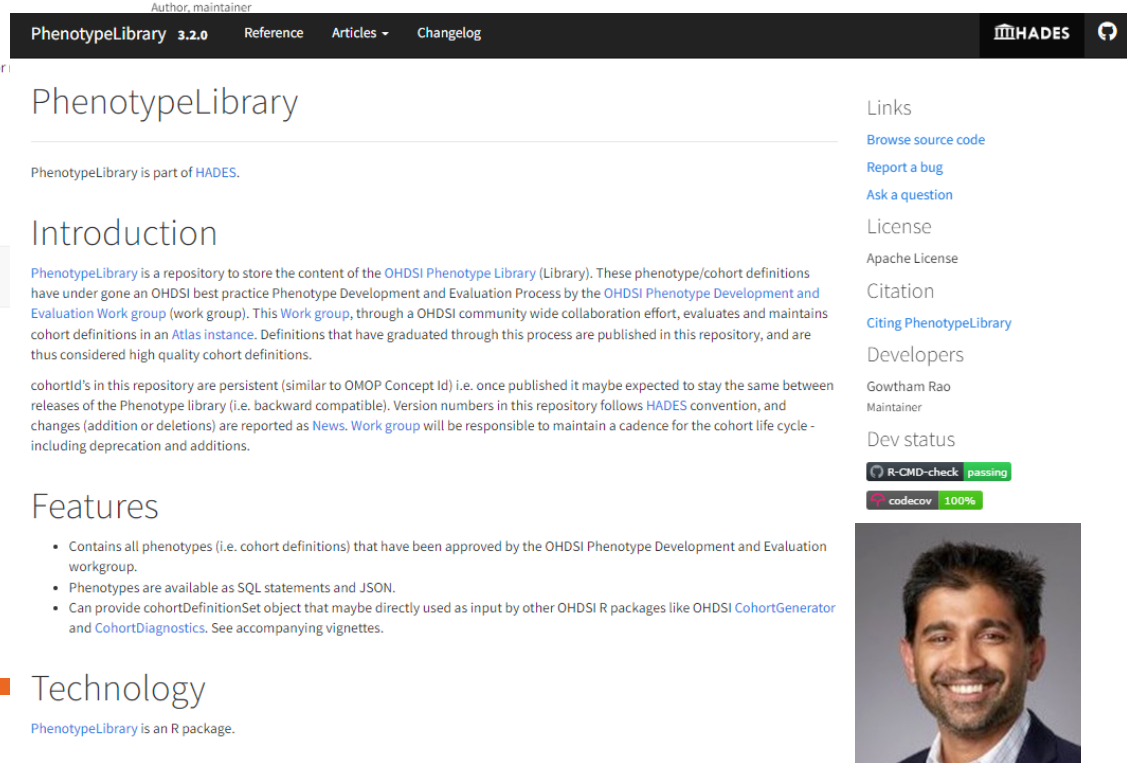
Check out Azza Shoaibi's Lightning Talk



Open-source tools to support phenotype development, evaluation, and dissemination







Journal of Biomedical Informatics 135 (2022) 104177

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation

Joel N. Swerdel [a,c,*], Martijn Schuemie [a,c], Gayle Murray [a], Patrick B. Ryan [a,b,c]

[a] Janssen Research and Development, Titusville, NJ, USA
[b] Columbia University, New York, NY, USA
[c] Observational Health Data Sciences and Informatics (OHDSI), New York, NY

ARTICLE INFO

Keywords:
Phenotype algorithms
Positive predictive value
Sensitivity
Specificity

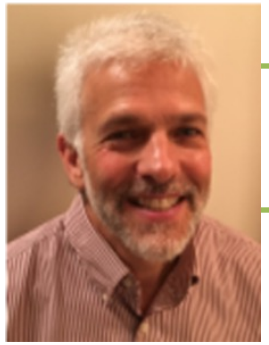
ABSTRACT

Purpose: Phenotype algorithms are central to performing analyses using observational data. These algorithms translate the clinical idea of a health condition into an executable set of rules allowing for queries of data elements from a database. PheValuator, a software package in the Observational Health Data Sciences and Informatics (OHDSI) tool stack, provides a method to assess the performance characteristics of these algorithms, namely, sensitivity, specificity, and positive and negative predictive value. It uses machine learning to develop predictive models for determining a probabilistic gold standard of subjects for assessment of cases and non-cases of health conditions. PheValuator was developed to complement or even replace the traditional approach of algorithm validation, i.e., by expert assessment of subject records through chart review. Results in our first PheValuator paper suggest a systematic underestimation of the PPV compared to previous results using chart review. In this paper we evaluate modifications made to the method designed to improve its performance.

Methods: The major changes to PheValuator included allowing all diagnostic conditions, clinical observations, drug prescriptions, and laboratory measurements to be included as predictors within the modeling process whereas in the prior version there were significant restrictions on the included predictors. We also have allowed the temporal relationships of the predictors in the model. To evaluate the performance of the ompared the results from the new and original methods against results found from the literal validation of algorithms for 19 phenotypes. We performed these tests using data from five ses.

sment aggregating all phenotype algorithms, the median difference between the PheValuator old standard estimate for PPV was reduced from −21 (IQR −34, −3) in Version 1.0 to 4 (IQR −3, 15) using Version 2.0. We found a median difference in specificity of 3 (IQR 1, 4.25) for Version 1.0 and 3 (IQR 1, 4) for Version 2.0. The median difference between the two versions of PheValuator and the gold standard for estimates of sensitivity was reduced from −39 (-51, −20) to −16 (-34, −6).

Conclusion: PheValuator 2.0 produces estimates for the performance characteristics for phenotype algorithms that are significantly closer to estimates from traditional validation through chart review compared to version 1.0. With this tool in researcher's toolkits, methods, such as quantitative bias analysis, may now be used to improve the reliability and reproducibility of research studies using observational data.



Check out Joel Swerdel's posters #66 and 67

Table 3
Differences in estimates for Positive Predictive Value from PheValuator Version 1.0 and 2.0 and the gold standard estimates from prior validation studies.

		Positive Predictive Value Difference	
Therapeutic		Version 1.0	Version 2.0
Area	Condition	Median (IQR)	Median (IQR)
Overall	Overall (all therapeutic areas)	−21 (-34, −3)	4 (-3, 15)
Cardiovascular	Overall (Cardiovascular)	−25 (-34, −18)	0 (-6, 4)
	Atrial Fibrillation	−32 (-39, −28)	−2 (-3, 1)
	Pulmonary Embolism	−38 (-46, −24.25)	−1.5 (-8.25, 3.25)
	Venous Thromboembolism	−18 (-25, −5)	−3 (-14, 3)
	Ischemic Stroke	−22.5 (-28, −19.75)	4 (2, 5)
	Myocardial Infarction	−31 (-39, −20)	−1 (-6.5, 0.5)
Immunology	Overall (Immunology)	−27 (-39, −4)	7 (-2, 30.25)
	Ankylosing spondylitis	−51 (-59, −41.5)	−2.5 (-10, 5)
	Atopic dermatitis	−8 (-16.5, −2)	39 (26.5, 42.5)
	Ulcerative Colitis	−31 (-37.5, −21.25)	−1 (-5.75, 4)
	Crohns Disease	−32.5 (-39, −26.25)	−2 (-6, 2.75)
	Rheumatoid Arthritis	6 (-11, 16)	38 (23, 48)
	Psoriasis	−23 (-34, −6.5)	9 (-3, 26)
	Systemic Lupus Erythematosus	−34 (-42.5, −30)	8 (0, 9)
Infectious Disease	Overall (Infectious disease)	−1 (-12.5, 3)	2 (-3, 6)
	Viral Hepatitis C	−2.5 (-11.5, 3)	0.5 (-3, 3.75)
	Viral Hepatitis B	1 (-14.5, 1)	5 (-5.5, 6)
Neurology	Overall (Neurology)	−24 (-31.75, −17)	7 (-8, 12.75)
	Autism	−31.5 (-36, −26.5)	2.5 (-5.25, 7.5)
	Bipolar	−28 (-31.5, −26.25)	13 (11.5, 14.75)
	Epilepsy	−20 (-24, −10.75)	−7 (-18.75, 9.5)
Oncology	Overall (Oncology)	−1 (-4, 4.5)	14 (10, 18)
	Multiple Myeloma	0 (-3.25, 6.5)	13.5 (10, 22)
	Prostate Cancer	−2 (-4, 1.25)	14 (9.5, 16.5)

Defining the valid analytic space for quantitative bias analysis in pharmacoepidemiology

James Weaver[1,2,3], Patrick B Ryan[2,3,4], Victoria Strauss[1,3], Marc A Suchard[3,5], Joel Swerdel[2,3], Daniel Prieto-Alhambra[1,3,6]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; [2]Observational Health Data Analytics, Global Epidemiology, Janssen Research and Development, Titusville, NJ, USA; [3]Observational Health Data Sciences and Informatics, New York, NY, USA; [4]Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA; [5]Department of Biostatistics, Fielding School of Public Health, and Department of Biomathematics, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA; [6]Medical Informatics, Erasmus Medical Centre, Rotterdam, Netherlands

BACKGROUND

- Bias from outcome misclassification is acknowledged but rarely corrected in observational comparative safety and effectiveness research
- Quantitative bias analysis (QBA) can correct effect estimates subject to outcome misclassification using incidence proportion and estimated measurement errors
- Certain QBA input combinations **can produce negative corrected event counts that invalidates results**

OBJECTIVE

- Determine which combinations of observed effect estimates, incidence proportions, sensitivity and specificity values produce valid and invalid corrections

METHODS

- Created grid space of:
  - 6 outcome incidence proportions (IP) [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]
  - 6 uncorrected odds ratios (OR) [1, 1.25, 1.50, 2, 4, 10]
  - 20 outcome sensitivity values [0.05 to 1.00 by 0.05]
  - Specificity precision is dependent on outcome IP, so specificity values were generated within each level of IP. 20 specificity values were defined as 1-incidence to 1.00 by 5%ile
- Complete space: 14,440 2x2 contingency tables, each with 1m target and 1m comparator exposures and associated inputs
- For each IP-OR combination, we computed a distribution of QBA-corrected ORs across combinations of sensitivity and specificity values and plotted their contours
- We estimated the sensitivity, specificity, and IP of ischemic stroke in 5 observational databases (labeled as Source in figure) using probabilistic reference standard validation and plotted their location on the analytic space

QBA produces implausible or invalid outcome misclassification-corrected estimates in most common comparative effect estimation scenarios

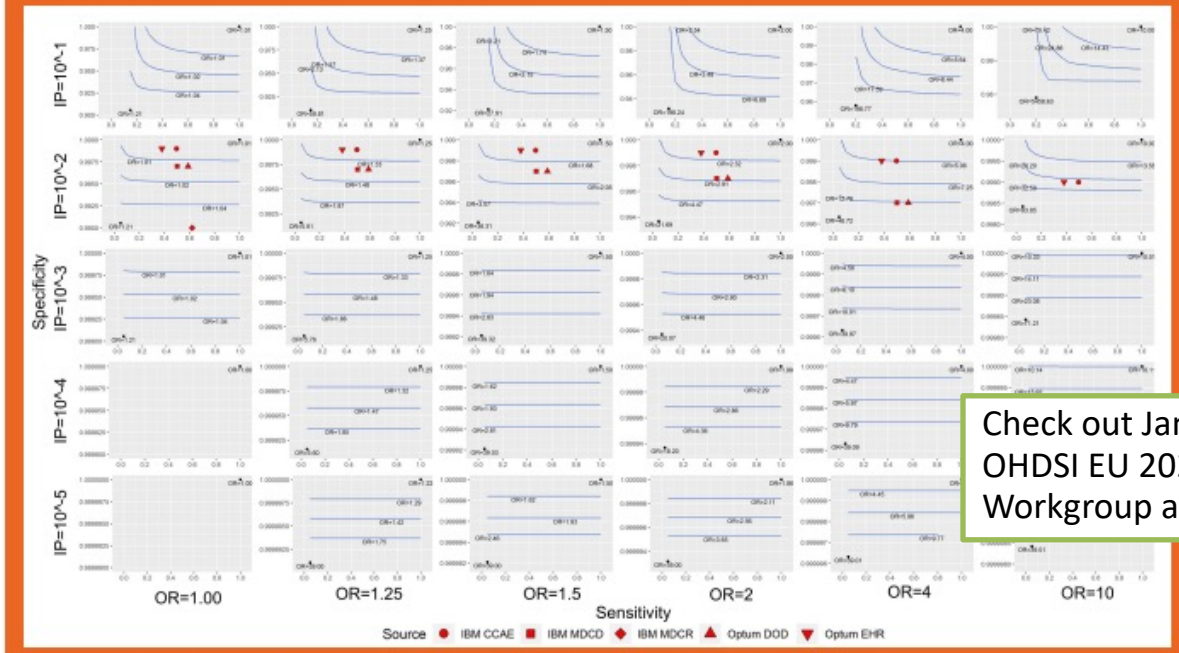


Figure 1. QBA-corrected OR contour plots across 4-dimensional grid space of IP, uncorrected OR, sensitivity, and specificity. Black data points are the uncorrected OR (sensitivity = specificity = 1) and maximum valid OR of the corrected OR distribution across sensitivity and specificity values for each IP-uncorrected OR combination. Blue lines display the corrected OR contour for the 25%ile, 50%ile, and 75%ile of the corrected OR distribution. Red data points are database-specific, empirical QBA-corrected estimates from a study assessing the risk of ischemic stroke between new users of angiotensin-converting enzyme inhibitors and angiotensin receptor blockers with hypertension.

RESULTS

- Minimum required specificity for valid QBA correction was inversely proportional to IP.
- Minimum specificity required for valid QBA correction is 0.91, observed where IP=10^{-1}.
- Where IP=10^{-5}, minimum required specificity is 0.9999911
- Lower value sensitivity variation at higher IP affected OR correction, but where incidence was ≤10^{-3}, only specificity materially affected correction
- Empirical results showed ischemic stroke IP as ~10^{-2} with measurement error variability across databases
- At higher uncorrected ORs, these measurement error values would considerably impact estimates
  - E.g., at uncorrected OR=4, the corrected estimate would be inflated >3x in three of five databases

DISCUSSION

- There is considerable IP-OR-sensitivity-specificity analytic space where QBA for outcome misclassification correction is implausible or invalid
- Correction with imprecise specificity is problematic because small specificity changes can make implausible large OR adjustments
- Impact of sensitivity on correction is limited where IP<10^{-2}

Check out Jamie Weaver's poster from OHDSI EU 2022, and join the Phenotype Workgroup activity this Sunday!

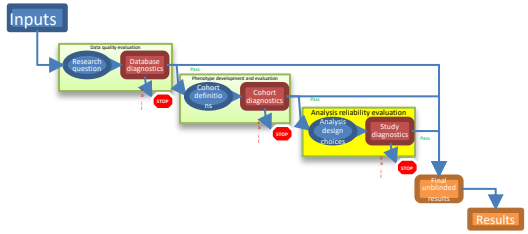
OHDSI OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

NDORMS NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

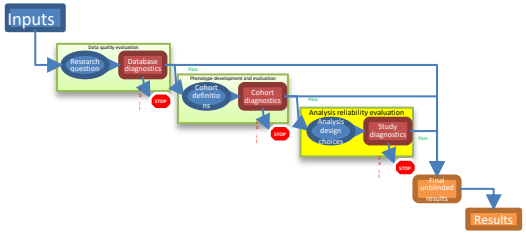
janssen PHARMACEUTICAL COMPANIES OF Johnson & Johnson

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

Study diagnostics



- Characterization
  - Feature summary, incidence, cohort pathways
    - Temporal stability, subpopulation heterogeneity, heterogeneity across data sources
- Population-level Estimation
  - Comparative cohort
    - Statistical power, comparator similarity, between-person confounding, generalizability, residual bias
  - Self-controlled case series
    - Statistical power, time-varying confounding, protopathic bias, residual bias
  - Meta-analysis
    - Statistical power, heterogeneity across data sources
- Patient-level prediction

Developing objective metrics to diagnose PatientLevelPrediction model designs

PRESENTER: **Jenna Reps**

INTRO:

- Prognostic models often fail to make any clinical impact. Investigations into why often identify poor methodology or applicability when developing models.
- PROBAST is a review guideline for methodology considerations when prognostic models are developed to identify potential causes of bias.
- In this study we propose objective measures taking into account PROBAST considerations that can highlight potential issues with a prediction study design.

METHODS

Given a model design and OMOP CDM database, metrics/plots were developed based on the four PROBAST aspects:

- **Participants:** investigates whether there are any issues in the design that may limit the generalizability of the model.
- **Predictors:** investigates whether the predictors (aka covariates/independent variables) are suitable. That is, do they only include data that would be available when the model is intended to be used?
- **Outcomes:** investigates whether the outcome definition is correct and is a commonly used definition. Is the database suitable for the outcome?
- **Design:** investigates whether the analysis is suitable. Is there sufficient data to learn a model?

Specific considerations are detailed in Figure 1.

RESULTS

A new function **diagnosePlp()** has been added into PatientLevelPrediction >= v5.4.0 to calculate these metrics/plots. This function can be used to identify potential model design issues.

PatientLevelPrediction now includes functions to **diagnose potential** issues (flaws) with **model design** when applied to a specific OMOP CDM database

Take a picture to **download** the **full paper**

Diagnostic results

probastId	probast category	consideration	explanation	cdm_truven_ccae_v2044
1.1	Participants	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	PatientLevelPrediction uses a cohort design	Pass
1.2.2	Participants	Were all inclusions and exclusions of participants appropriate?	Check the demographic differences between inclusion criteria in population settings and no additional inclusions (1 = similar, 0 = dissimilar)	0.999999999965047
1.2.4	Participants	Were all inclusions and exclusions of participants appropriate?	Check the demographic differences between inclusion criteria in population settings and no additional inclusions (1 = similar, 0 = dissimilar)	0.999999999752731
2.1	Predictors	Were predictors defined and assessed in a similar way for all participants?	PatientLevelPrediction uses standardized feature extraction to consistently engineer the predictors	Pass
2.2	Predictors	Were predictor assessments made without knowledge of outcome data?	Rule check is the last date used to engineer predictors before the time-at-risk start? (if it is then pass)	Pass
2.3	Predictors	Are all predictors available at the time the model is intended to be used?	This fails if the last date used to engineer predictors is after index but this requires the user to think about their data and design to determine whether this passes	Unknown
3.1	Outcome	Was the outcome determined appropriately?	This needs to be check via cohort diagnostic of the outcome cohort	NA
3.2	Outcome	Was a pre-specified or standard outcome definition used?	This passes if the outcome was from the OHDSI phenotype library (needs to be manually confirmed)	NA
3.3	Outcome	Were predictors excluded from the outcome definition?	Check Kaplan Meier plots to see whether outcomes occur close to index (which may mean predictors and outcome overlap) or not	NA
3.4	Outcome	Was the outcome defined and determined in a similar way for all participants?	PatientLevelPrediction determines the outcome based on an outcome phenotype, so the same logic is used for all patients.	Pass
3.5	Outcome	Was the outcome determined without knowledge of predictor information?	PatientLevelPrediction framework requires defining the outcome independently of the covariates - manually confirm ...	NA
3.6	Outcome	Was the time interval between predictor assessment and outcome determination appropriate?	Rule: does the time-at-risk start on or after target index? If so, pass.	Pass
4.1	Design	Were there a reasonable number of participants with the outcome?	Rule: If less than 100 fail.	Unknown

Figure 1: Example output of the new diagnostic functions as displayed in a diagnostic table in PatientLevelPrediction protocol.

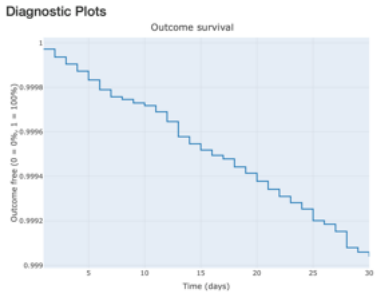
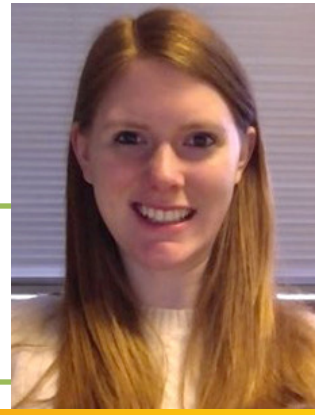


Figure 2: Example diagnostic plot to inspect the outcome definition and timing. In the example above the outcomes occur uniformly across the 30 days after index. Depending on the prediction task, this plot may help identify issues with the outcome definition or target population.

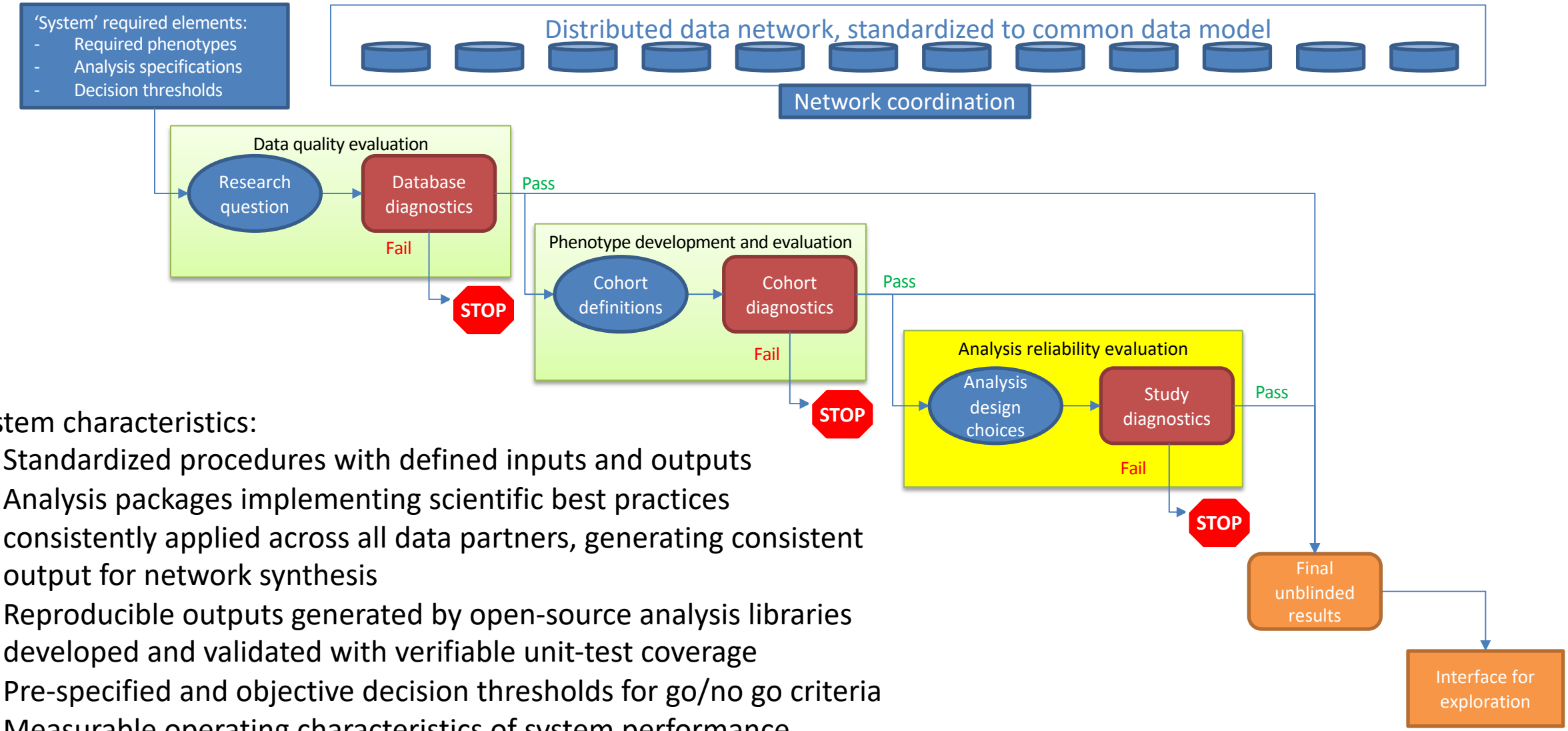
Check out Jenna Reps' poster #53



janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

OHDSI

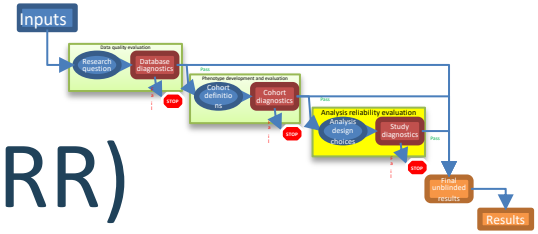
Focus for today: Establishing objective study diagnostics for comparative cohort analyses for population-level estimation



System characteristics:

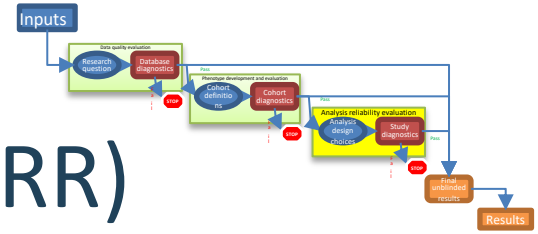
- Standardized procedures with defined inputs and outputs
- Analysis packages implementing scientific best practices consistently applied across all data partners, generating consistent output for network synthesis
- Reproducible outputs generated by open-source analysis libraries developed and validated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

Statistical power: Minimum detectable relative risk (MDRR)



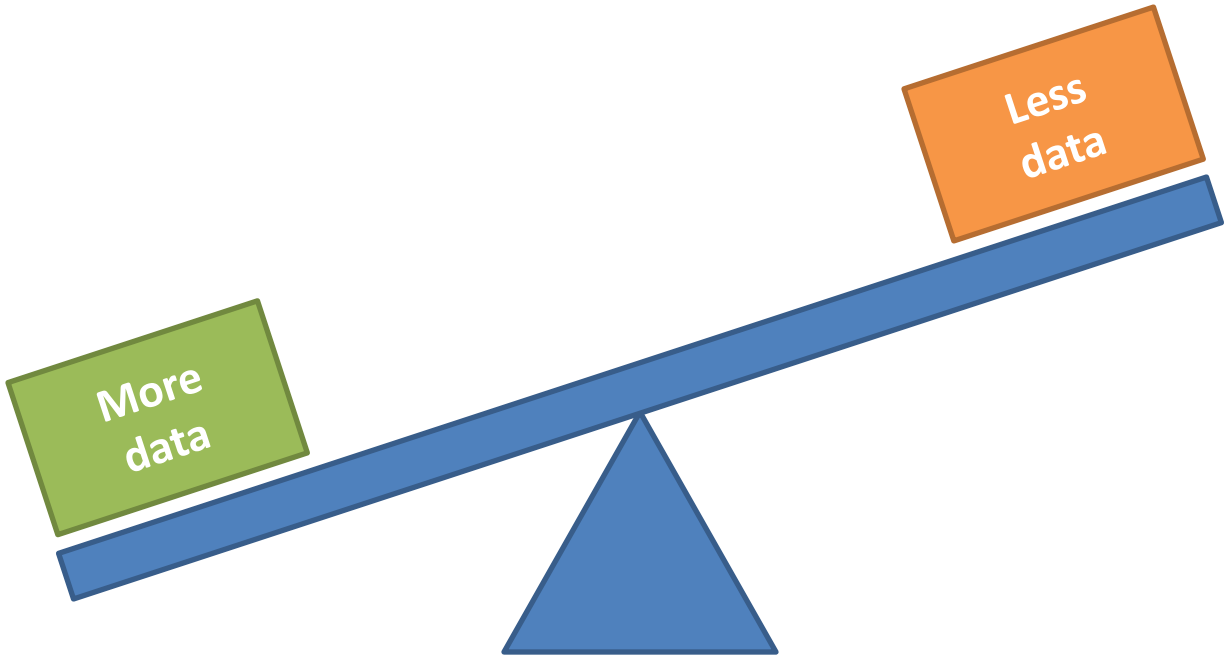
- Statistical power of a hypothesis test is the probability of detecting an effect if a true effect exists (1-Type II error)
  - Power analyses often conducted for interventional studies involving subject enrollment or non-interventional studies requiring primary data collection to determine the sample size that needs to be obtained, given the hypothesized effect size and background incidence
  - Given that sample size already exists when conducting non-interventional studies involving secondary use of existing clinical data, power analyses can be reformulated as: 'given the available data, what effect size would the analysis be able to detect?'
- More data provides greater power
  - Design and analysis choices impact how much data are used to generate estimates
- Potential diagnostic: how much data is sufficient to provide useful information?

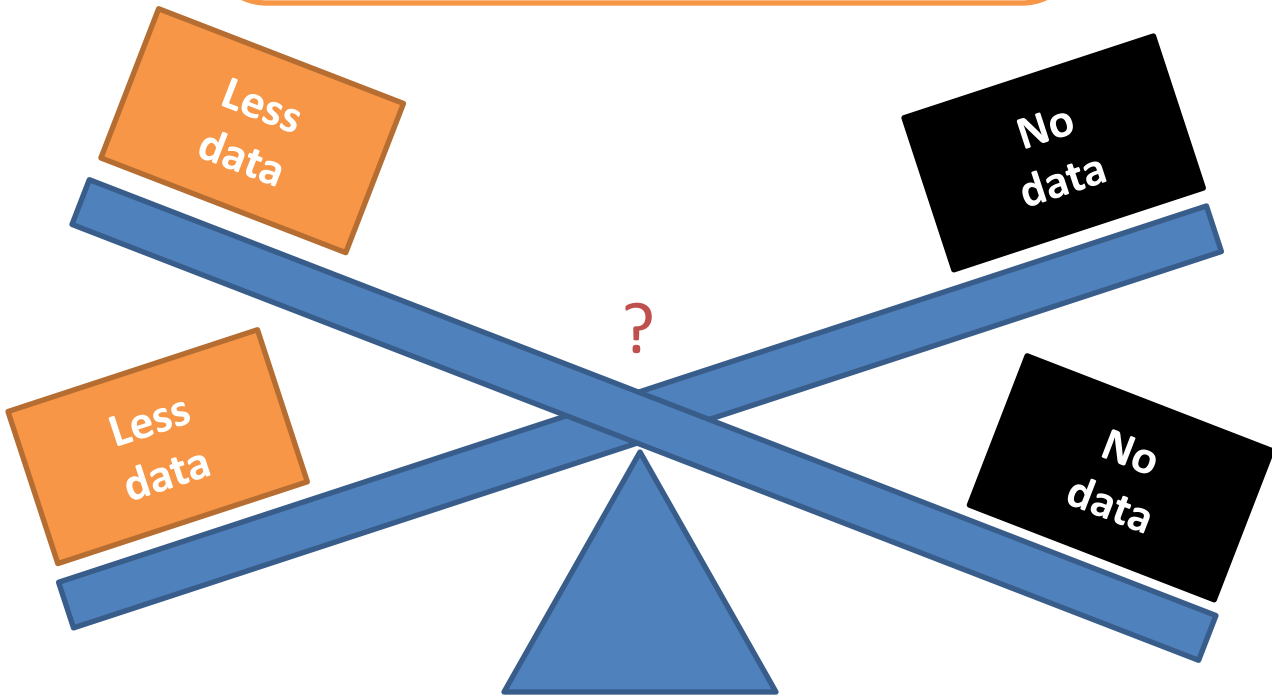
Statistical power: Minimum detectable relative risk (MDRR)



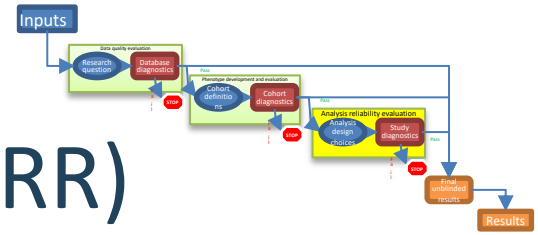
All else equal, we should all agree that it's better to have more data than less data, to improve precision in our estimates…

….but it's not straightforward to determine if less data is better than no data at all. Is the risk of misinterpreting underpowered estimates worth the marginal information gain?





Statistical power: Minimum detectable relative risk (MDRR)



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

Abstract

Observational databases are often used to study causal questions. Before being granted access to data or funding, researchers may need to prove that "the statistical power of their analysis will
imprecise estimates, will not be approved. This restrictive attitu
A key misunderstanding is the belief that the goal of a causa
that are either detected or undetected; causal effects are numeric
the effect as unbiasedly and precisely as possible, the solution t
observational analyses with imprecise estimates, but rather enco
have multiple studies with imprecise estimates than having no st
them and provide a more precise pooled effect estimate. Therefo
data cannot be that our estimates will be imprecise. Ethical a
which place individuals at risk are not transferable to observatio
If a causal question is important, analyze your data, publish y
The alternative is an unanswered question.

Journal of Clinical Epidemiology 141 (2022) 212–212

LETTER TO THE EDITOR

Sample size considerations are needed for the causal analyses of existing databases

Author statement

ptualization, writ-

ıad Ali Mansournia

Journal of Clinical Epidemiology 142 (2022) 261–263

A few things to consider not to conduct underpo

NTARY

s: the importance of understanding a before analysis (commentary on án)

urten van Smeden[b]

linical Trials and Methodology, London, UK
ty Medical Center, Utrecht University, Utrecht, Netherlands

able online 22 September 2021

cisely as possible." We note in passing that the hypothetical "socially alarmed" groups in the *Hypothetical example* may simply be interested in the binary signal of whether or not the unusual thrombotic events in vaccinated young people were made less unusual by the vaccine. However, we agree with the notion that the plausible magnitude of such an effect is important. Hernán's proposed solution is for groups to conduct causal analyses of several existing available data sources, which would subsequently be synthesised in a meta-analysis. This is a worthy goal. It also places a possibly-unbearable burden on systematic review-

Journal of Clinical Epidemiology 144 (2022) 193–193

COMMENTARY

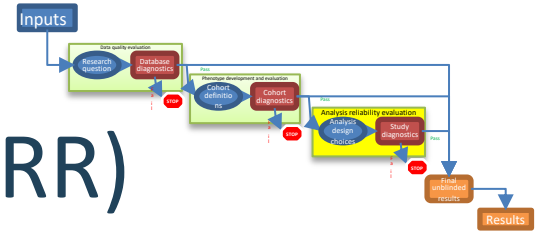
Causal analysis of existing databases: no power calculations required. Responses to Campbell, Morris and Mansournia, et al

Miguel A. Hernán[a,b,*]

[a] CAUSALab, Harvard T.H. Chan School of Public Health, Boston, MA
[b] Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston MA 6174320101

Accepted 30 November 2021; Available online 4 December 2021

Statistical power: Minimum detectable relative risk (MDRR) Examples from LEGEND-HTN



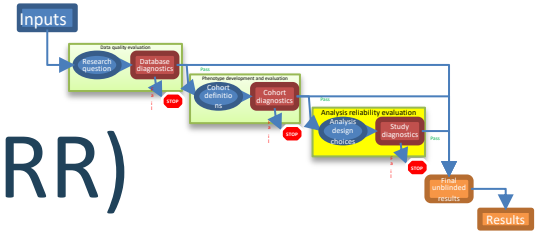
Good:
T = lisinopril
C = hydrochlorothiazide
O = cough

All databases have MDRR < 1.75 (ability to detect 75% increased risk if present), and 5 databases have MDRR < 1.1 (ability to detect 10% increased risk)

Table 1a. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate (IR) per 1,000 patient years (PY) in the target (*Lisinopril*) and comparator (*Hydrochlorothiazide*) group after stratification, as well as the minimum detectable relative risk (MDRR). Note that the IR does not account for any stratification.

Source	Target subjects	Comparator subjects	Target years	Comparator years	Target events	Comparator events	Target IR (per 1,000 PY)	Comparator IR (per 1,000 PY)	MDRR
CUMC	3,565	3,387	4,563	5,555	284	288	62.23	51.84	1.26
IMSG	2,980	1,443	2,034	683	96	26	47.19	38.04	1.72
MDCD	45,283	24,993	20,591	9,038	3,249	1,206	157.79	133.42	1.09
MDCR	60,853	28,461	48,503	22,586	4,831	1,514	99.60	67.03	1.08
Optum	364,307	154,543	261,838	100,906	25,947	7,631	99.10	75.62	1.03
CCAE	548,859	243,878	380,386	163,469	30,942	9,419	81.34	57.62	1.03
Panther	583,608	189,242	207,470	66,877	21,366	5,369	102.98	80.28	1.04
Summary	1,609,455	645,947	925,388	369,118	86,715	25,453	93.71	68.96	1.02

Statistical power: Minimum detectable relative risk (MDRR) Examples from LEGEND-HTN



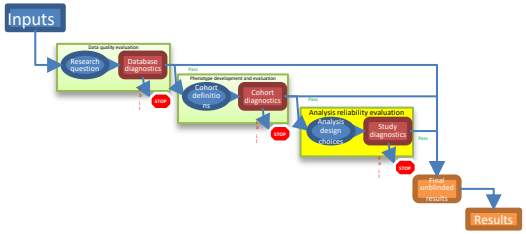
Bad:
T = candesartan
C = chlorthalidone
O = rhabdomyolysis

All databases have MDRR > 6 (underpowered to detect 600% increased risk if present), and two databases have MDRR > 15
<5 cases in target and comparator

Table 1a. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate (IR) per 1,000 patient years (PY) in the target (*Candesartan*) and comparator (*Chlorthalidone*) group after stratification, as well as the minimum detectable relative risk (MDRR). Note that the IR does not account for any stratification.

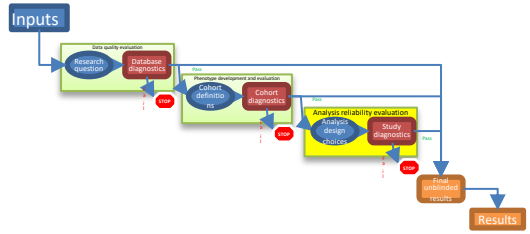
Source	Target subjects	Comparator subjects	Target years	Comparator years	Target events	Comparator events	Target IR (per 1,000 PY)	Comparator IR (per 1,000 PY)	MDRR
Optum	4,510	7,682	3,394	5,037	<5	<5	<1.47	<0.99	>6.27
CCAE	4,897	14,092	4,179	8,519	0	<5	0.00	<0.59	>17.53
Panther	3,148	15,105	877	5,626	0	<5	0.00	<0.89	>27.56

Empirical Equipoise: Preference score



- Randomized clinical trials assign treatment with each subject having the same probability of being each intervention
  - A 1:1 randomized head-to-head trial gives all subjects a 50% chance of being assigned to the target exposure and 50% chance of being assigned to the comparator, regardless of patient/provider characteristics
  - Randomization allows for assumption that persons assigned to target cohort are exchangeable at baseline with persons assigned to comparator cohort
- Non-interventional studies involve observing treatment choices, which can be influenced by patient or provider characteristics
  - Comparator selection is a pre-analysis design choice
  - Preference = probability of patient choosing target vs. comparator treatment, given baseline features
  - Preference = 50% means indifference between treatments for a patient, akin to random assignment
- Potential **pre-adjustment design diagnostic**: what proportion of the target population is close to treatment indifference?

Empirical Equipoise: Preference score



Comparative Effectiveness Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

A tool for assessing the feasibility of comparative effectiveness research

Alexander M Walker[1]
Amanda R Patrick[2]
Michael S Lauer[3]
Mark C Hornbrook[4]
Matthew G Marin[5]
Richard Platt[6]
Véronique L Roger[7]
Paul Stang[8]
Sebastian Schneeweiss[2]

[1]World Health Information Science Consultants, Newton, MA; [2]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, MA; [3]National Heart, Lung, and Blood Institute, National Institutes of Health,

This article was published in the following Dove Press journal:
Comparative Effectiveness Research
29 January 2013
Number of times this article has been viewed

Background: Comparative effectiveness research (CER) provides actionable information for health care decision-making. Randomized clinical trials cannot provide the patients, time horizons, or practice settings needed for all required CER. The need for comparative assessments and the infeasibility of conducting randomized clinical trials in all relevant areas is leading researchers and policy makers to non-randomized, retrospective CER. Such studies are possible when rich data exist on large populations receiving alternative therapies that are used as-if interchangeably in clinical practice. This setting we call "empirical equipoise."

Objectives: This study sought to provide a method for the systematic identification of settings it in which it is empirical equipoise that offers promised non-randomized CER.

Methods: We used a standardizing transformation of the propensity score called "preference" to assess pairs of common treatments for uncomplicated community-acquired pneumonia and new-onset heart failure in a population of low-income elderly people in Pennsylvania, for whom we had access to de-identified insurance records. Treatment pairs were considered suitable for CER if at least half of the dispensings of each treatment-pair member fell within a preference range of 30% to 70%.

Methods

A prioritization tool

We propose the following algorithm.

1. Identify an environment with longitudinal health care data for a large population in which CER may be relevant. The

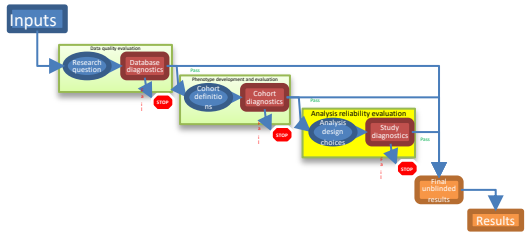
...

5. Accept drug pairs as emerging from empirical equipoise if at least half of the dispensings of each of the drugs are to patients with a preference score of between 0.3 and 0.7.

and health characteristics. This preference score was obtained by subtracting the natural logarithm of Treatment A prevalence divided by Treatment B prevalence from the logit of the propensity score, and taking the anti-logit (expit) of the result. In the resulting equation (Equation 1), in the universe of persons receiving either Treatment A or B, F and S are the preference score and propensity score for receiving Treatment A, respectively, and P is the fraction of persons receiving Treatment A:

$$\ln\left(\frac{F}{1-F}\right) = \ln\left(\frac{S}{1-S}\right) - \ln\left(\frac{P}{1-P}\right) \qquad (1)$$

Empirical Equipoise: Preference score Examples from LEGEND-HTN



Good:
T = valsartan
C = olmesartan
DB = CCAE

Even with >40,000 patients on each drug, large-scale propensity score model could not meaningfully discriminate between the two treatments; >90% of persons in 'empirical equipoise' with a preference score between 0.3 and 0.7

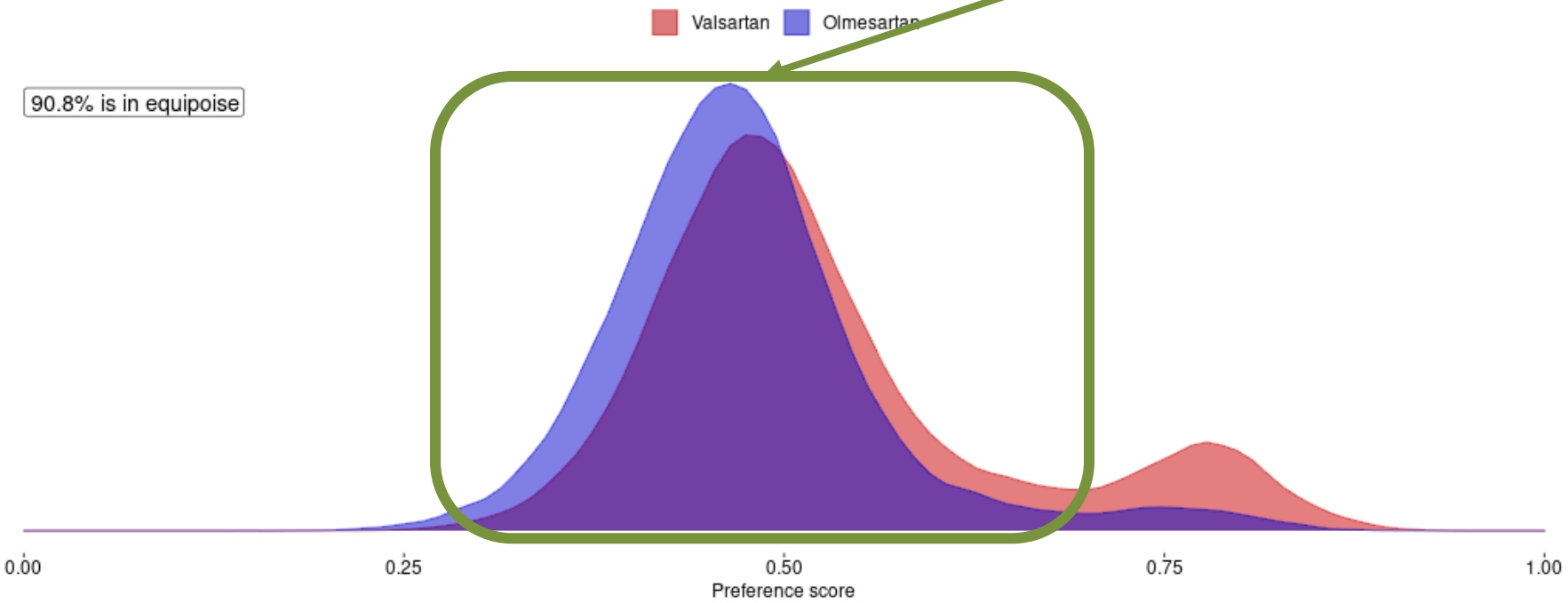
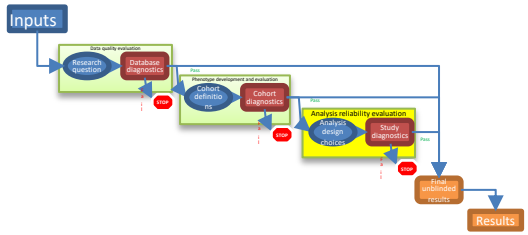


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Empirical Equipoise: Preference score Examples from LEGEND-HTN



Bad:
T = valsartan
C = chlorthalidone
DB = CCAE

Baseline characteristics can clearly discriminate most new users of valsartan vs. chlorthalidone; <30% of persons in 'empirical equipoise' with a preference score between 0.3 and 0.7

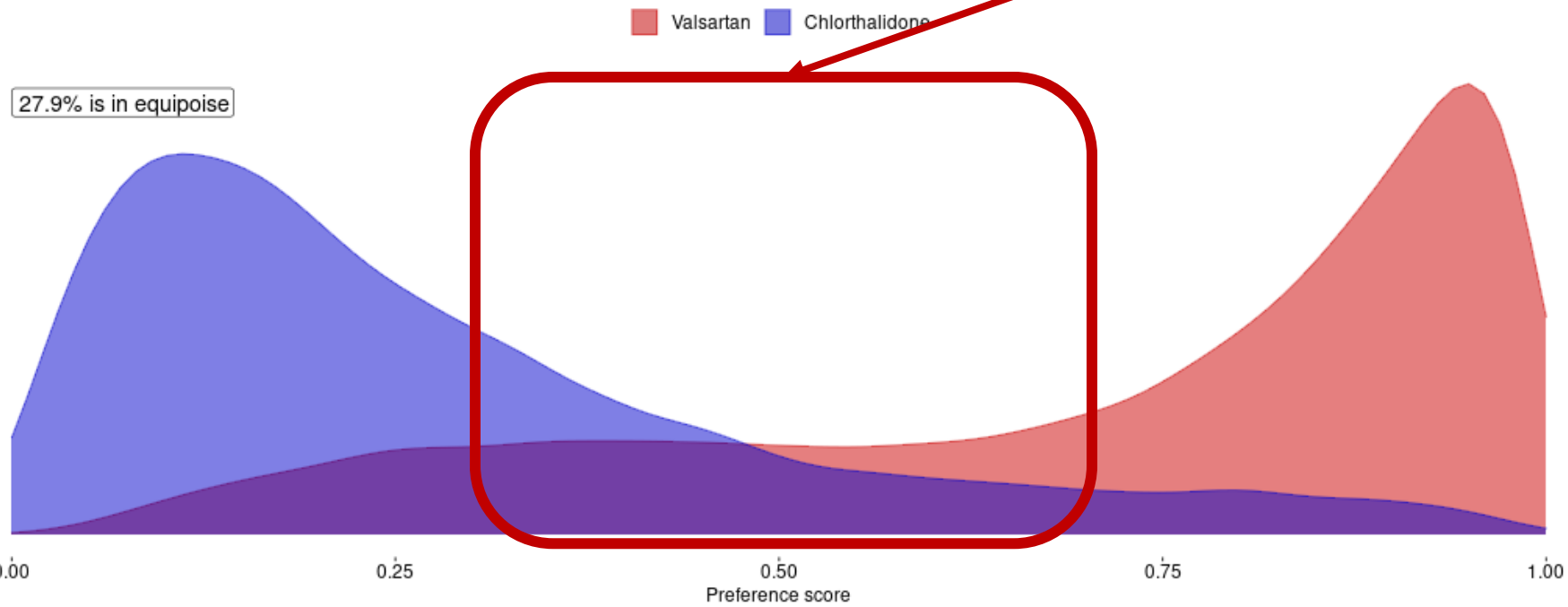
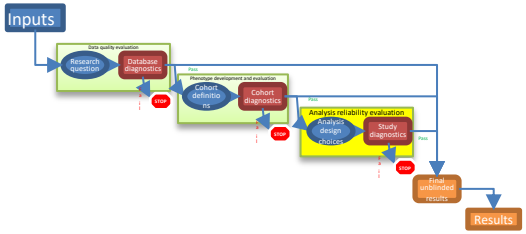
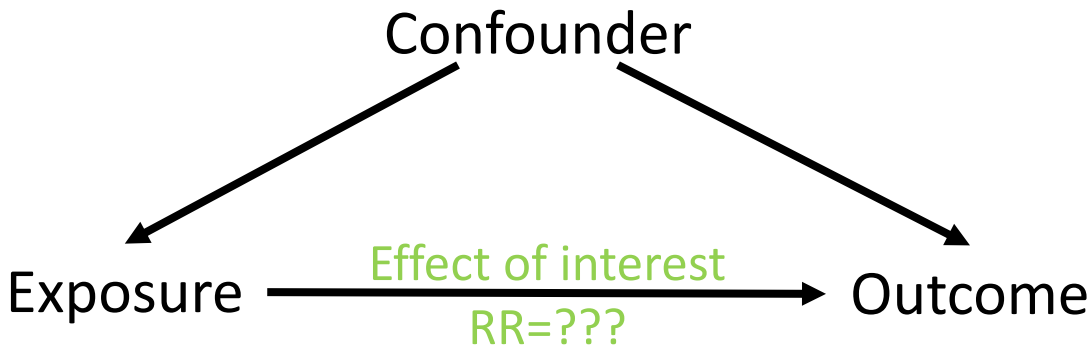


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

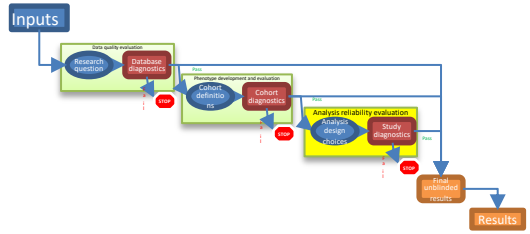
Covariate balance: Standardized mean difference





- Confounding variables associated with both exposure and outcome can bias effect estimates if not properly addressed
- Various design and analysis choices (restriction, matching, propensity score adjustment) offer strategies to reduce the effect of confounding by balancing confounder prevalence in target and comparator cohort
- Potential **post-adjustment analytic diagnostic**: are all observed baseline characteristics sufficiently similar between target and comparator cohorts?

Covariate balance: Standardized mean difference



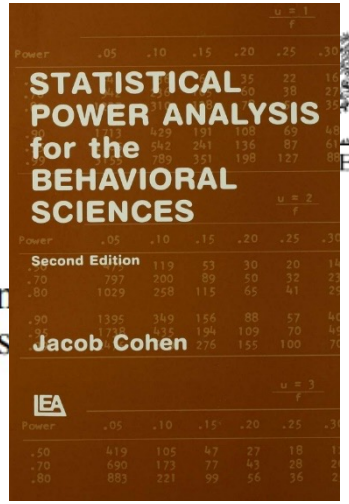
STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3083–3107
Published online 15 September 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3697

Balance diagnostics for comparing the distribution covariates between treatment groups in propens matched samples

Peter C. Austin[1, 2, 3, *, †]

[1]*Institute for Clinical Evaluative Sciences,*
[2]*Dalla Lana School of Public Health*
[3]*Department of Health Policy, Manage*

Standardized differences are increasingly between treated and untreated subjects in their use is lack of consensus as to what residual imbalance between treated and unt no clear consensus on this issue, some resea 0.1 (10 per cent) denotes meaningful imbal threshold for acceptable imbalance depends covariate in question. Ho *et al.* suggest tha important covariates than for weak predicto



The terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation (see Sections 1.4 and 11.1). In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available.

SMALL EFFECT SIZE: **d** = .2. In new areas of research inquiry, effect sizes are likely to be small (when they are not zero!). This is because the phenomena under study are typically not under good experimental or measurement control or both. When phenomena are studied which cannot be brought into the laboratory, the influence of uncontrollable extraneous variables ("noise") makes the size of the effect small relative to these (makes the "signal" difficult to detect).

ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 387–398

Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores

Sharon-Lise T. Normand[a,b,*], Mary Beth Landrum[a], Edward Guadagnoli[a], John Z. Ayanian[a,c], Thomas J. Ryan[d], Paul D. Cleary[a,c], Barbara J. McNeil[a,f]

[a]*Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, 02115-5899, USA*
of Biostatistics, Harvard School of Public Health, Boston, MA, USA
Social Medicine, Harvard School of Public Health, Boston, MA, USA
Boston University School of Medicine, Boston, MA, USA
ine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
of Radiology, Brigham and Women's Hospital, Boston, MA, USA

1 2000; received in revised form 8 July 2000; accepted 8 August 2000

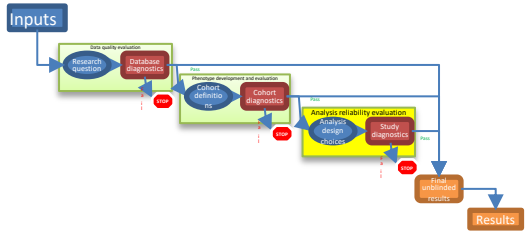
2.4.2. *Quantifying bias*

We assessed the degree of imbalance within the matched pairs as well as between the matched and unmatched patients using the standardized difference, d_i, in covariate means [21]. For each covariate, we calculated

$$d_i = 100 \times (x_{ci} - x_{nci}) / \sqrt{\{(s^2_{ci} + s^2_{nci})/2\}}$$

where x_{ci} and x_{nci} are the sample means in the catheterized and noncatheterized groups of the ith covariate, respectively, and s^2_{ci} and s^2_{nci} are the corresponding sample variances. Small (<10%) absolute values of d_i support the assumption of balance [25] between treatment groups.

Covariate balance: Standardized mean difference Examples from LEGEND-HTN



Good:
T = amlodipine
C = atenolol
A = PS matching, on-treatment
DB = CCAE

>45,000 baseline covariates evaluated, many with SMD > 0.1 before matching, but after matching all covariates have SMD <= 0.03

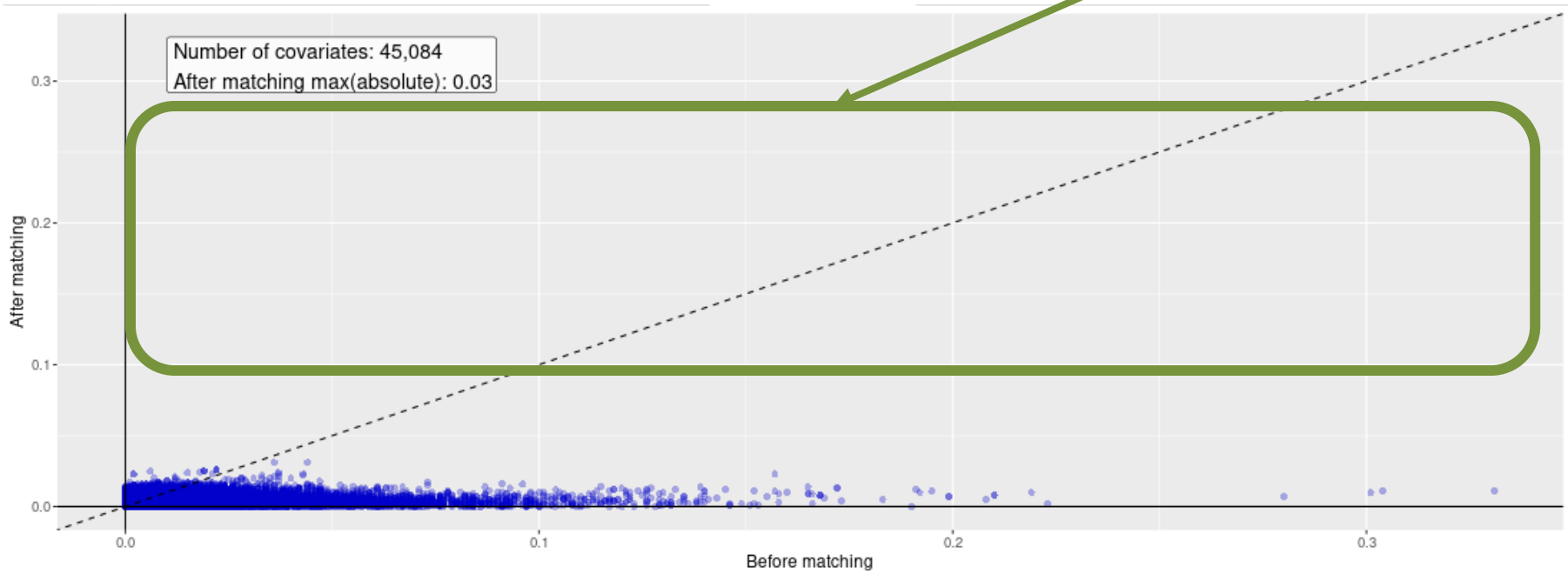
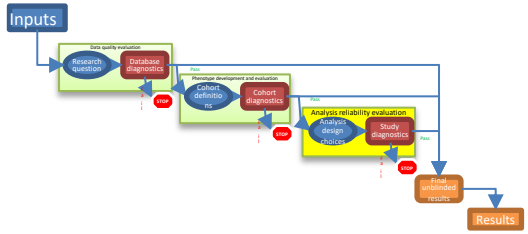


Figure 3. Covariate balance before and after matching. Each dot represents the standardized difference of means for a single covariate before and after matching on the propensity score. Move the mouse arrow over a dot for more details.

Covariate balance: Standardized mean difference Examples from LEGEND-HTN



Bad:
T = candesartan
C = atenolol
A = PS stratification, on-treatment
DB = CCAE

>50,000 baseline covariates evaluated, many with SMD > 0.1 before stratification. After stratification, many covariates have higher SMD than pre-stratification, many covariates with SMD > 0.1

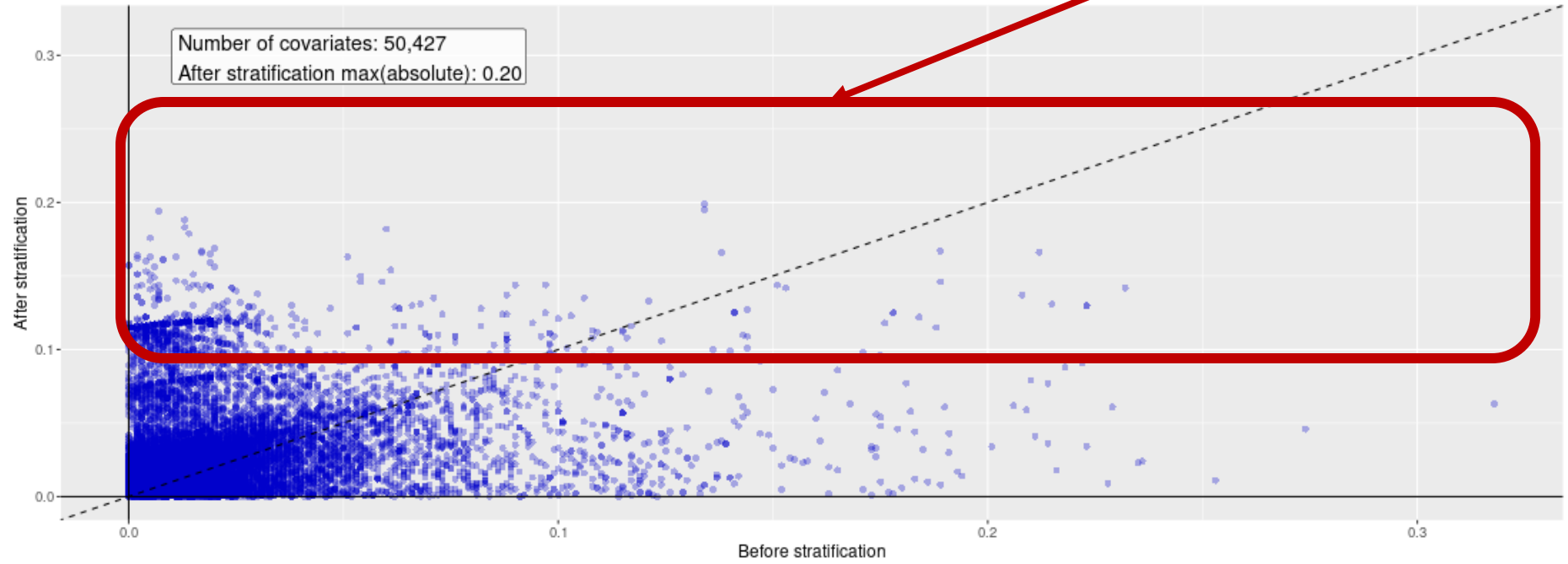
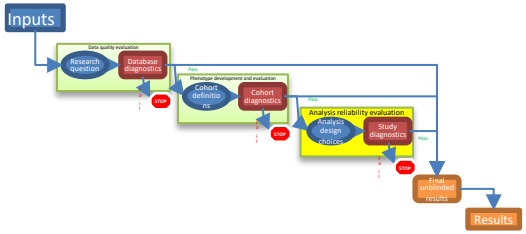


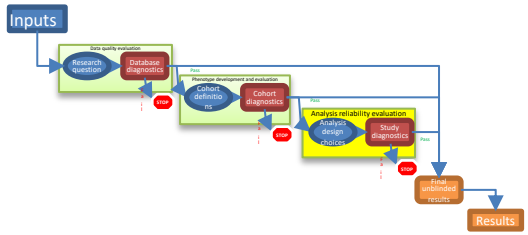
Figure 3. Covariate balance before and after stratification. Each dot represents the standardized difference of means for a single covariate before and after stratification on the propensity score. Move the mouse arrow over a dot for more details.

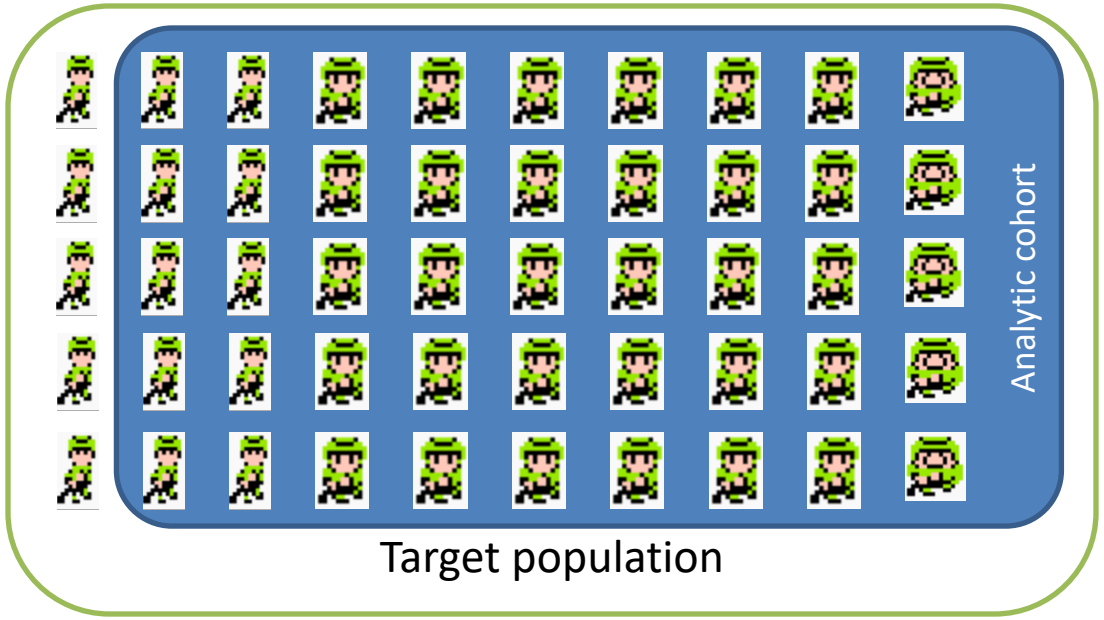
Generalizability: Standardized mean difference

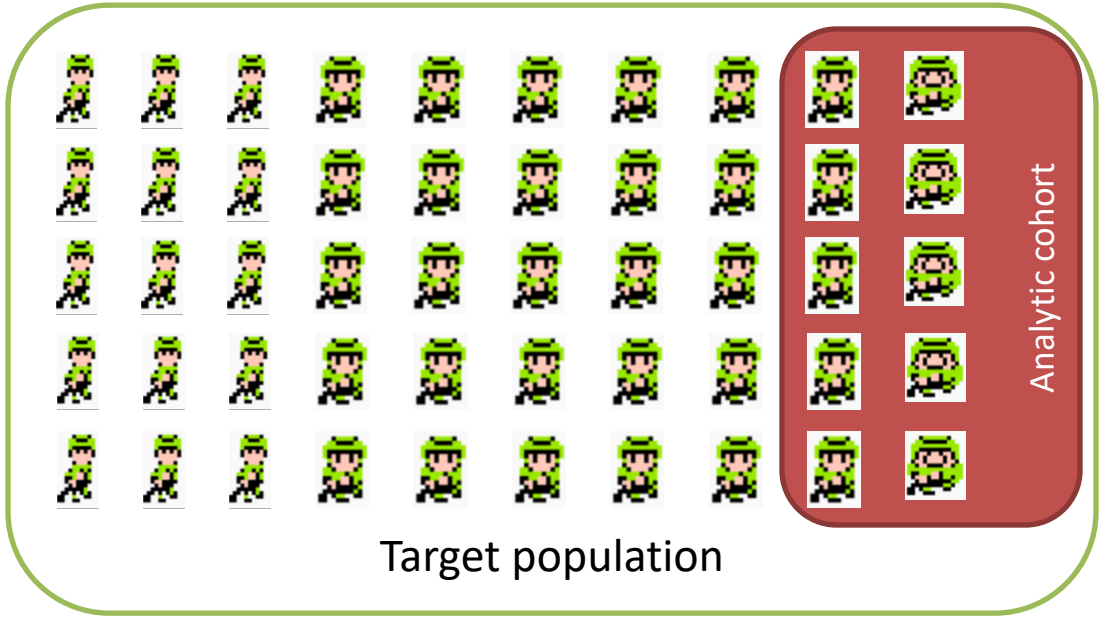


- Generalizability is the extent to which a study result can be applied to a target population of interest
- The same design and analytic strategies employed to reduce confounding (such as restriction, matching, propensity score adjustment) can potentially shift the composition of the analytic cohort away from the original target
- Potential **post-adjustment analytic diagnostic**: are all observed baseline characteristics sufficiently similar between the pre-adjustment target and post-adjustment analytic cohorts?

Generalizability: Standardized mean difference



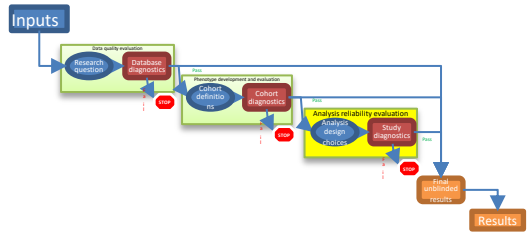




	Target	Analytic	SMD
	30%	22%	0.18
	60%	67%	-0.15
	10%	11%	-0.03
Total	50 (100%)	45 (90%)	

	Target	Analytic	SMD
	30%	0%	0.93
	60%	50%	0.20
	10%	50%	-0.97
Total	50 (100%)	10 (20%)	

Generalizability: Standardized mean difference



Article

Evaluation Review
2017, Vol. 41(5) 472-505

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]

Assessment

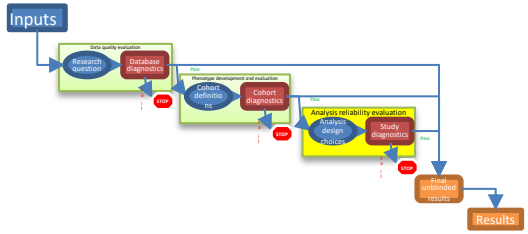
Statistics for assessing generalizability. When assessing generalizability, three global measures and one covariate-level measure are often used. We begin here with the covariate-level measure—the ASMD defined for each covariate X_j ($j = 1, \ldots, p$) as:

$$|d_j| = \left| \frac{\bar{X}_{jS} - \mu_{jP}}{\sigma_{jP}} \right|, \quad (1)$$

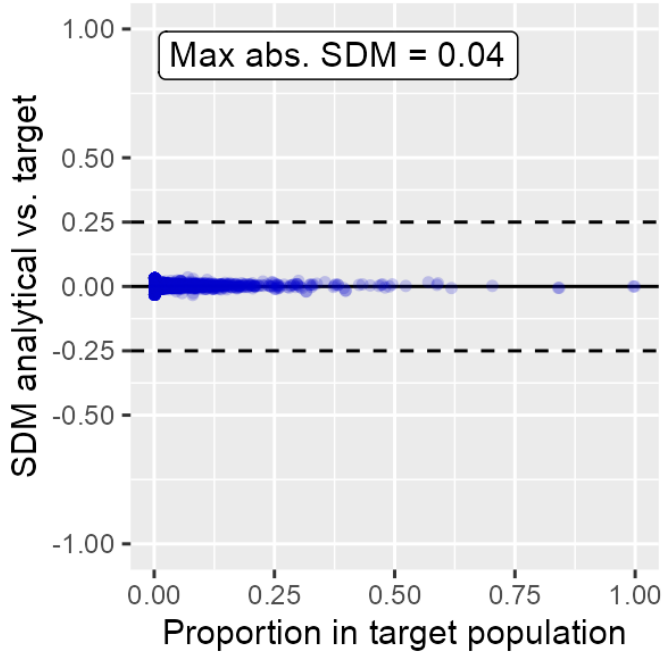
where $\bar{X}_{jS}$ is the mean in the sample, μ_{jP} is the population mean, and σ_{jP} is the population standard deviation. This is calculated for each of the p covariates included in the propensity score. We focus here on the absolute value since, following the literature, we are rarely interested in direction but instead in magnitude. When assessing similarity, researchers are often interested in both covariate by covariate comparisons (each value of $|d_j|$) and aggregates of these (e.g., $|d| = \sum_{j=1}^{p} \Sigma |d_j|/p$).

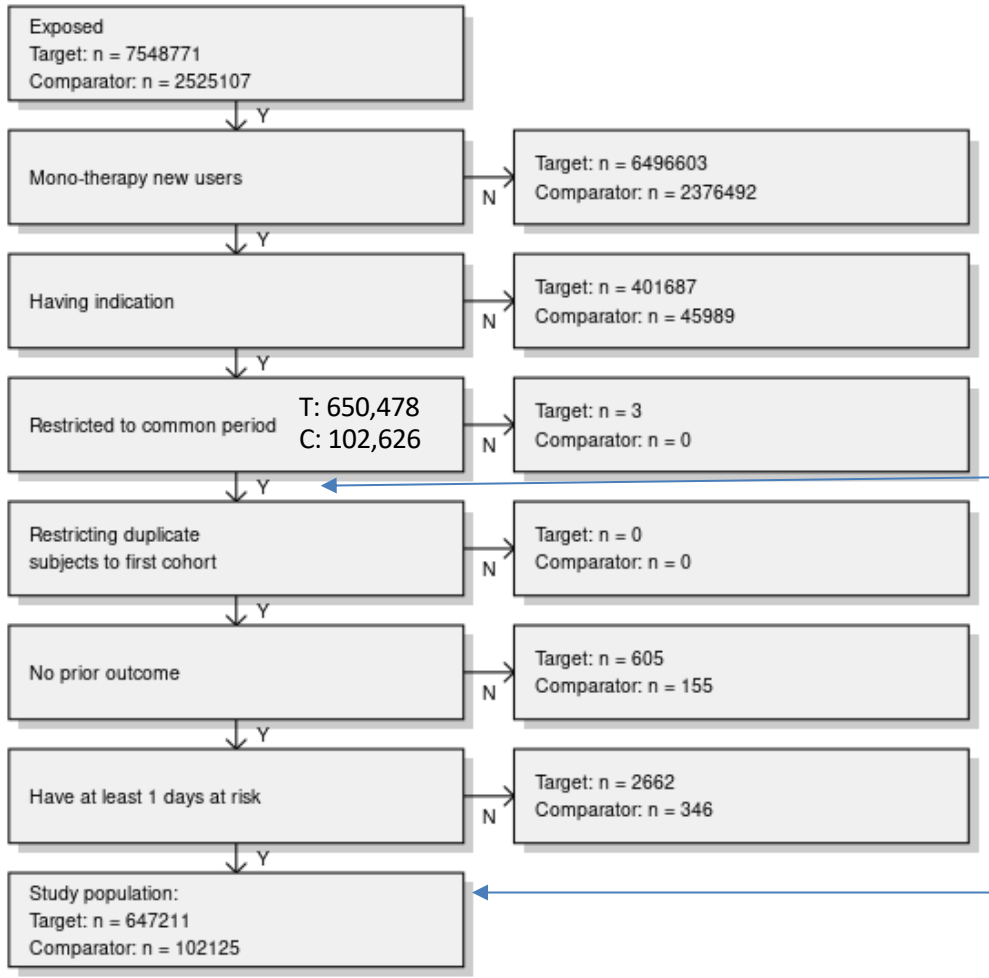
Each of these four statistics can then be used to determine if the sample is similar enough to the population (on observables) to warrant generalization of the experimental findings. For the ASMD and the logit SMD, researchers have borrowed rules of thumb generated in observational studies, with similarity achieved when the values are smaller than 0.25 (Rubin, 2001) or 0.10 (e.g., Normand et al., 2001). For the generalizability index, Tipton

Generalizability: Standardized mean difference



Good:
T = lisinopril
C = losartan
O = angioedema
A = PS stratification, on-treatment
DB = CCAE



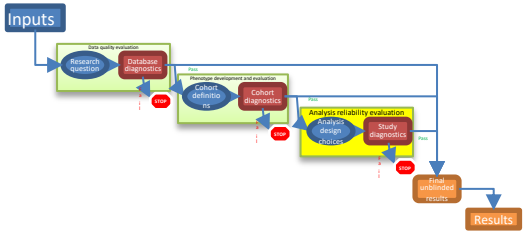


'Target' cohort

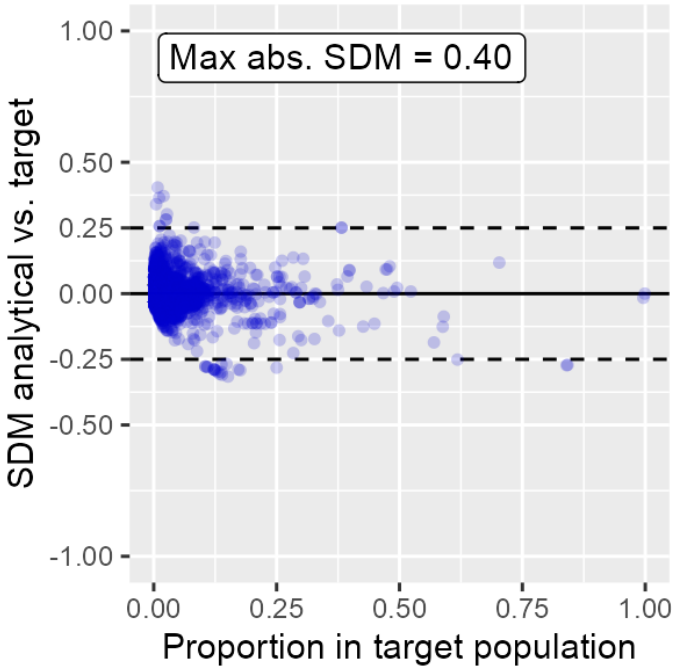
>99% of target population remains in analysis cohort

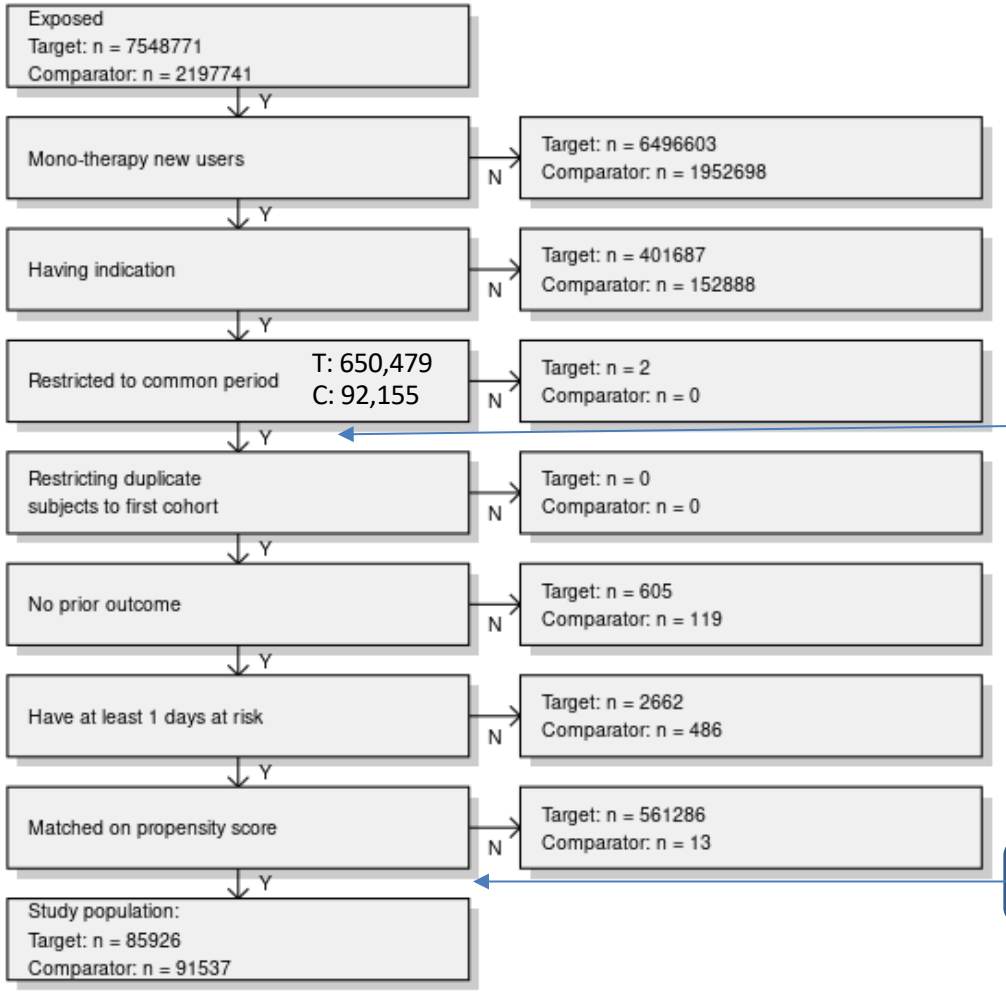
'Analysis' cohort

Generalizability: Standardized mean difference



Bad:
T = lisinopril
C = atenolol
O = angioedema
A = PS matching, on-treatment
DB = CCAE



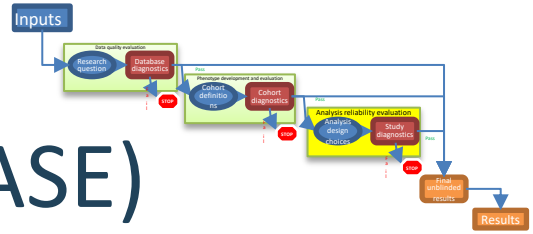


'Target' cohort

Only 13% of target population remains in analysis cohort

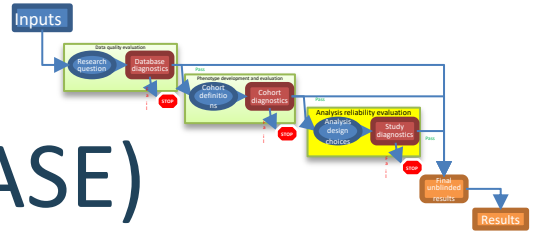
'Analysis' cohort

Residual bias: Expected Absolute Systematic Error (EASE)



- Design and analysis choices aim to produce unbiased estimates, but residual systematic error can 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 bias:
Expected Absolute Systematic Error (EASE)

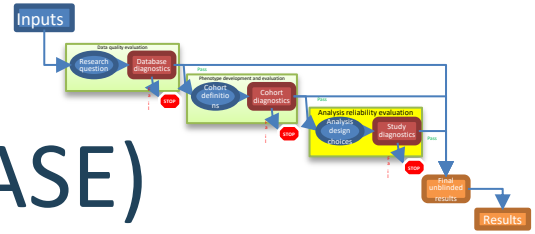


You trust your scale to estimate an accurate weight. If it's off by a couple pounds, you may think it's 'good enough' since its probably directionally correct and you can adjust the weight by how much you think the scale is miscalibrated.

But what if you weighed a standard 100lb weight and your scale read 126.4?

Would you 'calibrate' by adjusting your own weight by 26 lbs or would you dismiss the scale's estimate altogether?

Residual bias: Expected Absolute Systematic Error (EASE)



COLLOQUIUM PAPER

STATISTICS

MEDICAL SCIENCES

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 same question. One source of discrepancies is error, both random caused by sampling variability and systematic (for example, because of confounding, selection bias, and measurement error). Only random error is typically quantified but converges to zero as databases become larger, whereas systematic error persists independent from sample size and therefore, increases in relative importance. Negative controls are exposure–outcome pairs, where one believes no causal effect exists; they can be used to detect multiple sources of systematic error, but interpreting their results is not always straightforward. Previously, we have shown that an empirical null distribution can be derived from a sample of negative controls and used to calibrate *P* values, accounting for both random and systematic error. Here, we extend this work to calibration of confidence intervals (CIs). CIs require positive controls, which we synthesize by modifying negative controls. We show that our CI calibration restores nominal characteristics, such as 95% coverage of the true effect size by the 95% CI. We furthermore show that CI calibration reduces disagreement in replications of two pairs of conflicting observational studies: one related to dabigatran, warfarin, and gastrointestinal bleeding and one related to selective serotonin reuptake inhibitors and upper gastrointestinal bleeding. We recommend CI calibration to improve reproducibility of observational studies.

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 for specific sources of bias, there has been comparatively little work in devising approaches to empirically estimate the magnitude of systematic error or clinical applications that show how to integrate this error into effect estimation methods.

The acuity of this problem is only exacerbated as the size of observational databases grow: random error (the only component that is typically quantified) converges to zero as sample size increases, but systematic error persists independent from sample size. Some sources of systematic error may potentially increase if expanding the size of a data source comes with compromise in the depth or quality of the data captured. Therefore, the hype of "big data" has brought with it an increased number of studies with vanishingly narrow CIs, while our collective uncertainty about the accuracy of any given observational estimate has steadily increased. While we expect an accurate 95% CI to have a 95% coverage probability—the proportion of time that an interval contains the true value of interest—we have little empirical evidence to support that observational estimates exhibit this basic, nominal operating characteristic.

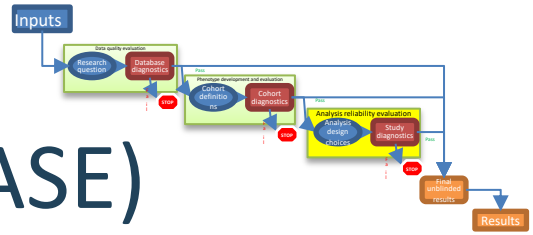
A promising development toward better explication of systematic error has been recent proposals and examples to apply negative controls as a diagnostic tool or "falsification hypothesis" (2–4). Negative controls are exposure–outcome pairs where one believes no causal effect exists. Executing a study on negative con-

CI Calibration. For CI calibration, we build on our previous work in calibrating *P* values (10). Using the computed effect size estimates for the negative and positive controls, we observe to what extent random error alone explains the difference between the estimates and their true effect sizes. Systematic error explains any additional difference. We fit a systematic error model using the effect size estimates for the controls and subsequently use this model to compute calibrated CIs for the effect sizes of interest. The model assumes that systematic error follows a Gaussian probability distribution around the true effect size. We have found that a Gaussian distribution provides a good approximation, and more complex models, such as mixtures of Gaussians and nonparametric density estimation, do not improve results. Let $\hat{\theta}_i$ denote the computed log effect estimate (e.g., hazard ratio) from the ith negative or positive control, and let $\hat{\tau}_i$ denote its corresponding estimated SE for $i = 1, \ldots, n$. Let θ_i denote the true log effect size, and let β_i denote the asymptotic bias associated with pair i: specifically, the difference between the log of the true effect size and the log of the estimate that the study would have returned for control i had it been infinitely large. As in the standard CI computation, we assume that $\hat{\theta}_i$ is normally distributed with mean $\theta_i + \beta_i$ and variance $\hat{\tau}_i^2$. Note that the traditional CI calculation always assumes $\beta_i = 0$ but that we assume that β_i for all i arises from a normal distribution with mean $\mu(\theta_i)$ and SD $\sigma(\theta_i)$ that follow linear models, after appropriate transformation, with unknown intercepts a and c and slopes b and d, respectively. Specifically, we model

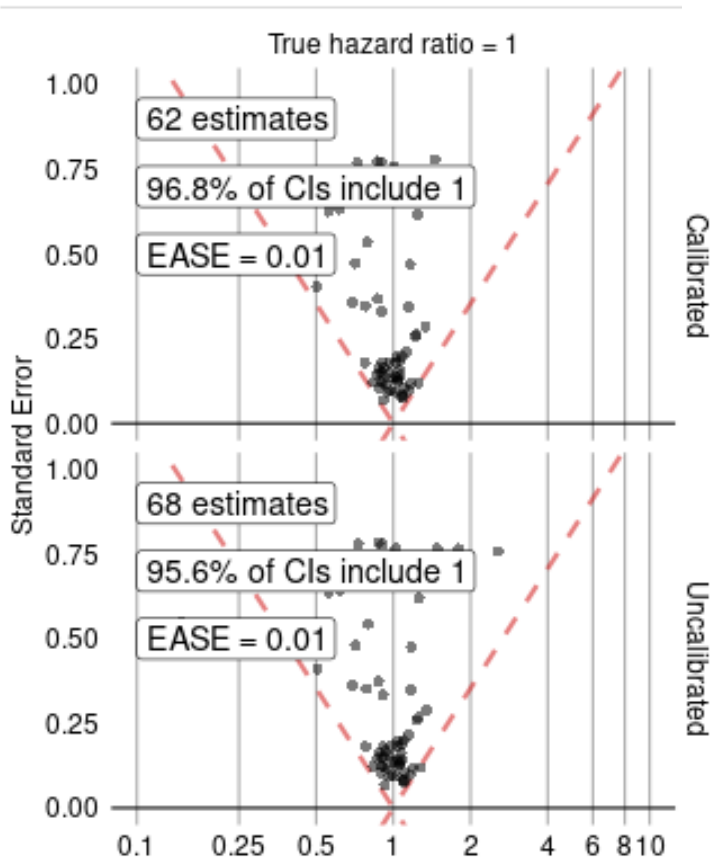
$$\beta_i \sim N(\mu(\theta_i), \sigma^2(\theta_i)) \text{ and}$$
$$\hat{\theta}_i \sim N(\theta_i + \beta_i, \hat{\tau}_i^2), \qquad [1]$$

Expected absolute systematic error (EASE) = $|\beta_i|$

Residual bias: Expected Absolute Systematic Error (EASE)



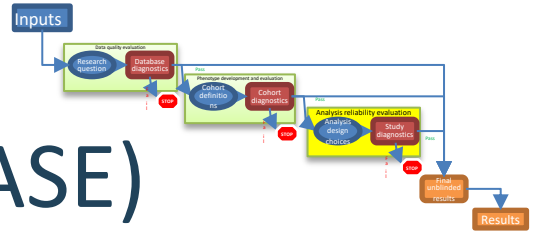
Good:
T = hydrochlorothiazide
C = chlorthalidone
O = acute myocardial infarction
A = PS stratification, on-treatment
DB = CCAE



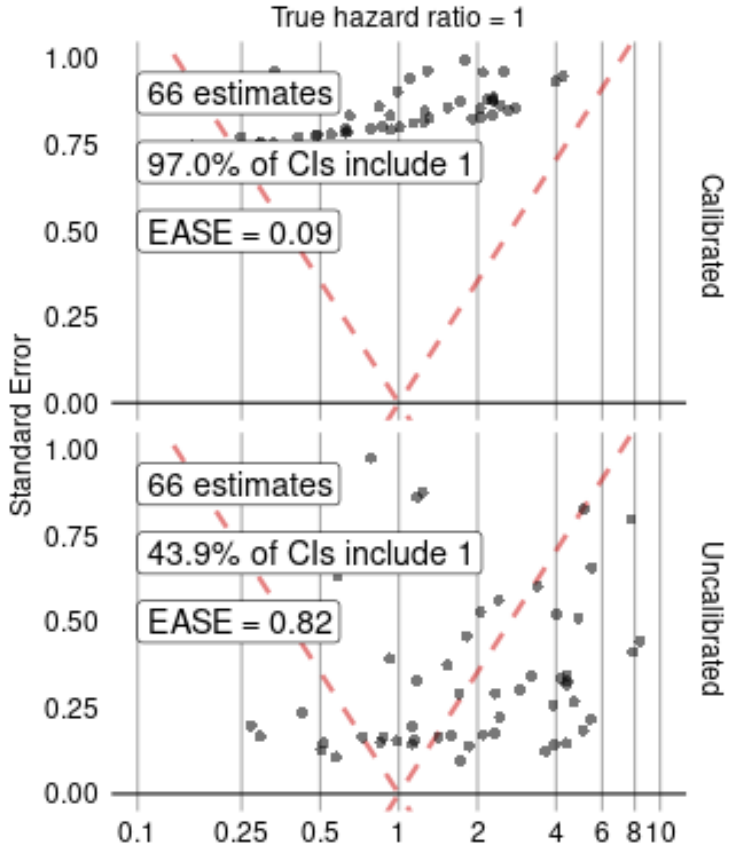
Little residual bias observed (EASE=0.01), so calibration has very little impact on effect estimate (HR=1.54 → HR=1.51)

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	CCAE	1.54	0.88	3.00	0.17	1.51	0.82	2.79	0.18

Residual bias: Expected Absolute Systematic Error (EASE)



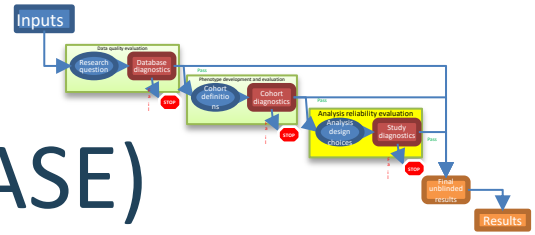
Bad:
T = furosemide
C = labetalol
O = acute myocardial infarction
A = PS stratification, on-treatment
DB = CCAE



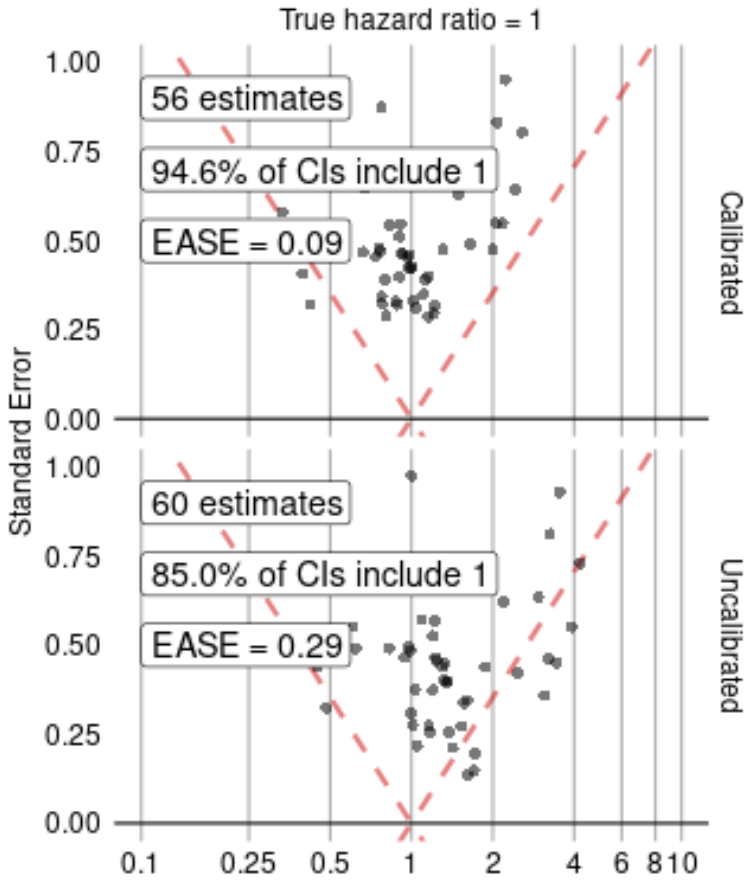
Substantial positive bias and variance observed (EASE=0.82), so calibration has substantial impact on effect estimate (HR=5.55, p<0.01 → HR=2.86, p<0.20)

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	CCAE	5.55	3.50	9.25	0.00	2.86	0.59	17.78	0.20

Residual bias: Expected Absolute Systematic Error (EASE)



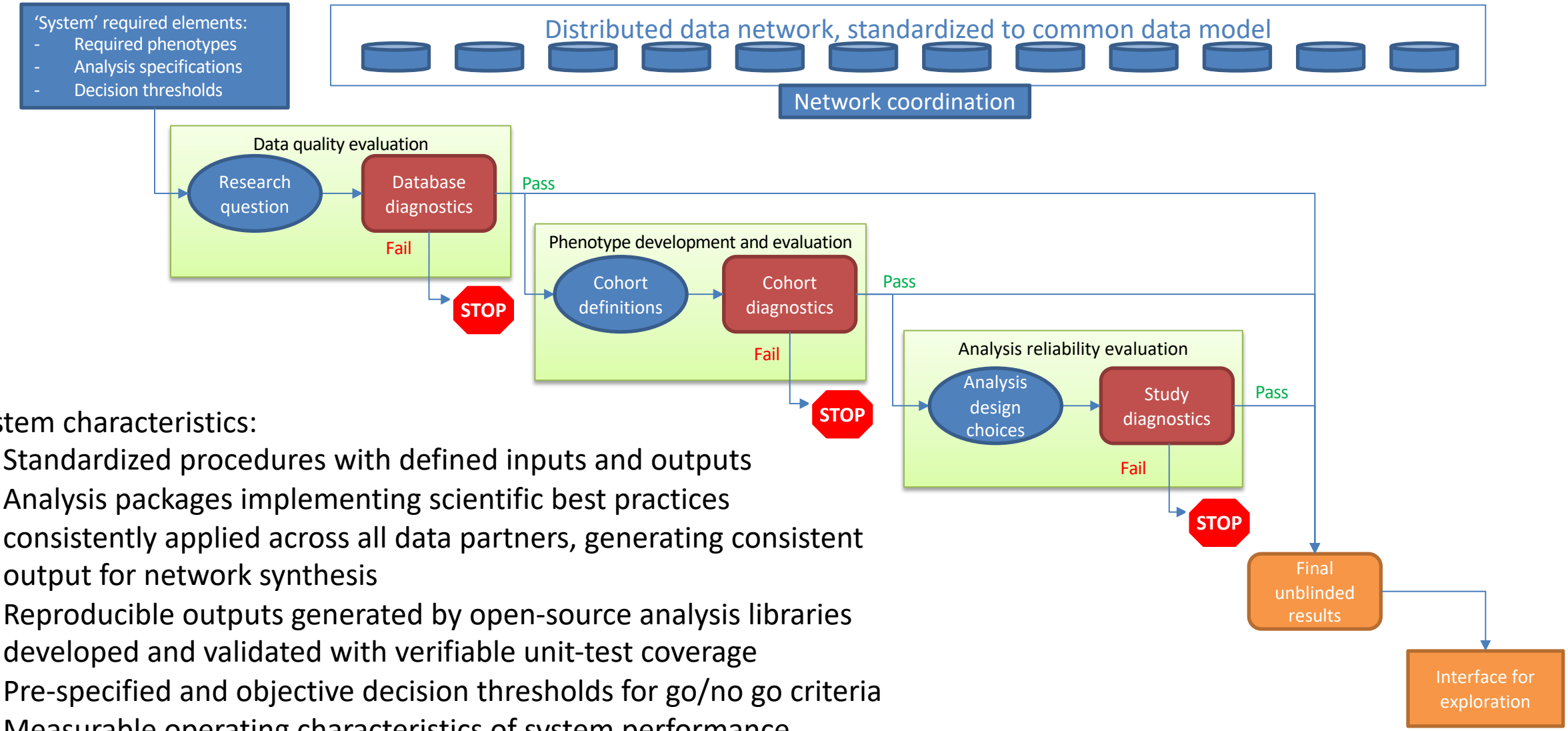
Maybe?
T = furosemide
C = chlorthalidone
O = abdominal pain
A = PS stratification, on-treatment
DB = Optum EHR



Some positive bias and variance observed (EASE=0.29), so calibration has some impact on effect estimate (HR=1.64 → HR=1.17) and changes statistical significance

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	Panther	1.64	1.34	2.02	0.00	1.17	0.76	2.25	0.25

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

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?

An empirical evaluation of study diagnostics

LEGEND viewer

LEGEND Basic Viewer

About | Specific research questions

Indication

Hypertension

Exposure group

Drug major class

☐ Include combination exposures

Target

ACE inhibitors

Comparator

Angiotensin receptor blockers (ARBs)

Outcome

Angioedema

Data source

- ☑ CCAE
- ☑ CUMC
- ☑ IMSG
- ☑ JMDC
- ☑ MDCD

Show 15 entries

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS stratification, on-treatment	CCAE	3.96	3.09	5.17	0.00	3.80	2.65	6.01	0.00
PS stratification, on-treatment	CUMC	1.60	0.54	6.88	0.47	1.61	0.51	5.66	0.40
PS stratification, on-treatment	MDCD	9.77	3.09	59.27	0.00	8.76	2.10	NA	0.00
PS stratification, on-treatment	MDCR	3.03	1.83	5.35	0.00	3.26	1.81	6.63	0.00
PS stratification, on-treatment	NHIS_NSC	0.15	NA	2.62	NA	NA	NA	NA	NA
PS stratification, on-treatment	Optum	3.18	2.45	4.20	0.00	3.14	2.21	4.79	0.00
PS stratification, on-treatment	Panther	3.78	2.37	6.45	0.00	2.80	1.66	7.45	0.00
PS matching, on-treatment	CCAE	4.36	2.94	6.70	0.00	4.44	2.71	7.68	0.00
PS matching, on-treatment	CUMC	3.00	0.38	60.62	0.39	2.98	0.22	NA	0.41
PS matching, on-treatment	MDCD	14.00	2.81	NA	0.02	15.88	1.58	NA	0.02
PS matching, on-treatment	MDCR	7.93	2.76	33.48	0.00	9.59	2.42	NA	0.00
PS matching, on-treatment	NHIS_NSC	NA	NA	NA	NA	NA	NA	NA	NA
PS matching, on-treatment	Optum	3.48	2.23	5.65	0.00	3.55	2.12	6.23	0.00
PS matching, on-treatment	Panther	2.64	1.36	5.51	0.01	2.07	1.14	4.56	0.01

Showing 1 to 14 of 14 entries Previous 1 Next

Power | Attrition | Population characteristics | Propensity scores | Covariate balance | Systematic error | Kaplan-Meier

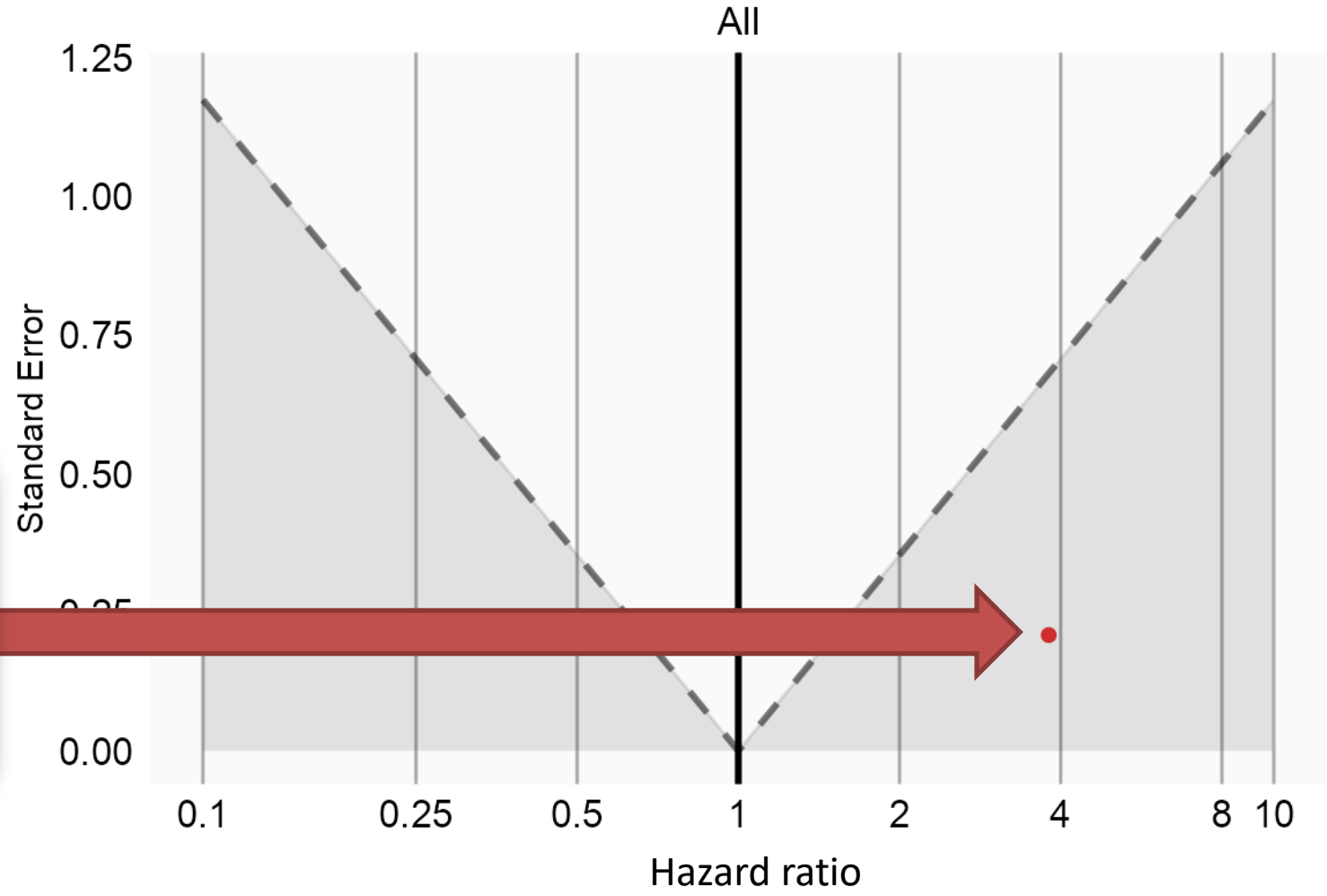
Table 1a. Number of subjects, follow-up time (in years), number of outcome events, and event incidence rate (IR) per 1,000 patient years (PY) in the target (*ACE inhibitors*) and comparator (*Angiotensin receptor blockers (ARBs)*) group after stratification, as well as the minimum detectable relative risk (MDRR). Note that the IR does not account for any

https://data.ohdsi.org/LegendBasicViewer/

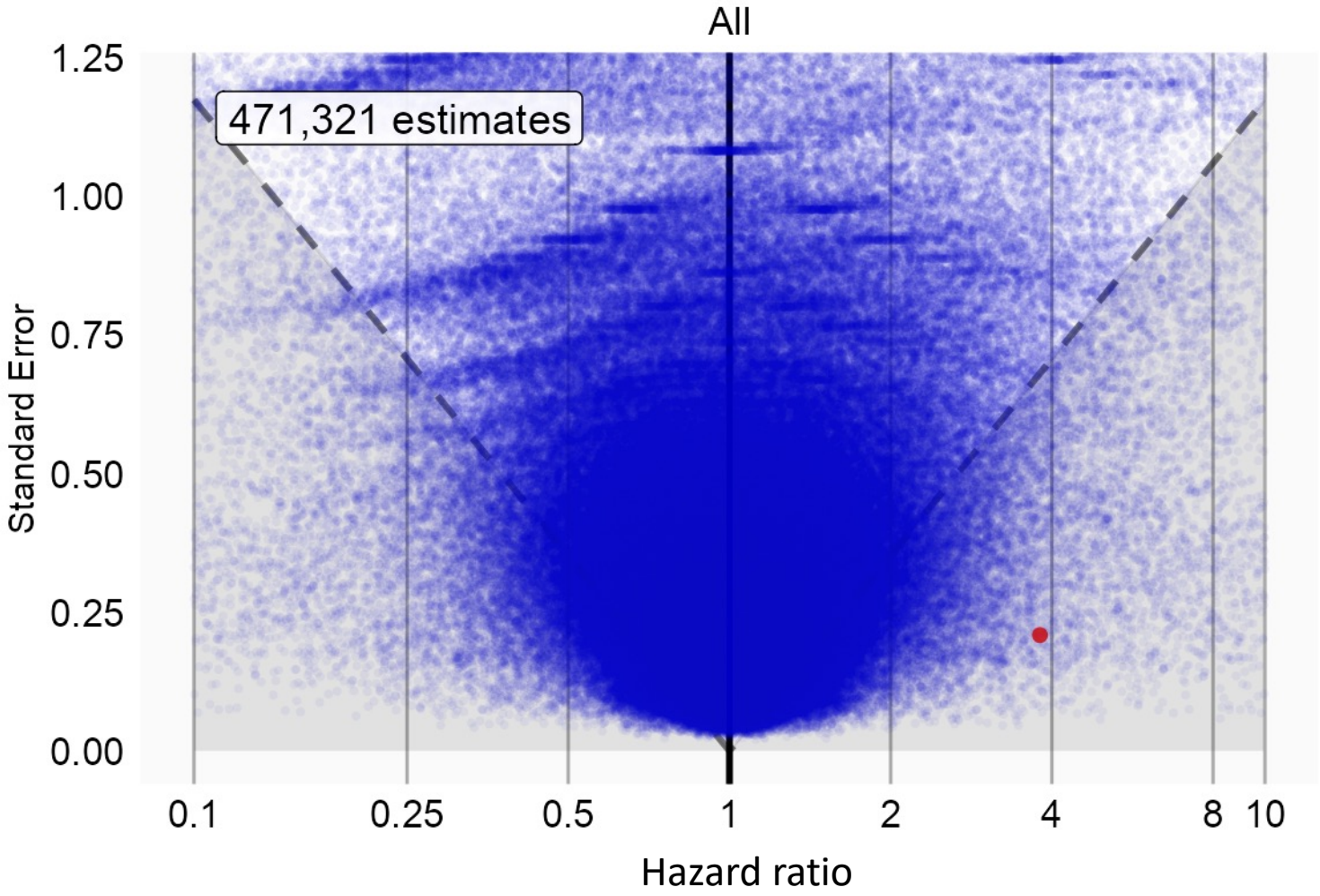
LEGEND estimates

ACEs vs ARBs for angioedema using PS stratification in CCAE

Calibrated HR = 3.80 (2.65-3.80)



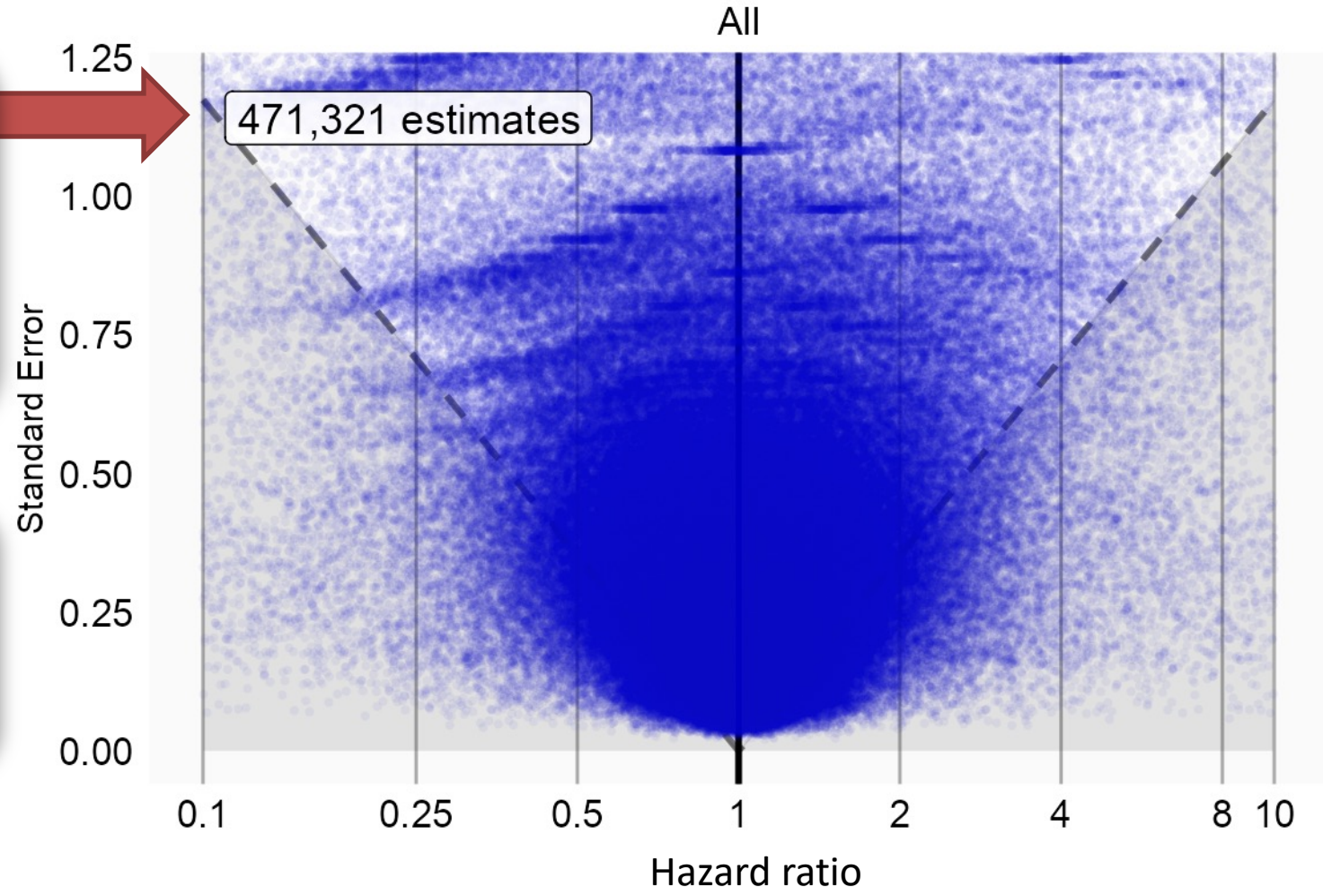
LEGEND estimates



LEGEND estimates

Showing all 471,321 calibrated estimates from LEGEND Hypertension (restricting to mono-therapy comparisons only, using on-treatment time-at-risk)

It is good that we see no evidence of publication bias or p-hacking, but is this otherwise good or bad?



LEGEND estimates where no effect is expected

- Hypertension medications are well studied
- 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 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

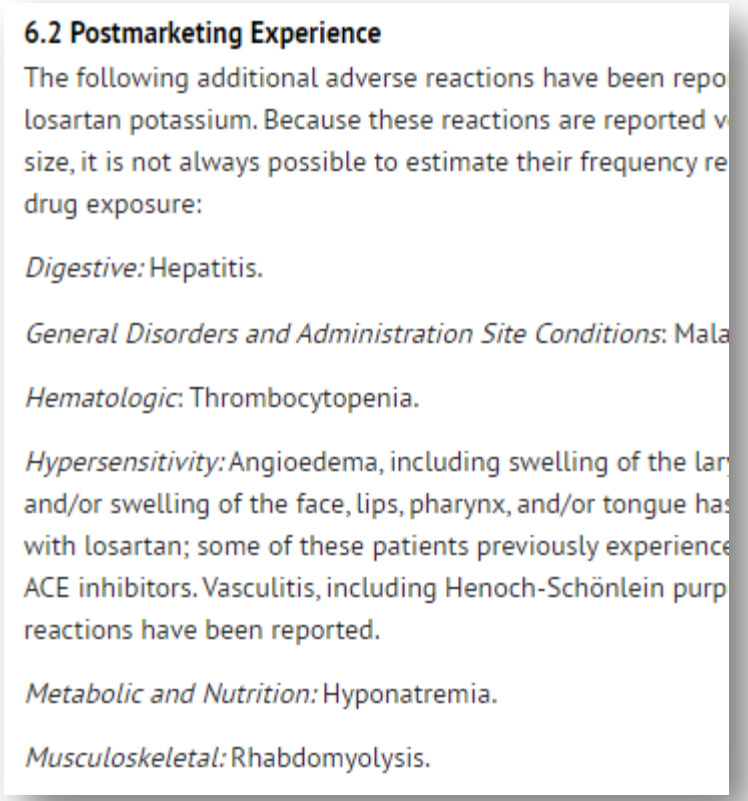
Patients taking concomitant mTOR inhibitor (e.g. temsir
neprilysin inhibitor may be at increased risk for angioed

Hydralazine (vasodilator) vs **prazosin** (alpha blocker) for **angioedema** is **likely null**

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.

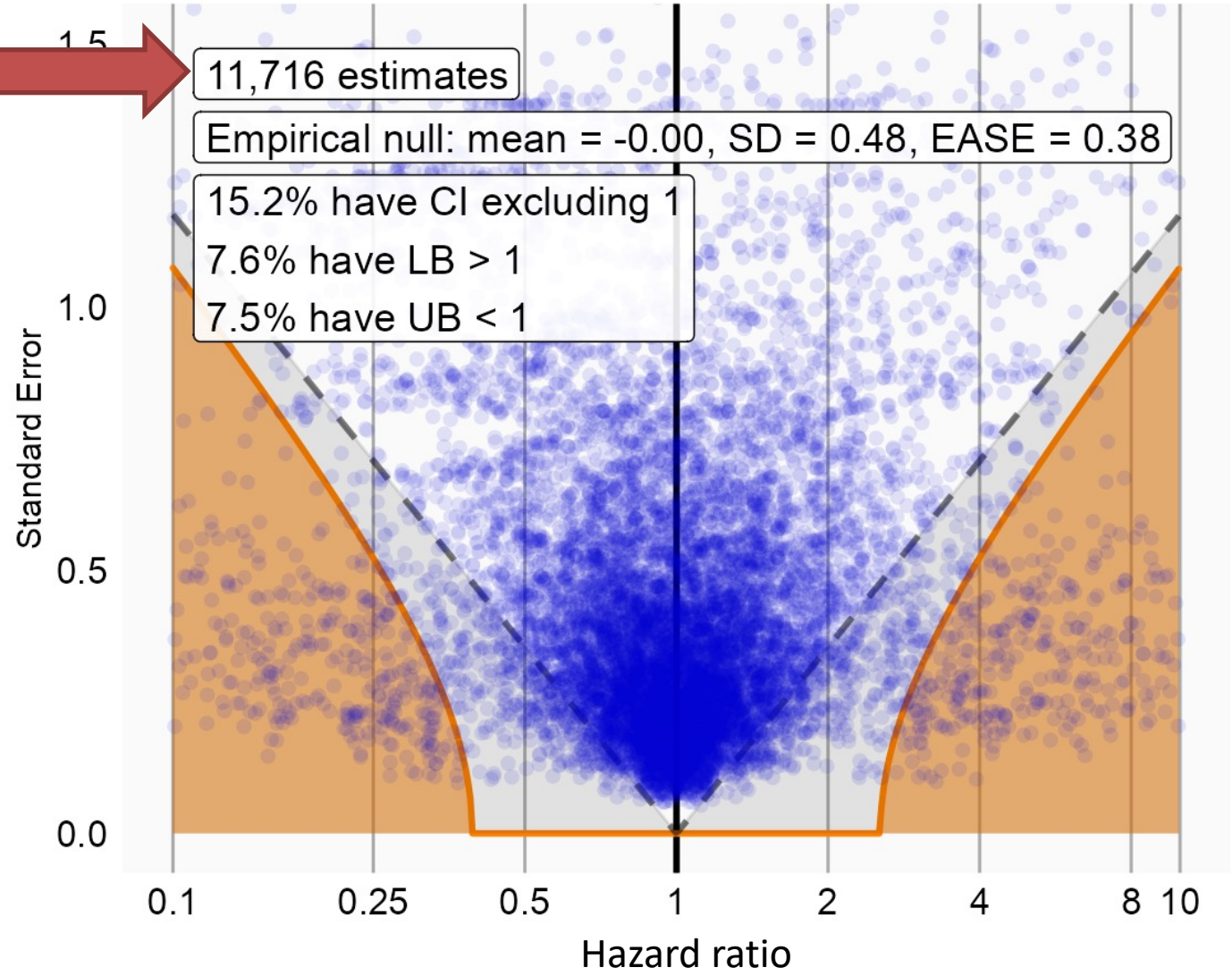
Musculoskeletal: Rhabdomyolysis.

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 more similar to the outcomes of interest: better exchangeability?
  - Using full outcome phenotypes instead of ‘occurrence of concept’

LEGEND estimates when null is likely true

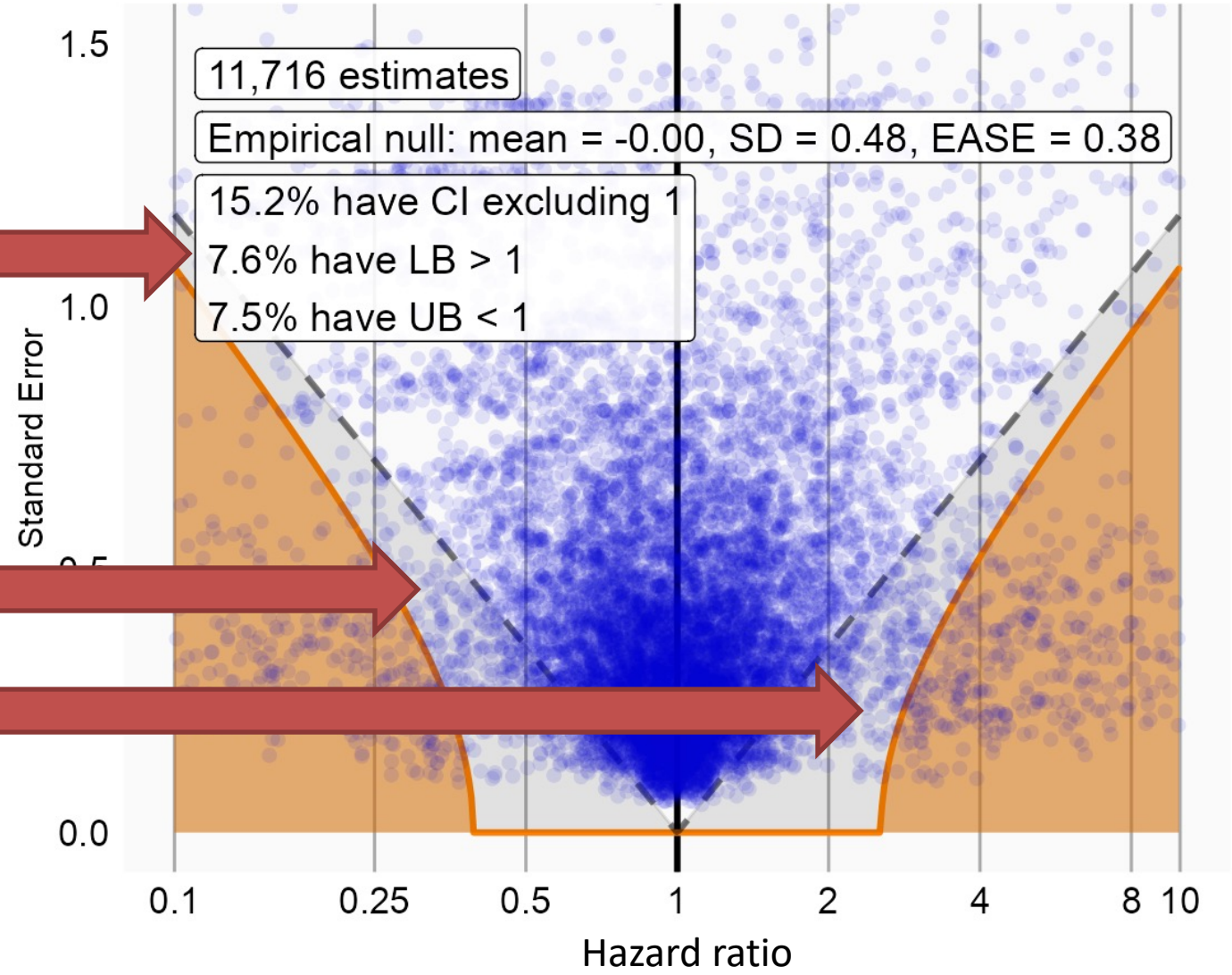
Showing all 11,716 calibrated estimates from LEGEND Hypertension where believe the null to be true



LEGEND estimates when null is likely true

If the null is true for all, we would expect 5% of CIs to not include 1

Estimates below the dashed line have confidence interval (CI) excluding 1

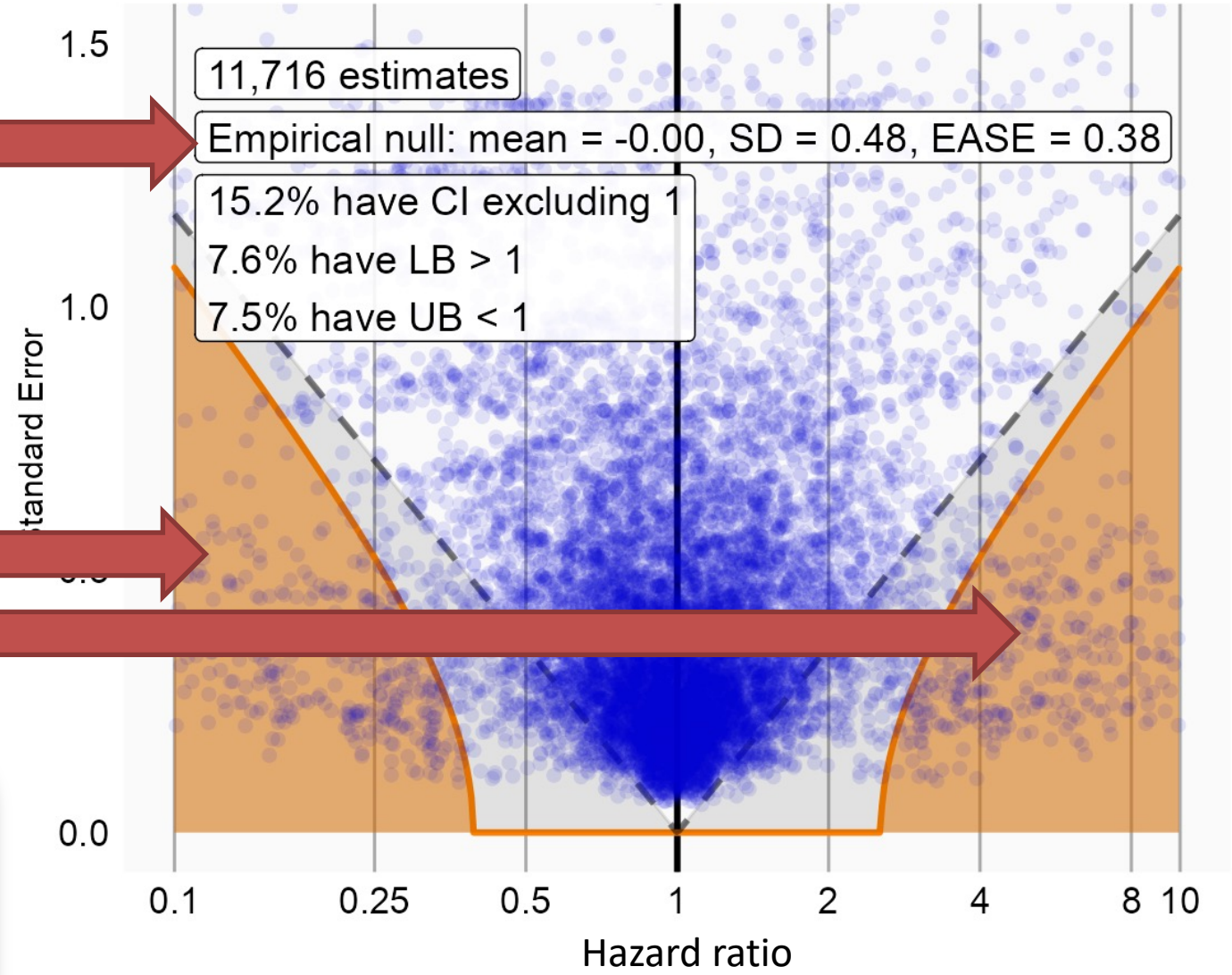


LEGEND estimates when null is likely true

A more accurate way to look at deviation from the null is the fitted parameters of the null distribution. We can summarize these as EASE.

Fitted null distribution also visualized as orange areas, where (newly) calibrated CI doesn't include 1

EASE = 0.0 means it seems the null is true for all, and there is only random error (as expressed in the CIs), no systematic error.



Evaluating the effect of diagnostics rules

Statistical power

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

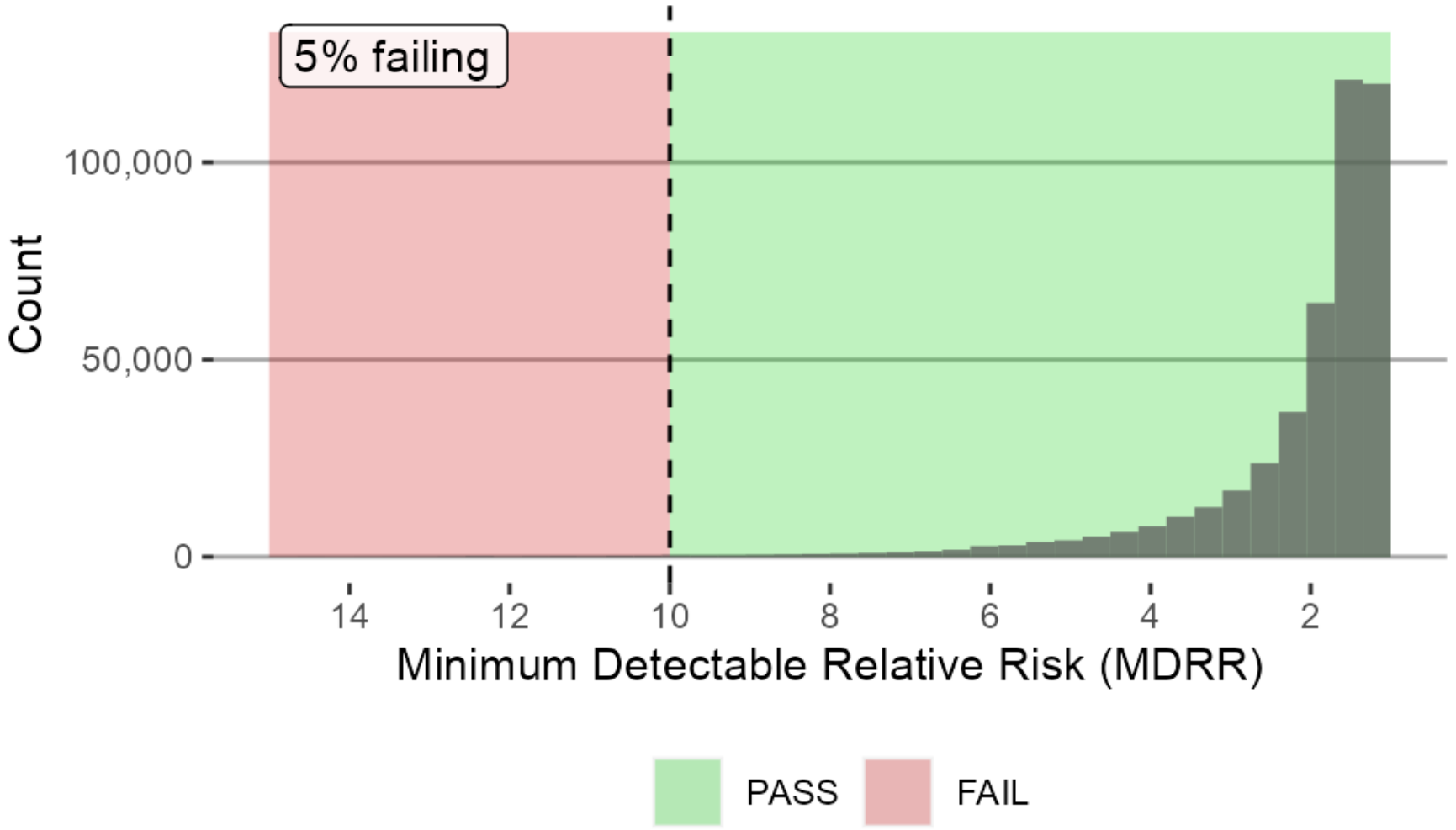
Reasoning:

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

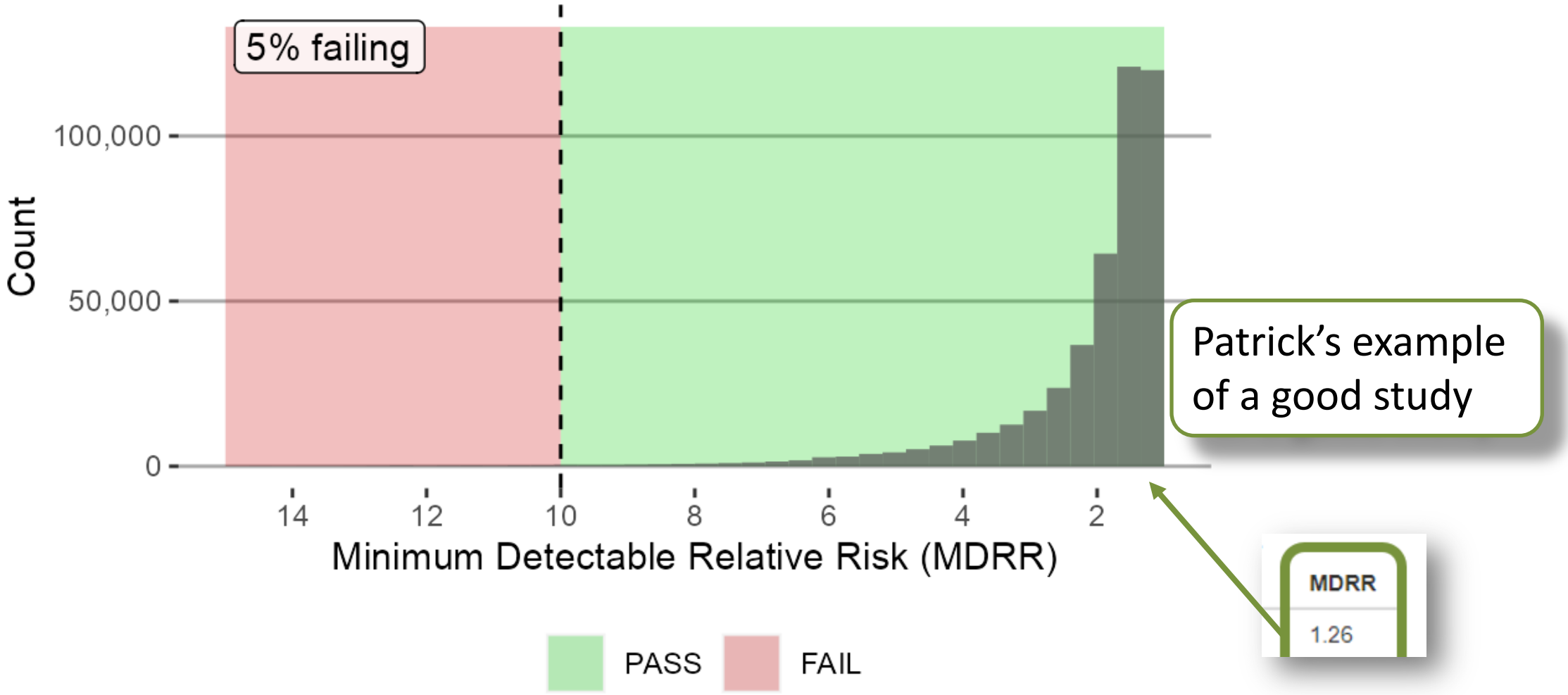
Note:

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

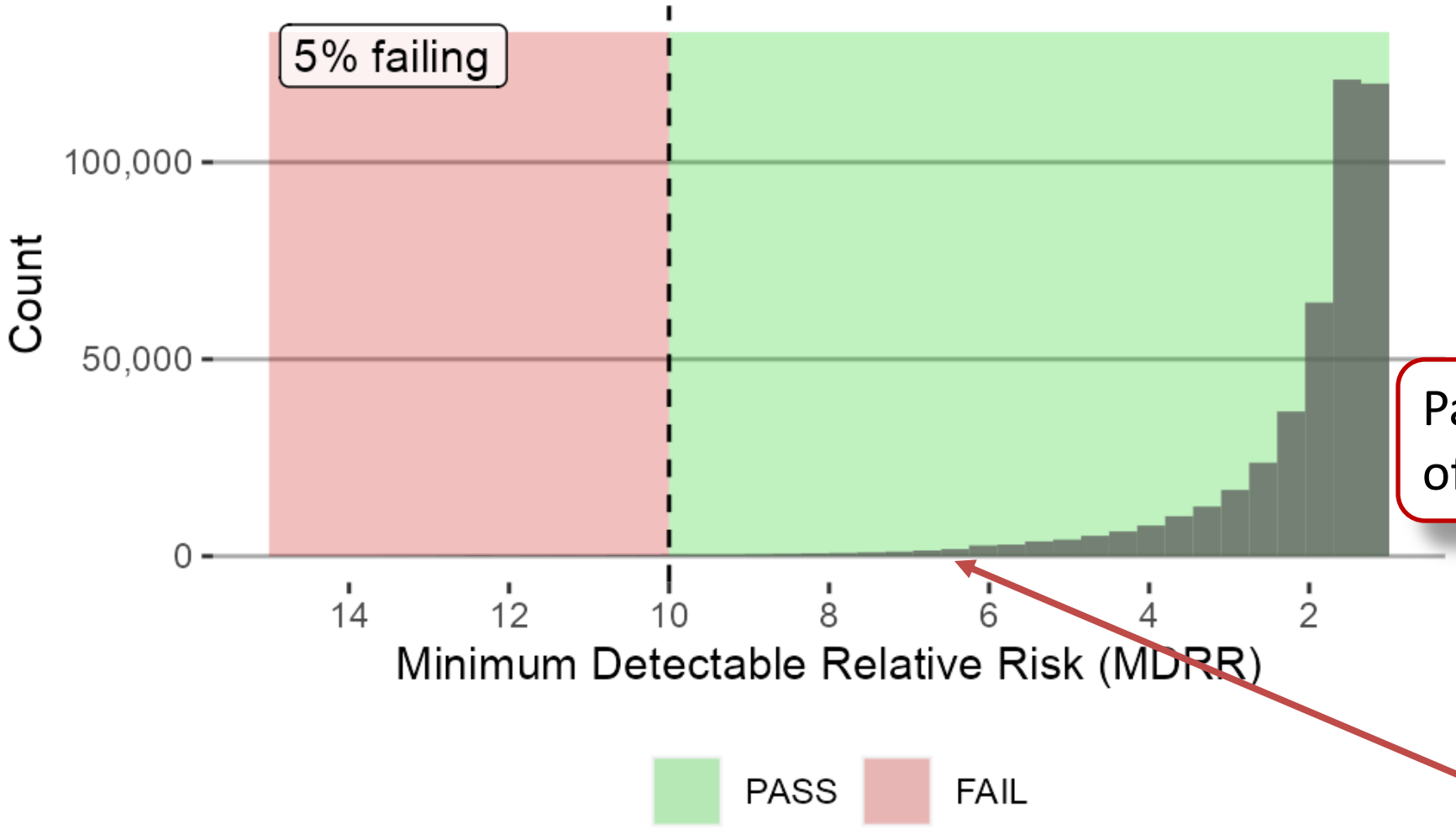
Statistical power



Statistical power

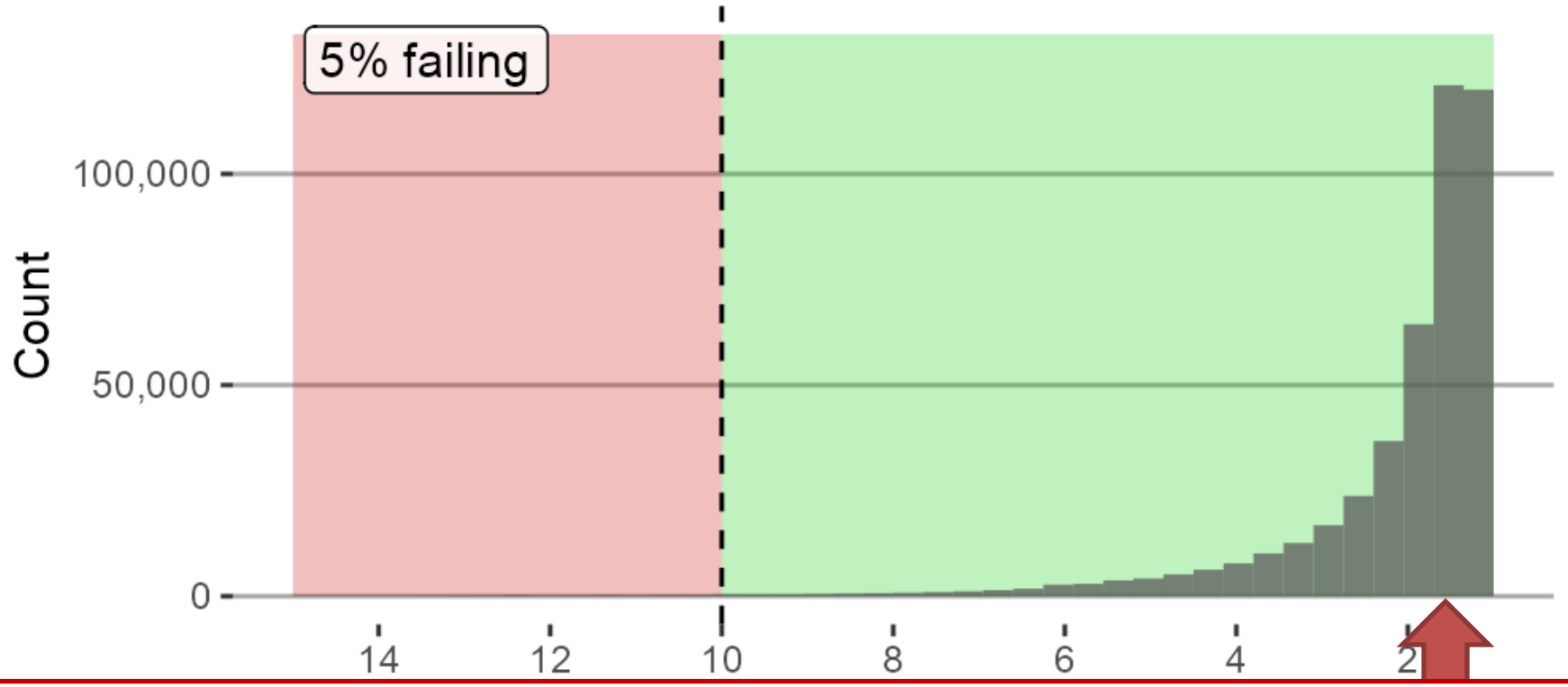


Statistical power



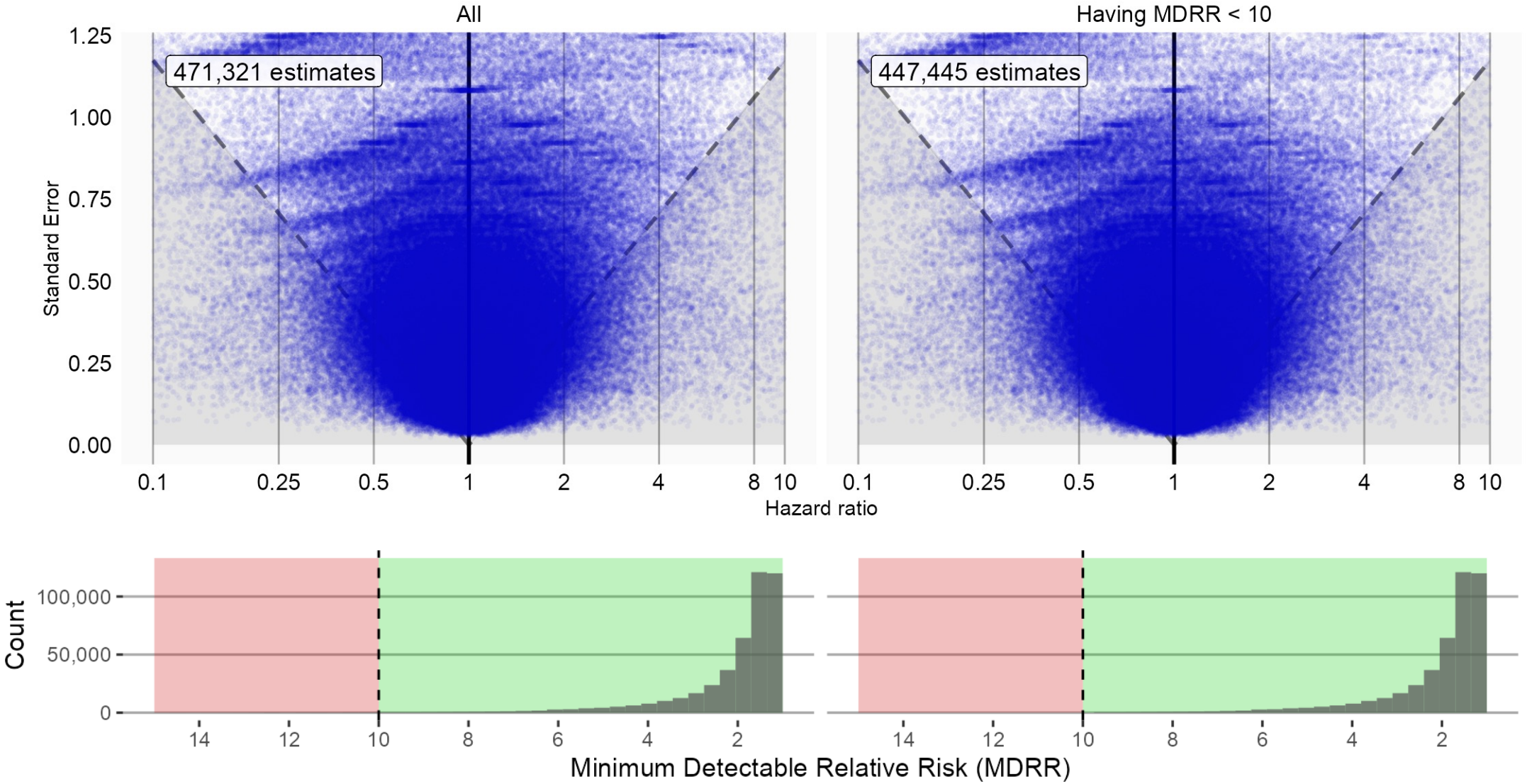
Patrick's example of a bad study

Statistical power

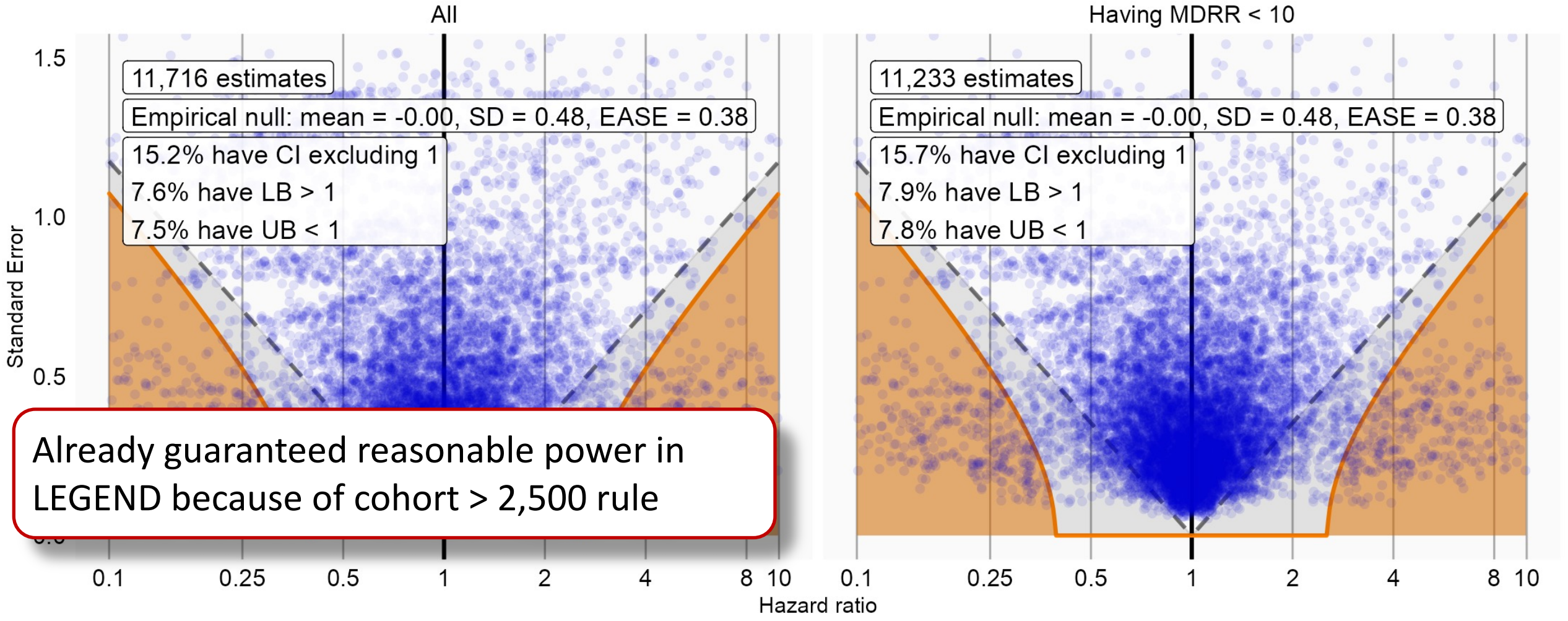


Most comparative effects in antihypertensives have HR < 2.
We're not ensuring we are powered to answer real questions. (we are)
We're trying to avoid reporting hard-to-interpret estimates. (e.g. HR = 5.1 (0.7-36.2))

Statistical power



Statistical power



Already guaranteed reasonable power in LEGEND because of cohort > 2,500 rule

Equipoise

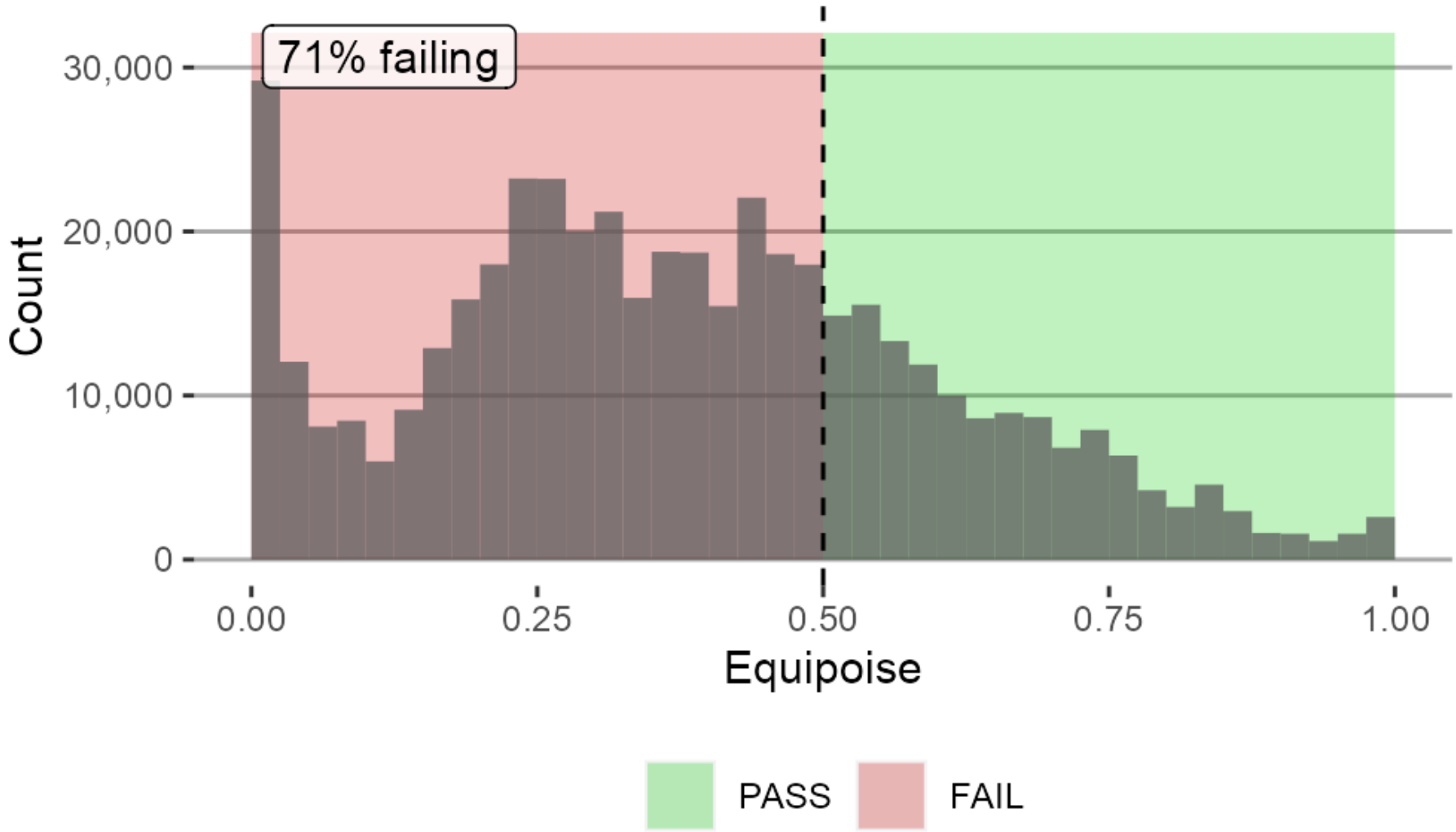
Rule: **Equipoise > 0.5**

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

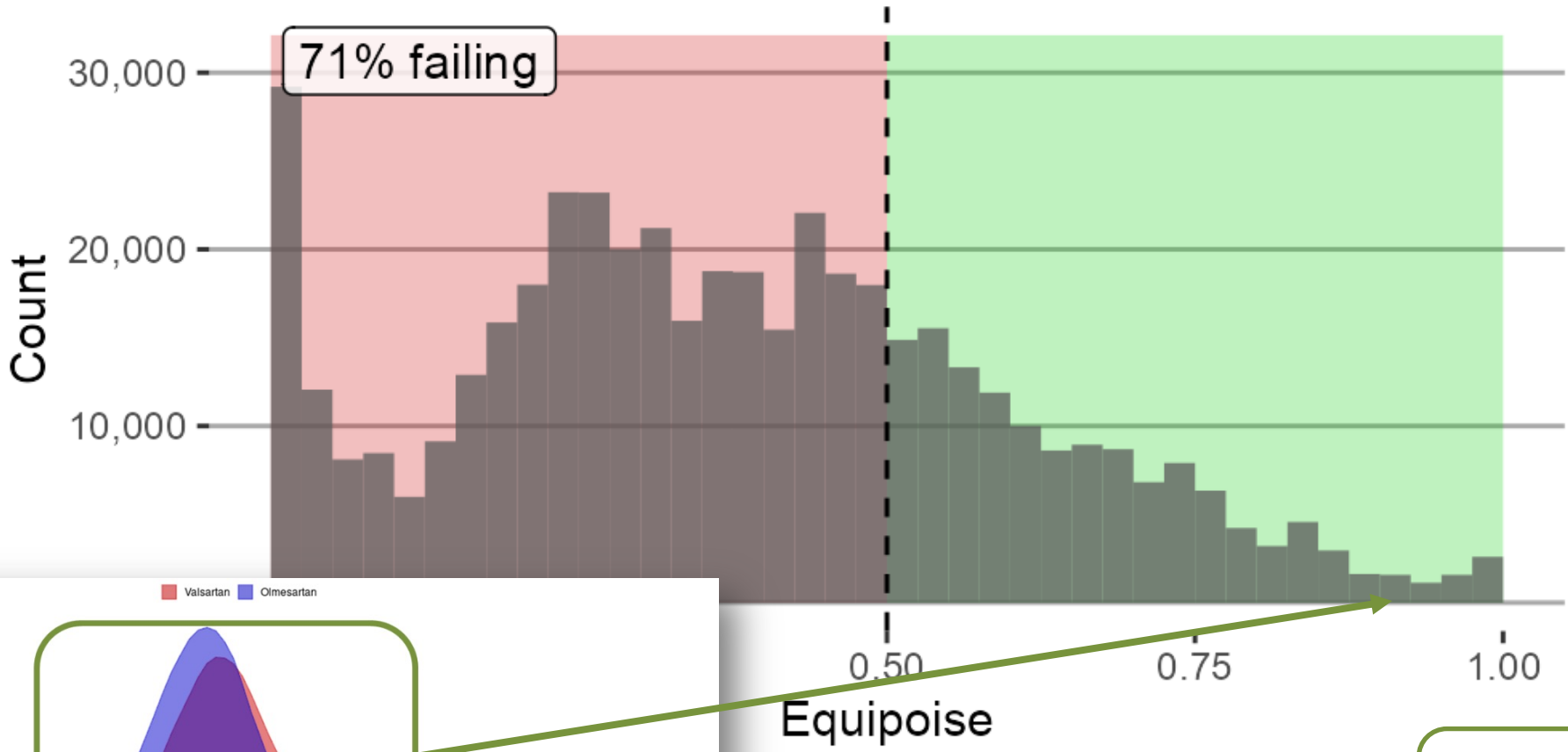
Reasoning:

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

Equipoise


71% failing
30,000
20,000
10,000
0
Count
0.00
0.25
0.50
0.75
1.00
Equipoise
PASS
FAIL

Equipoise



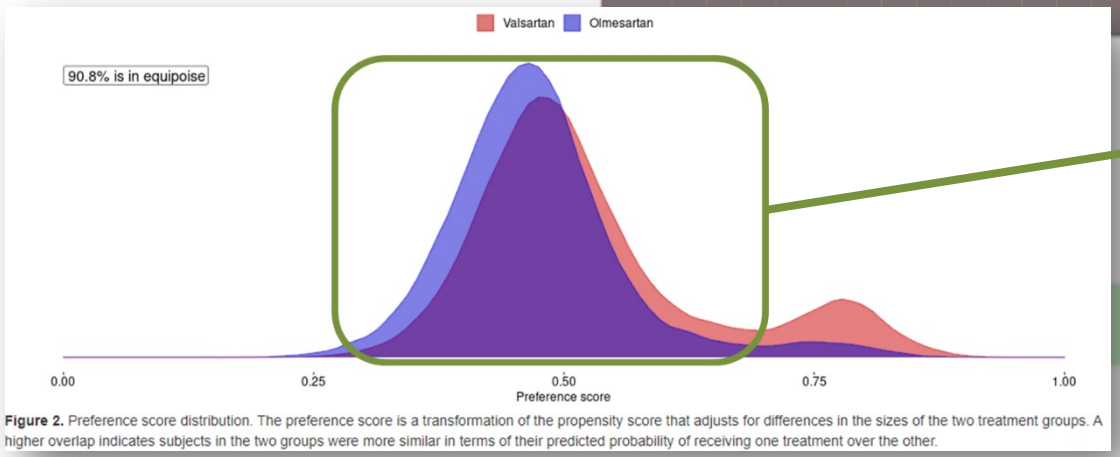


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

Patrick's example of a good study

Equipoise

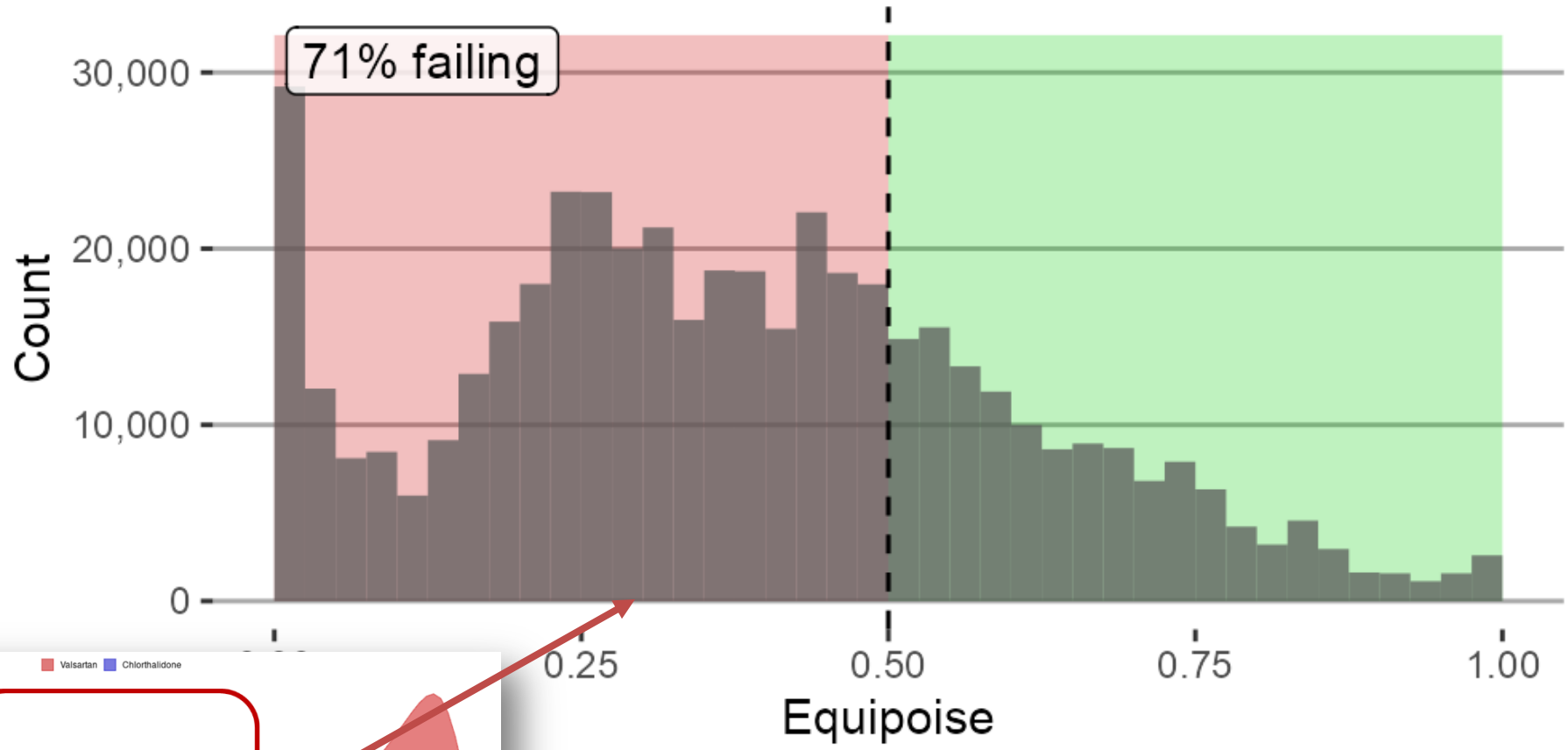
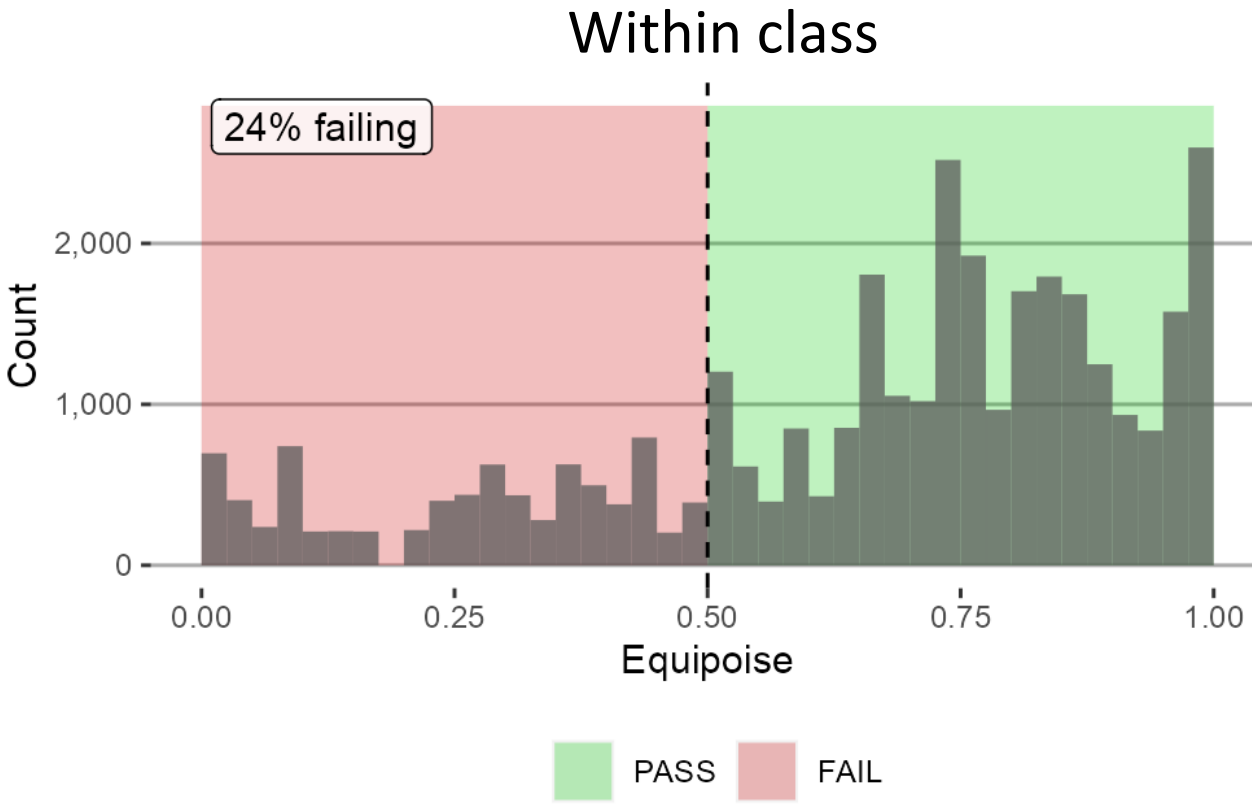
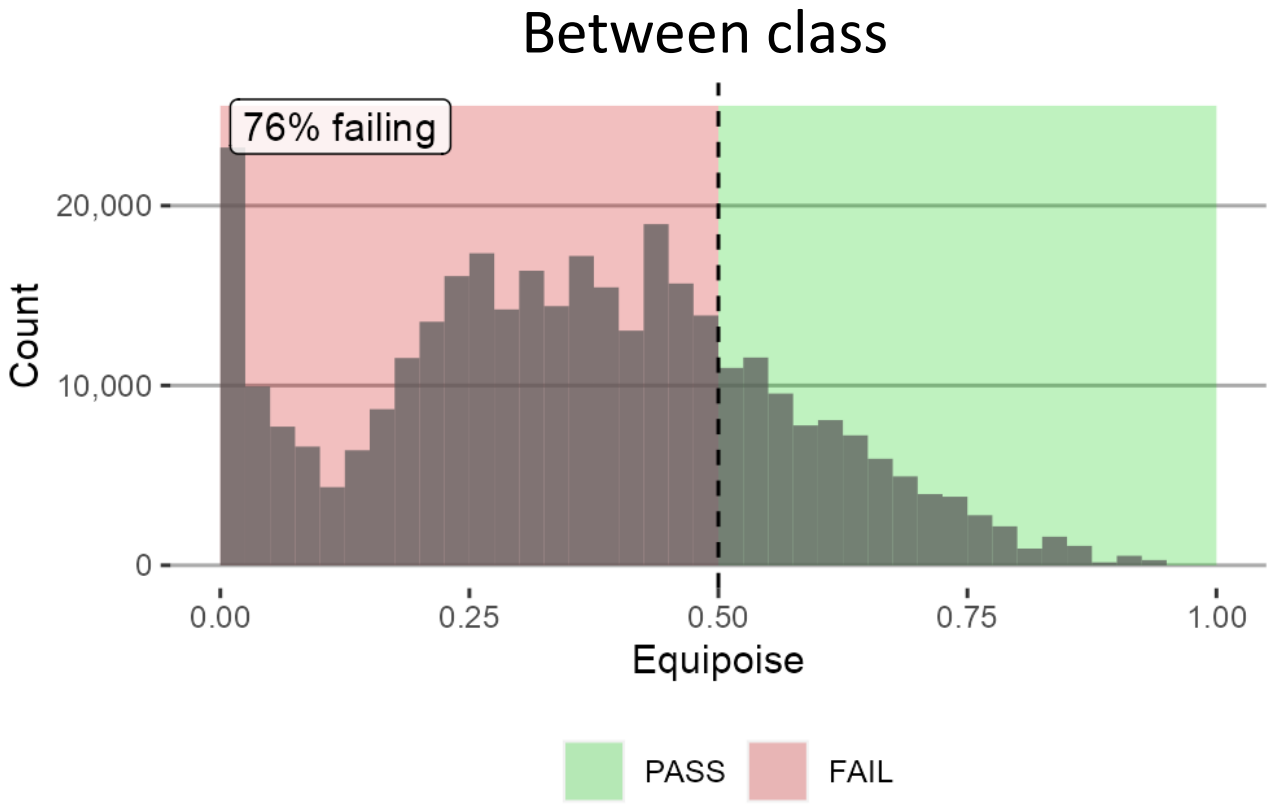


Figure 2. Preference score distribution. The preference score is a transformation of the propensity score that adjusts for differences in the sizes of the two treatment groups. A higher overlap indicates subjects in the two groups were more similar in terms of their predicted probability of receiving one treatment over the other.

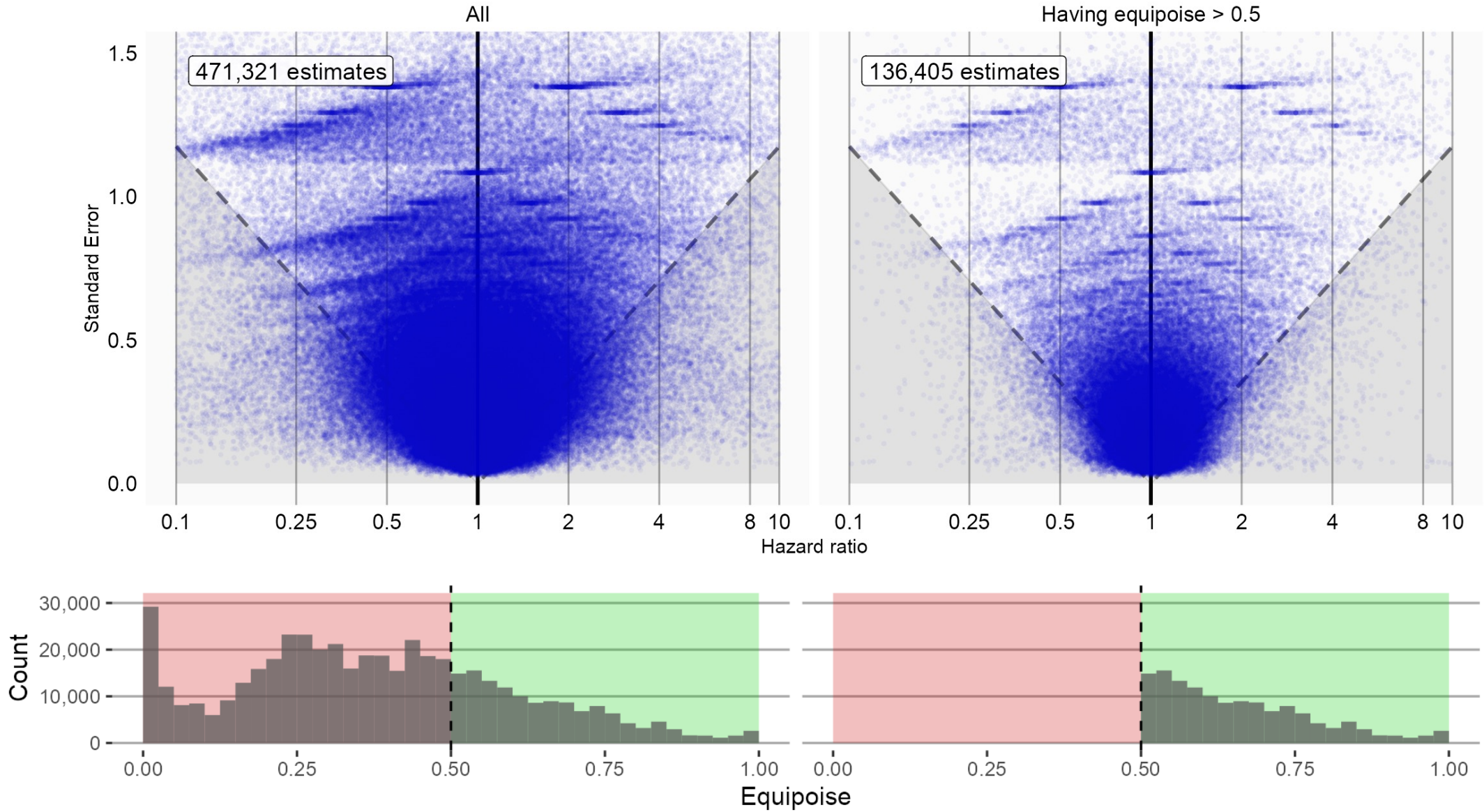
Patrick's example of a bad study

Equipoise

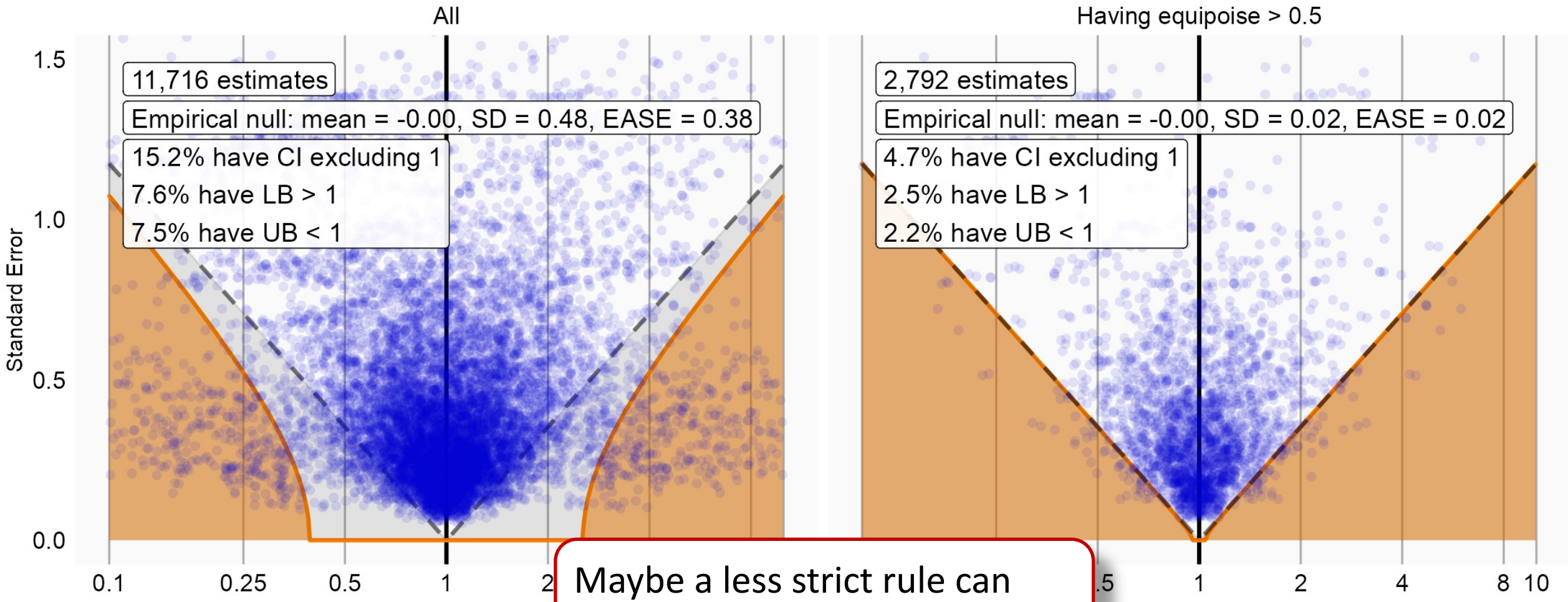

Within class
24% failing
2,000
1,000
0
Count
0.00
0.25
0.50
0.75
1.00
Equipoise
PASS
FAIL


Between class
76% failing
20,000
10,000
0
Count
0.00
0.25
0.50
0.75
1.00
Equipoise
PASS
FAIL

Equipoise

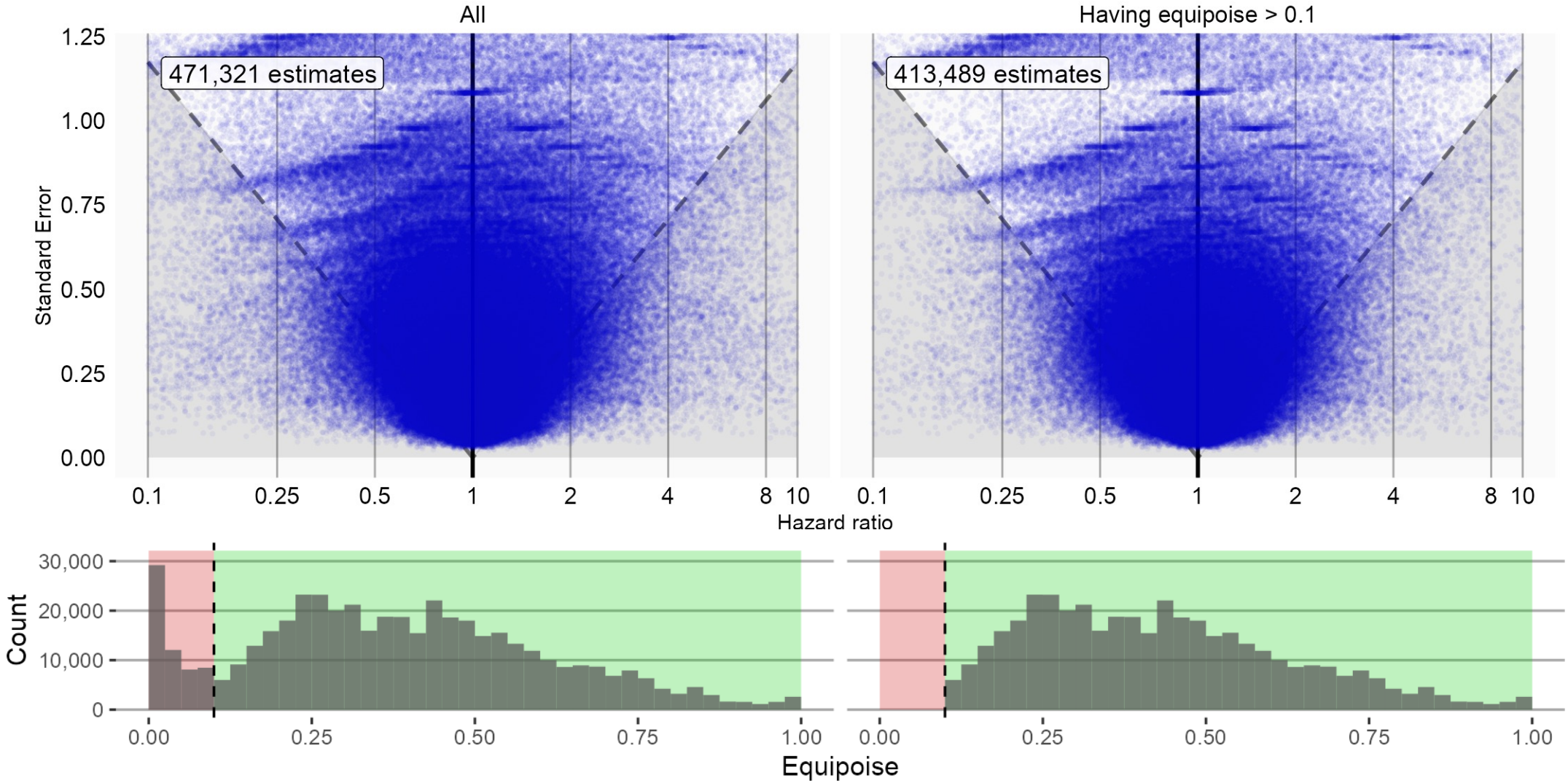

All
Having equipoise > 0.5
471,321 estimates
136,405 estimates
Standard Error
1.5
1.0
0.5
0.0
0.1
0.25
0.5
1
2
4
8
10
Hazard ratio
Count
30,000
20,000
10,000
0
0.00
0.25
0.50
0.75
1.00
Equipoise

Equipoise

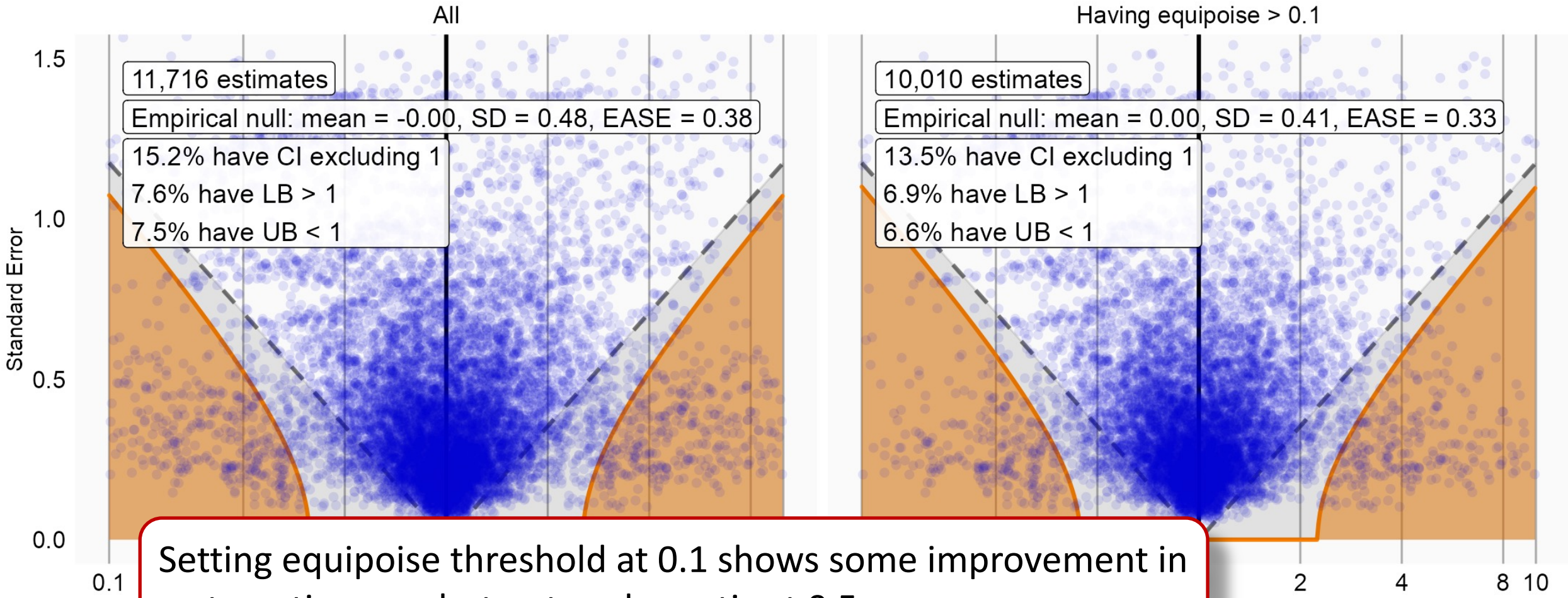


Maybe a less strict rule can keep more estimates?

Equipoise relaxing to > 0.1



Equipoise relaxing to > 0.1



Setting equipoise threshold at 0.1 shows some improvement in systematic error, but not as dramatic at 0.5.

Covariate balance

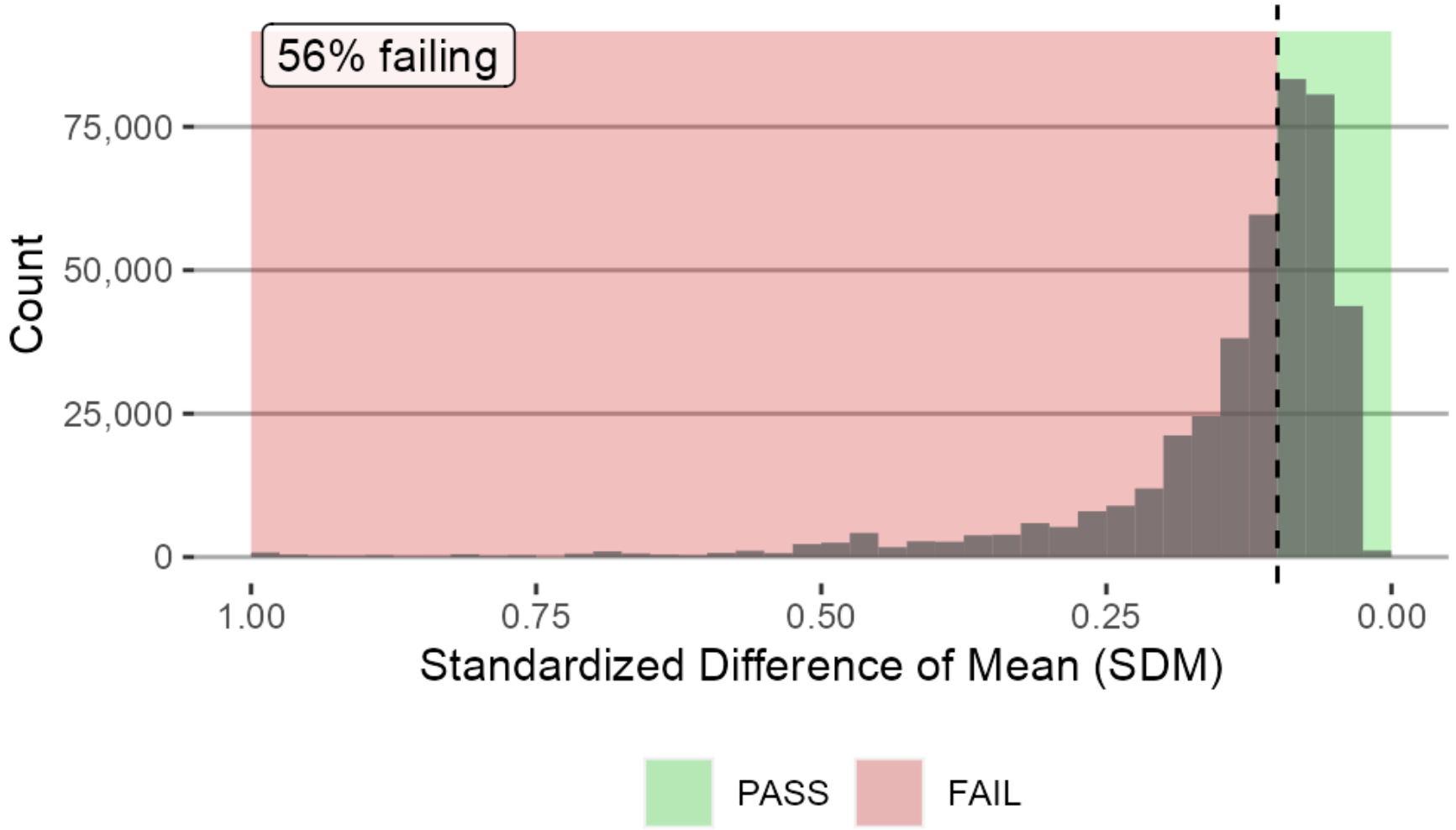
Rule: **Max standardized difference of mean (SDM) < 0.1**

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

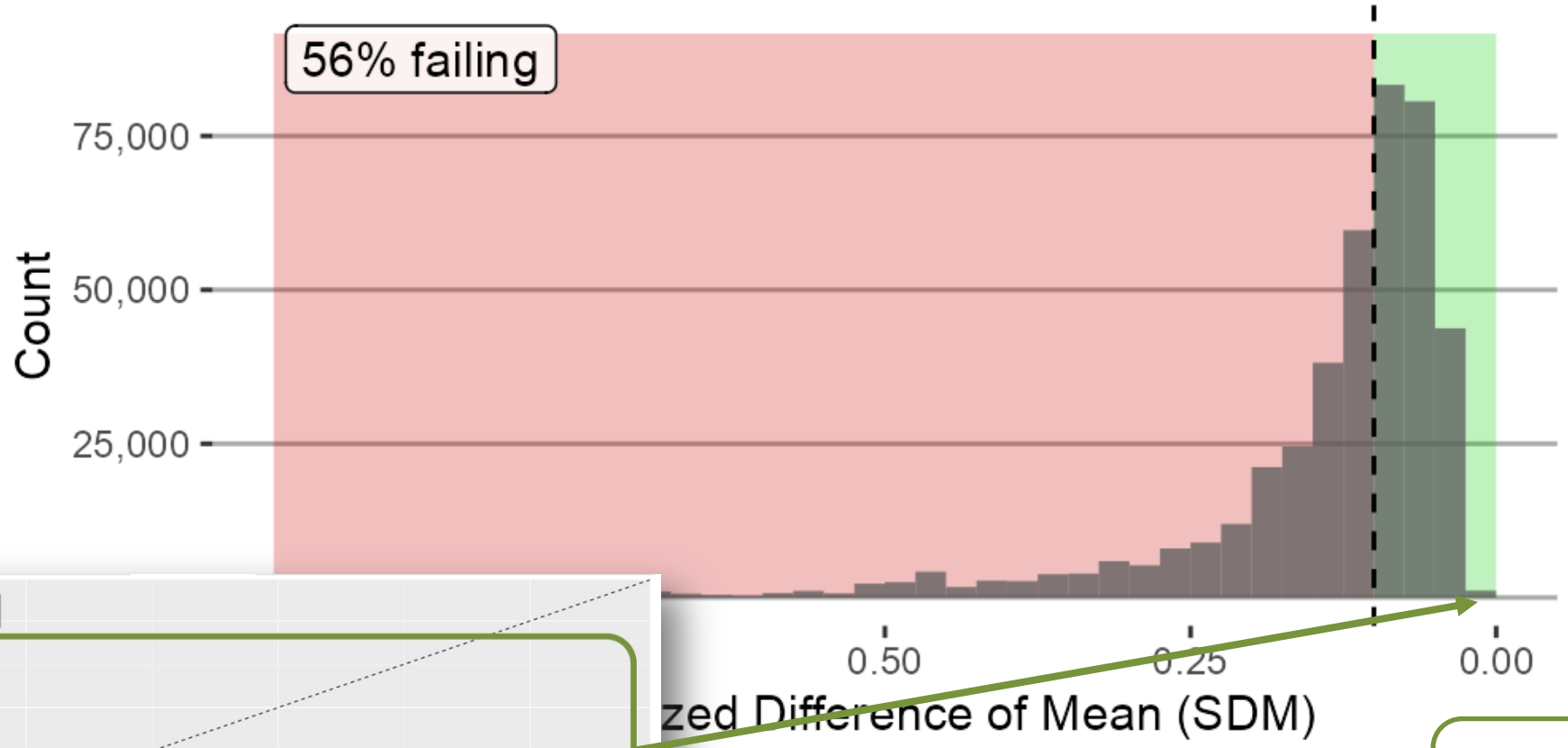
Reasoning:

If covariates are unbalanced there may be confounding.

Covariate balance



Covariate balance



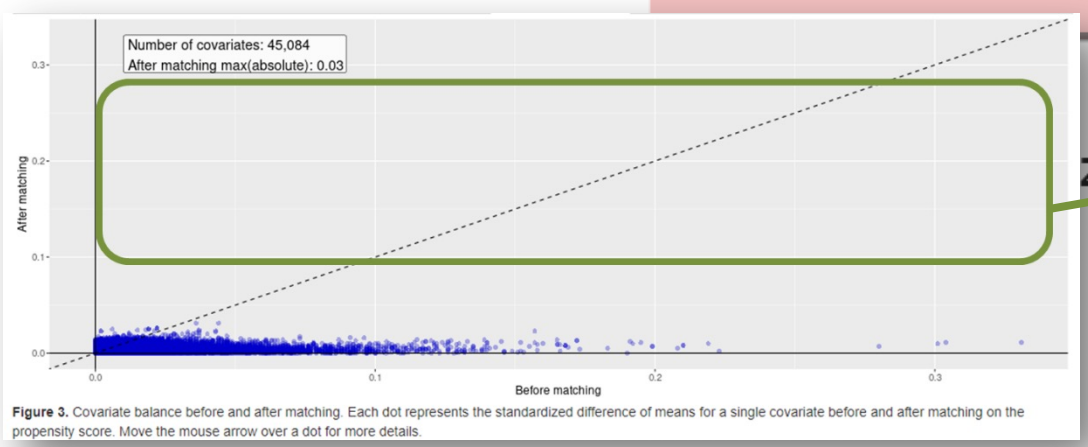
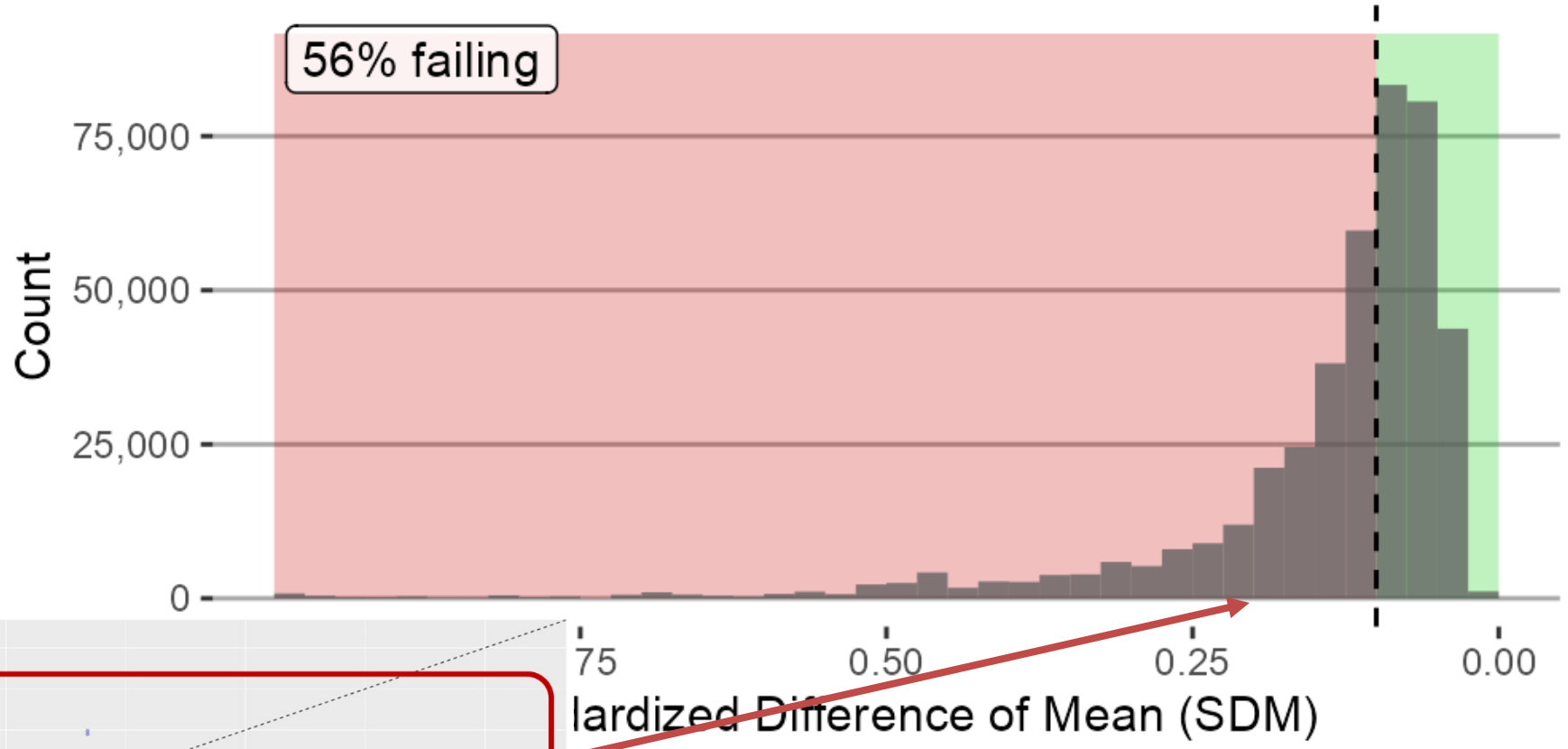


Figure 3. Covariate balance before and after matching. Each dot represents the standardized difference of means for a single covariate before and after matching on the propensity score. Move the mouse arrow over a dot for more details.

Patrick's example of a good study

Covariate balance



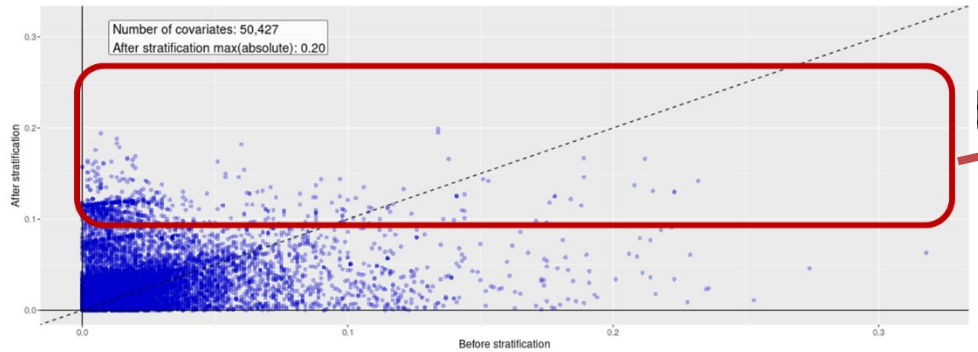
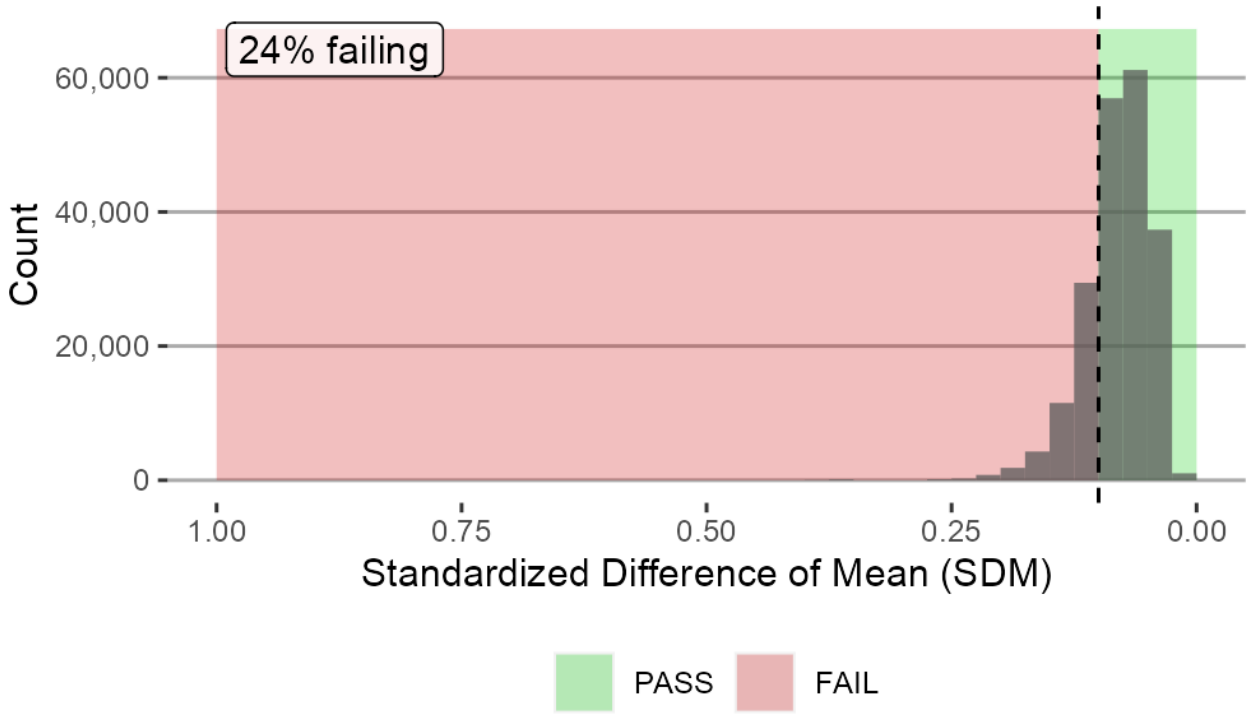


Figure 3. Covariate balance before and after stratification. Each dot represents the standardized difference of means for a single covariate before and after stratification on the propensity score. Move the mouse arrow over a dot for more details.

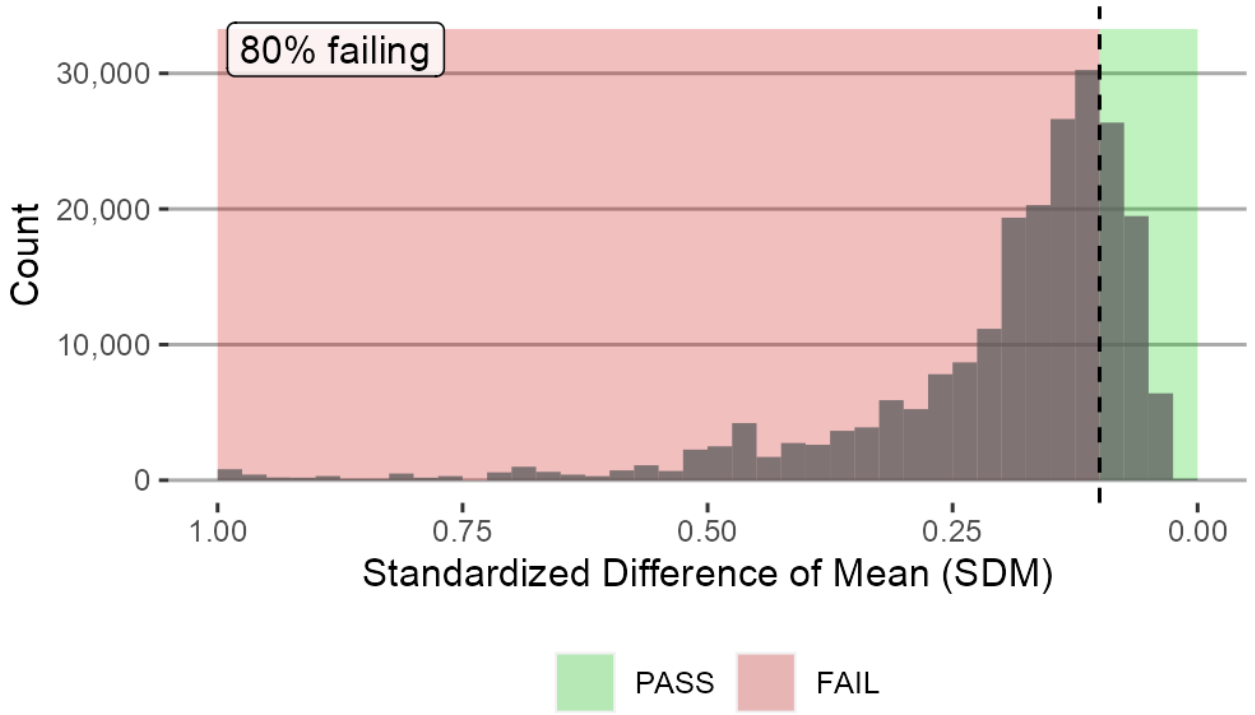
Patrick's example of a bad study

Covariate balance

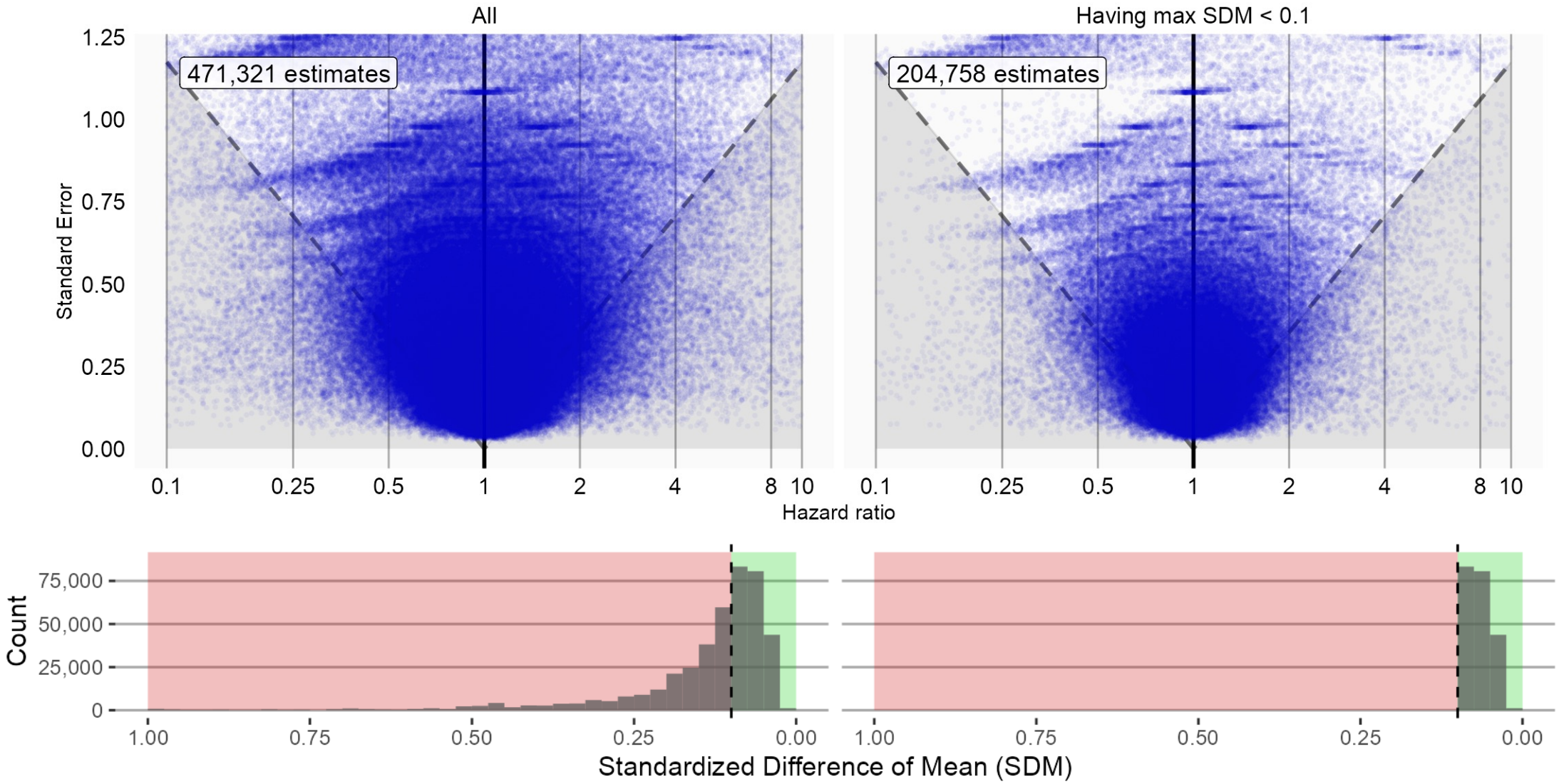
Propensity score matching



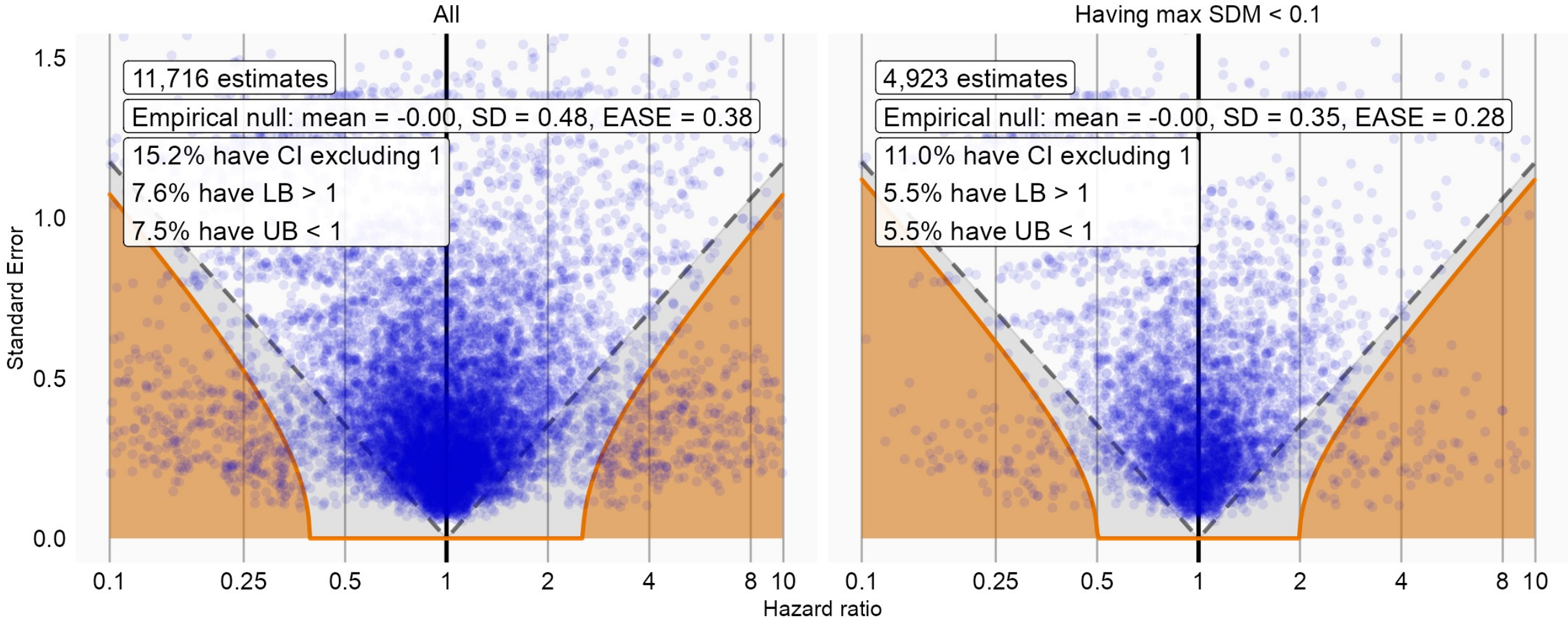
Propensity score stratification



Covariate balance



Covariate balance



Generalizability

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

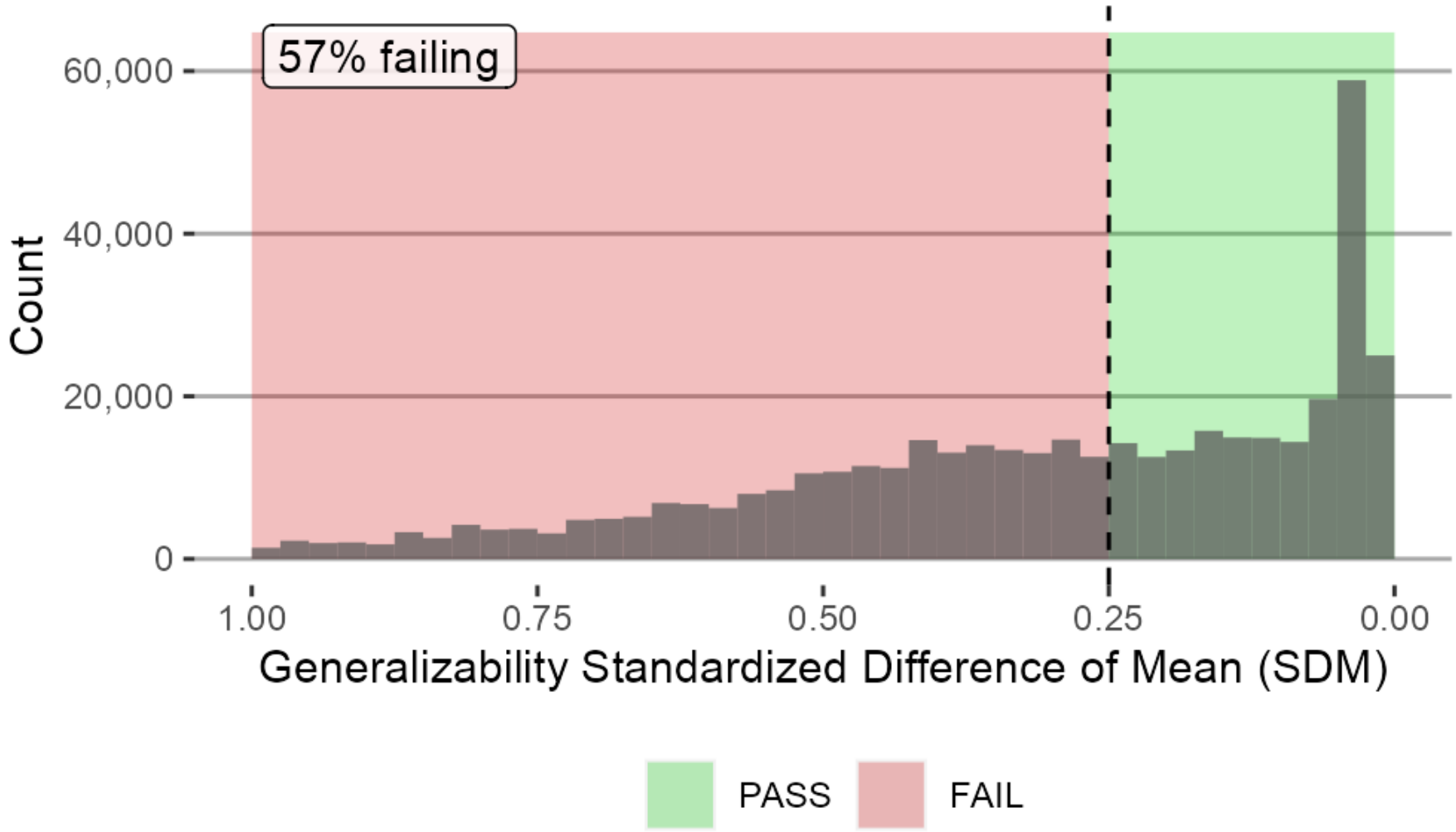
(target cohort: the cohort we started with (those exposed))

(analytic cohort: the cohort after all adjustments)

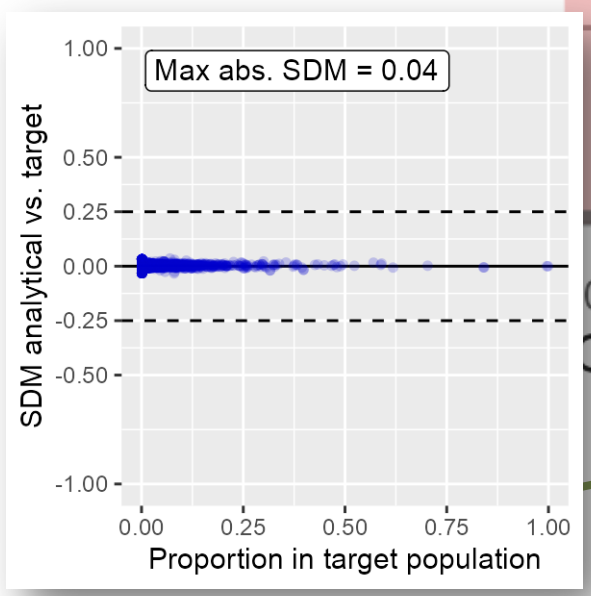
Reasoning:

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

Generalizability

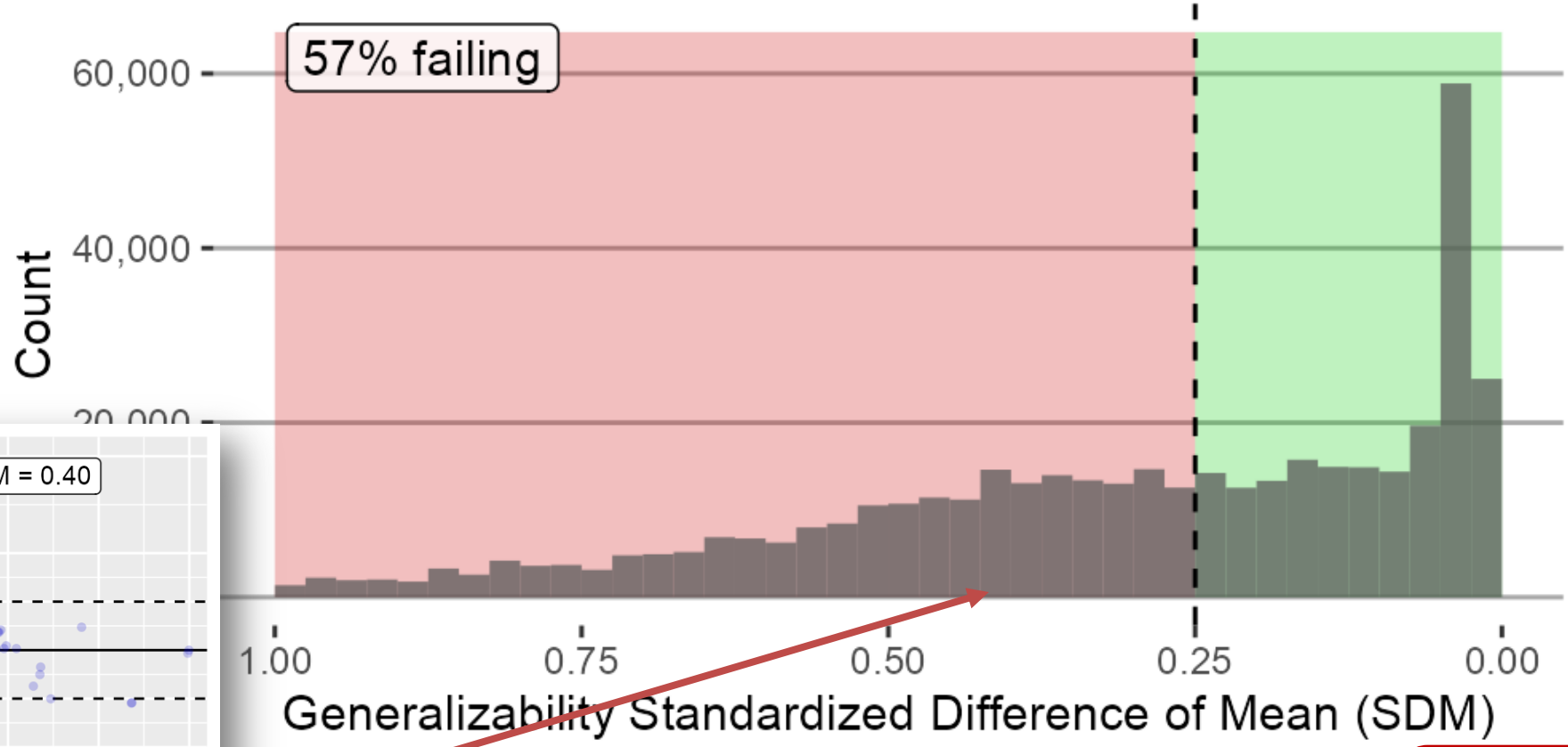


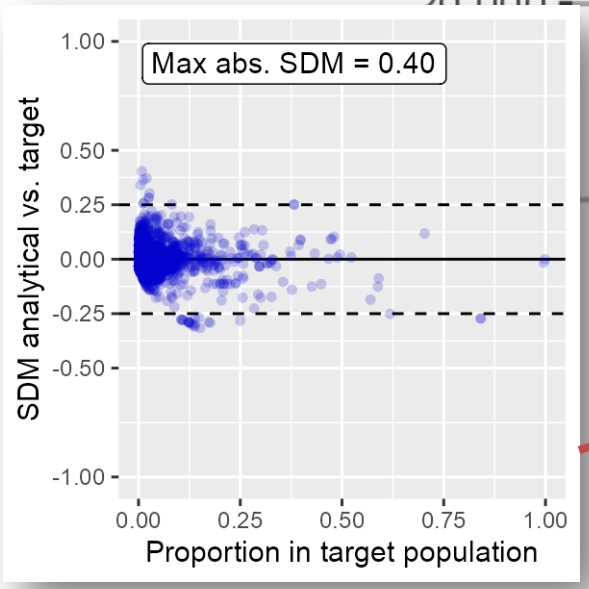
Generalizability



Patrick's example of a good study

Generalizability

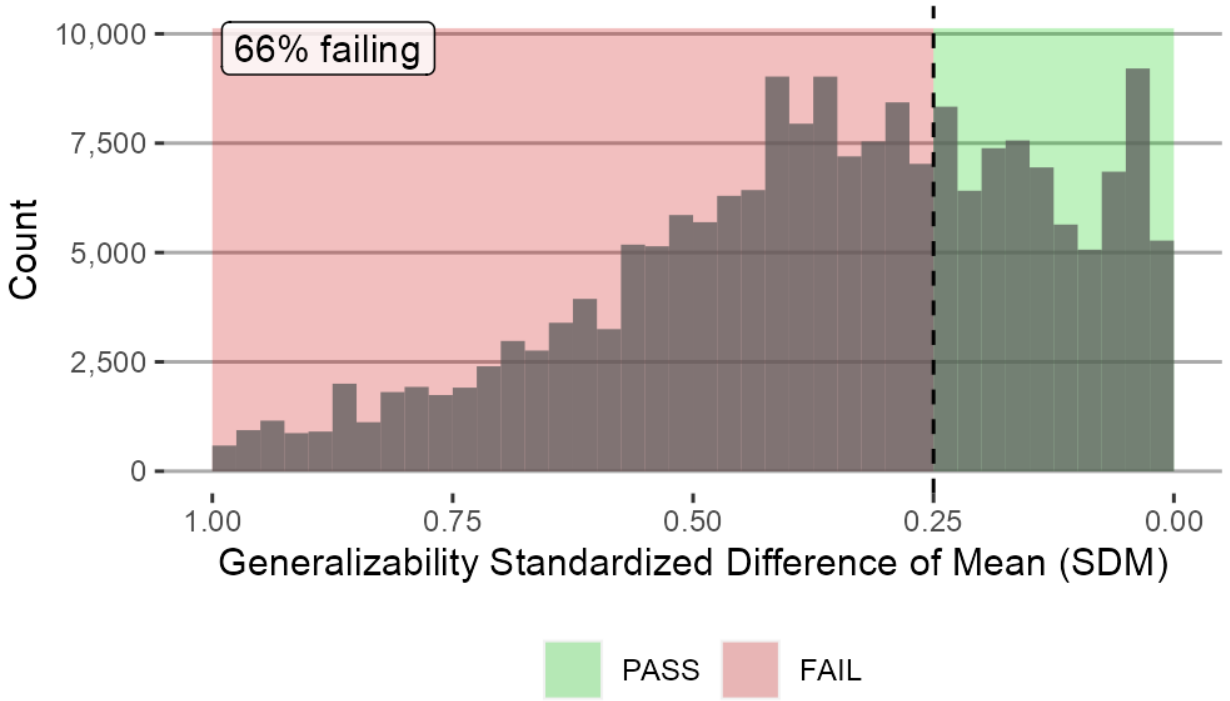




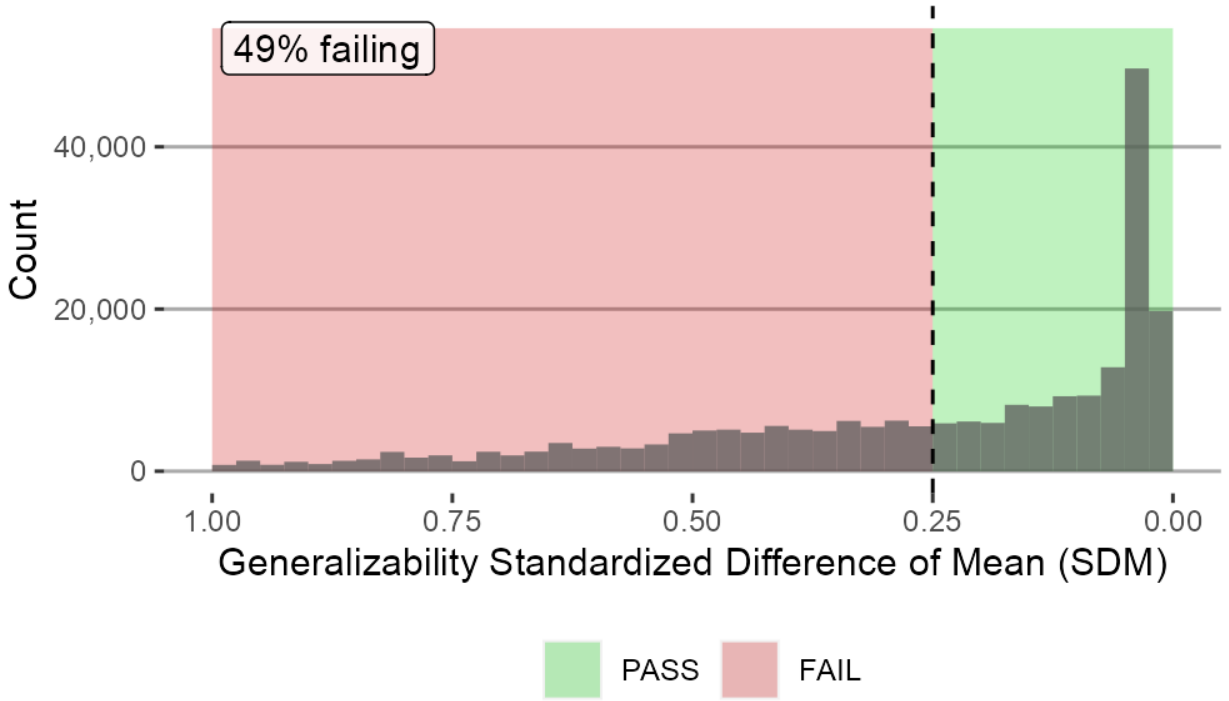
Patrick's example of a bad study

Generalizability

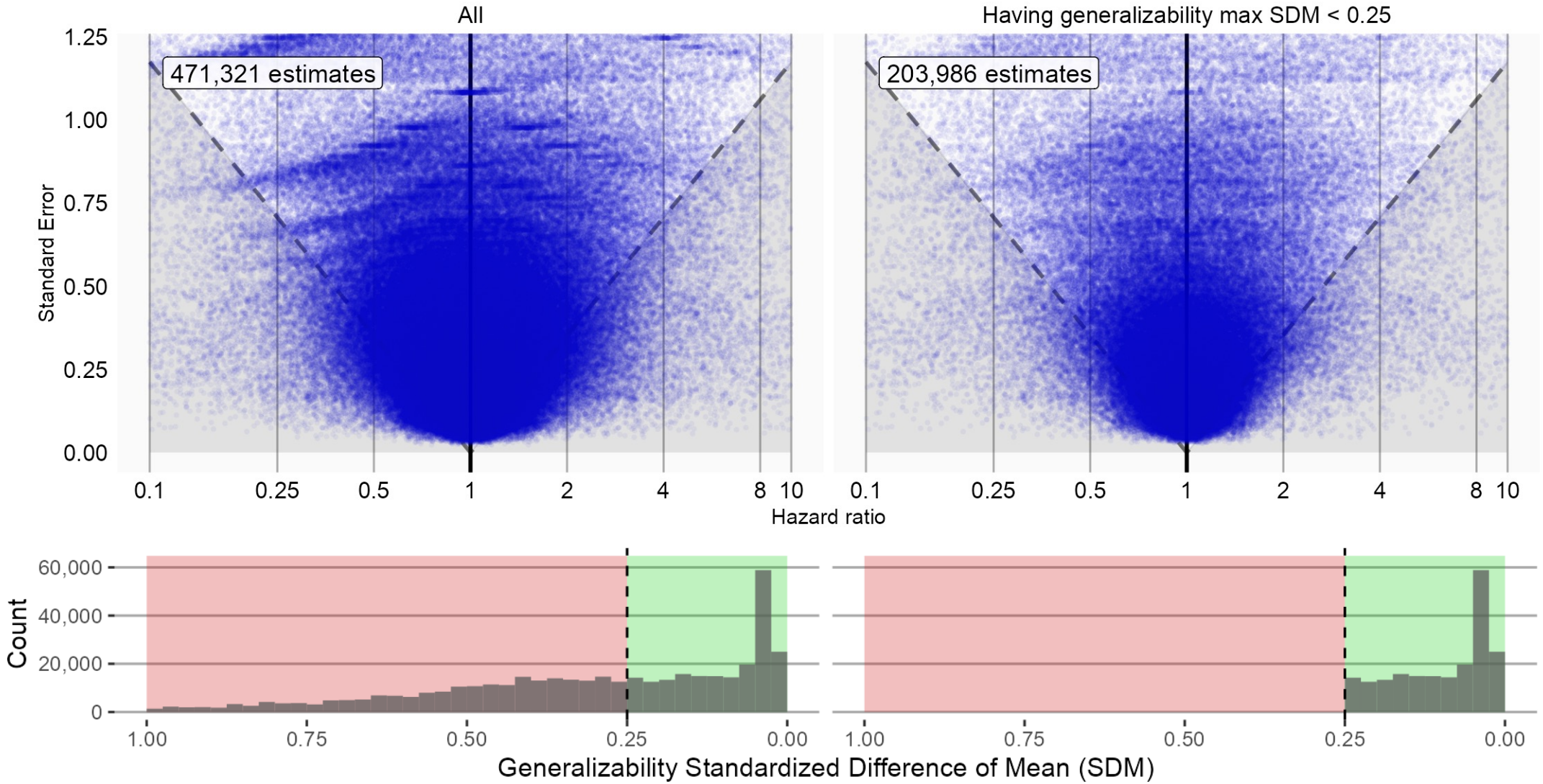
Propensity score matching



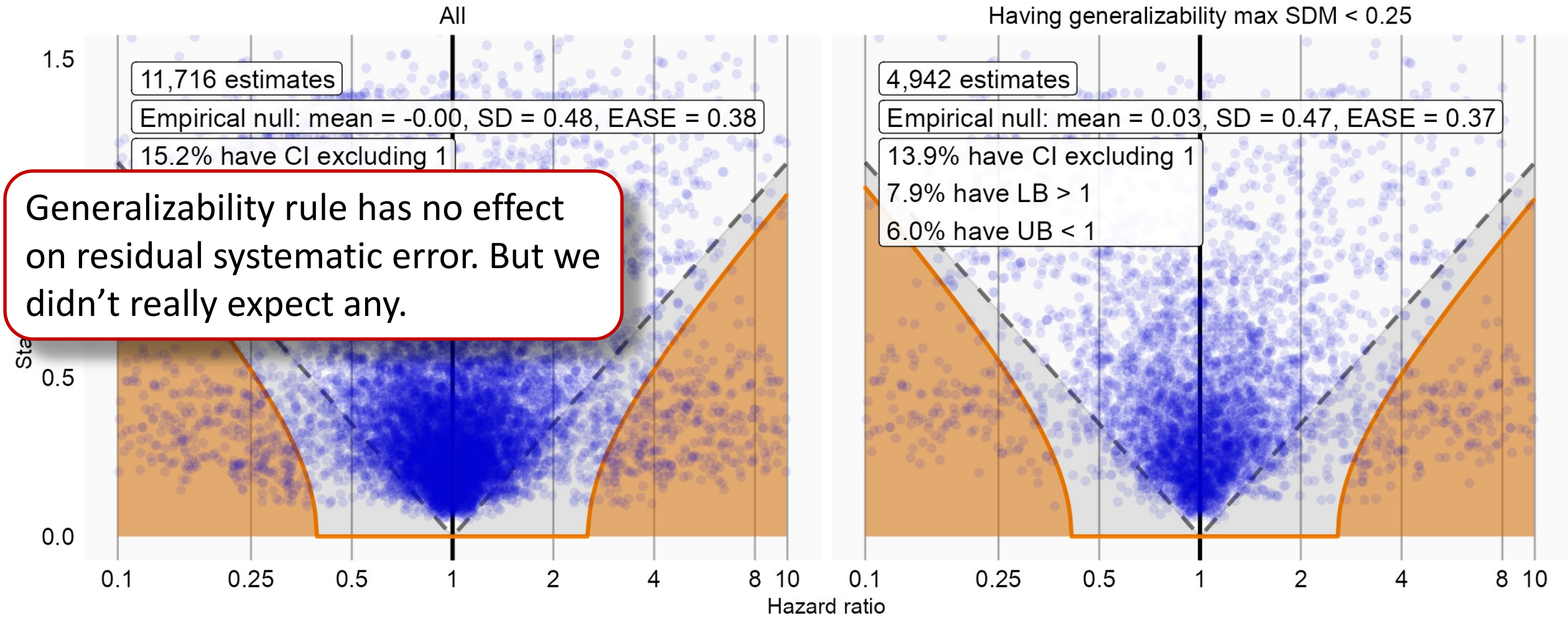
Propensity score stratification



Generalizability

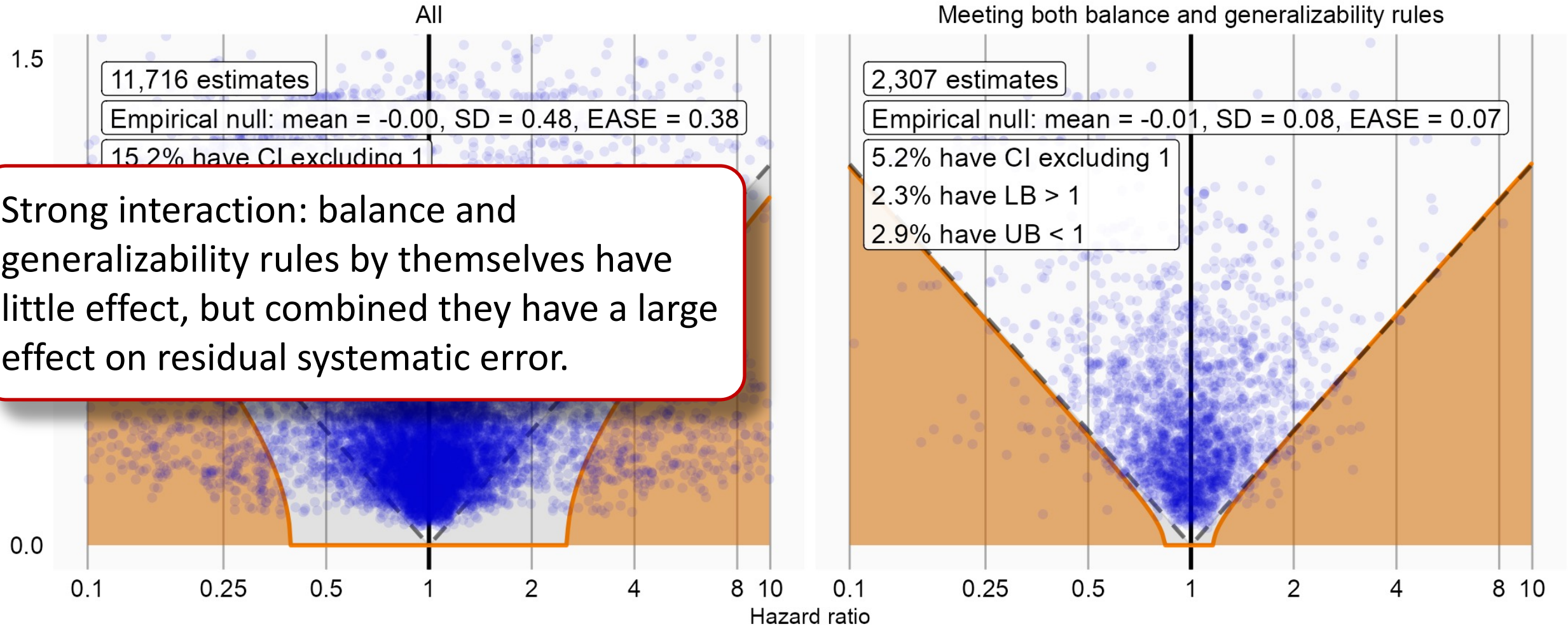


Generalizability



Generalizability rule has no effect on residual systematic error. But we didn't really expect any.

Strong interaction effect between Covariate balance and generalizability



Strong interaction: balance and generalizability rules by themselves have little effect, but combined they have a large effect on residual systematic error.

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)

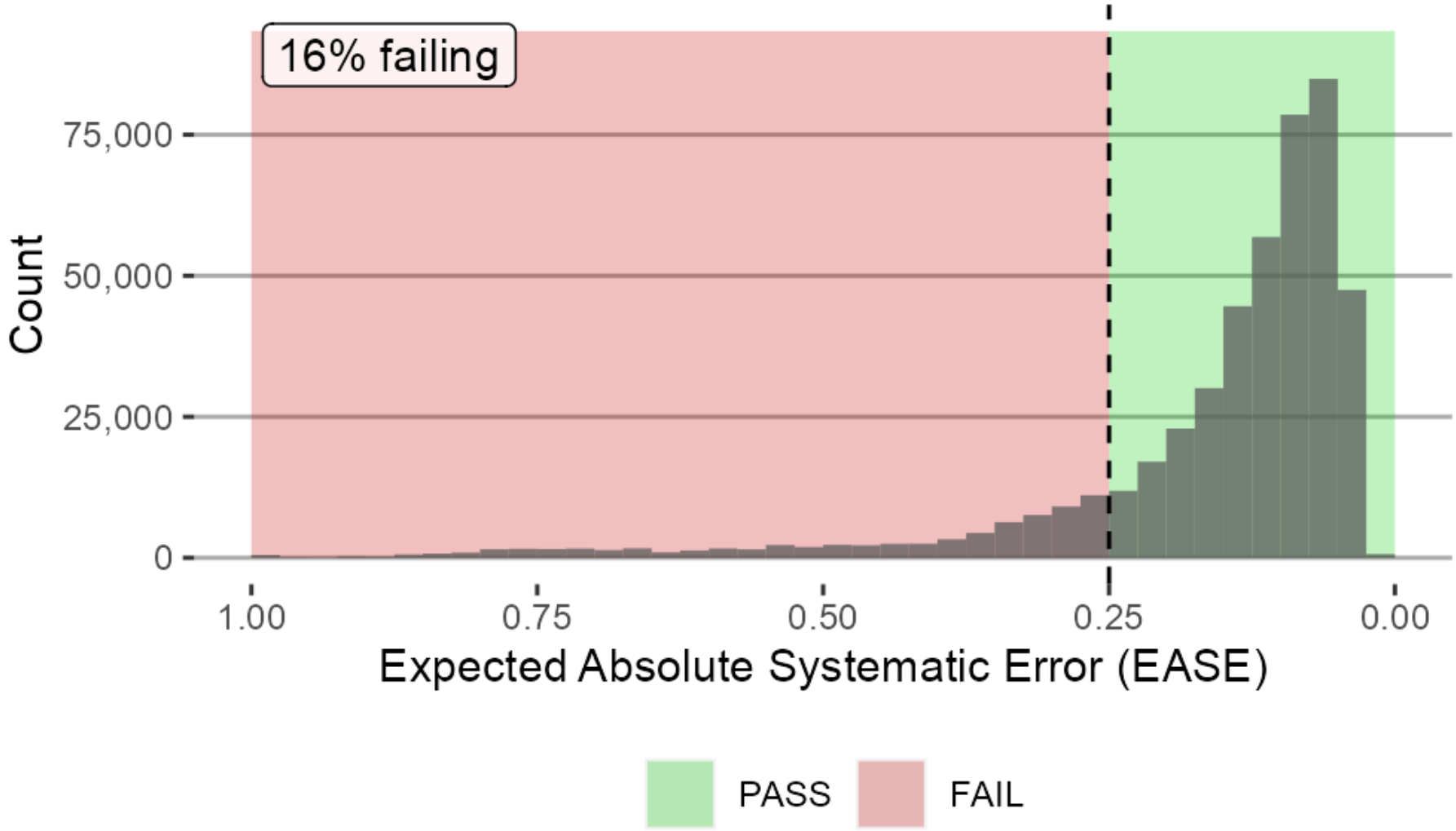
Reasoning:

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.

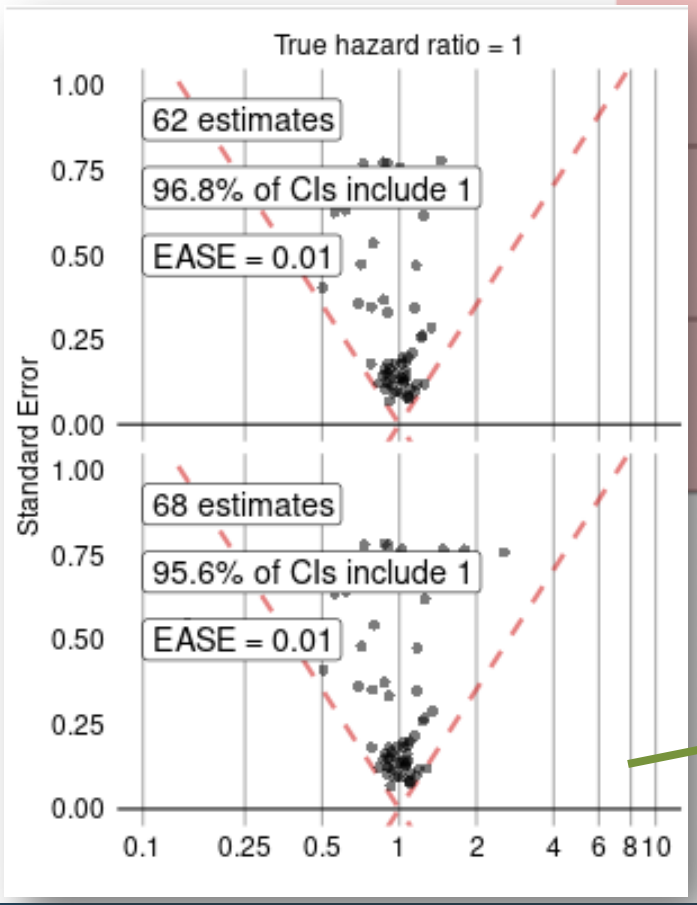
Note:

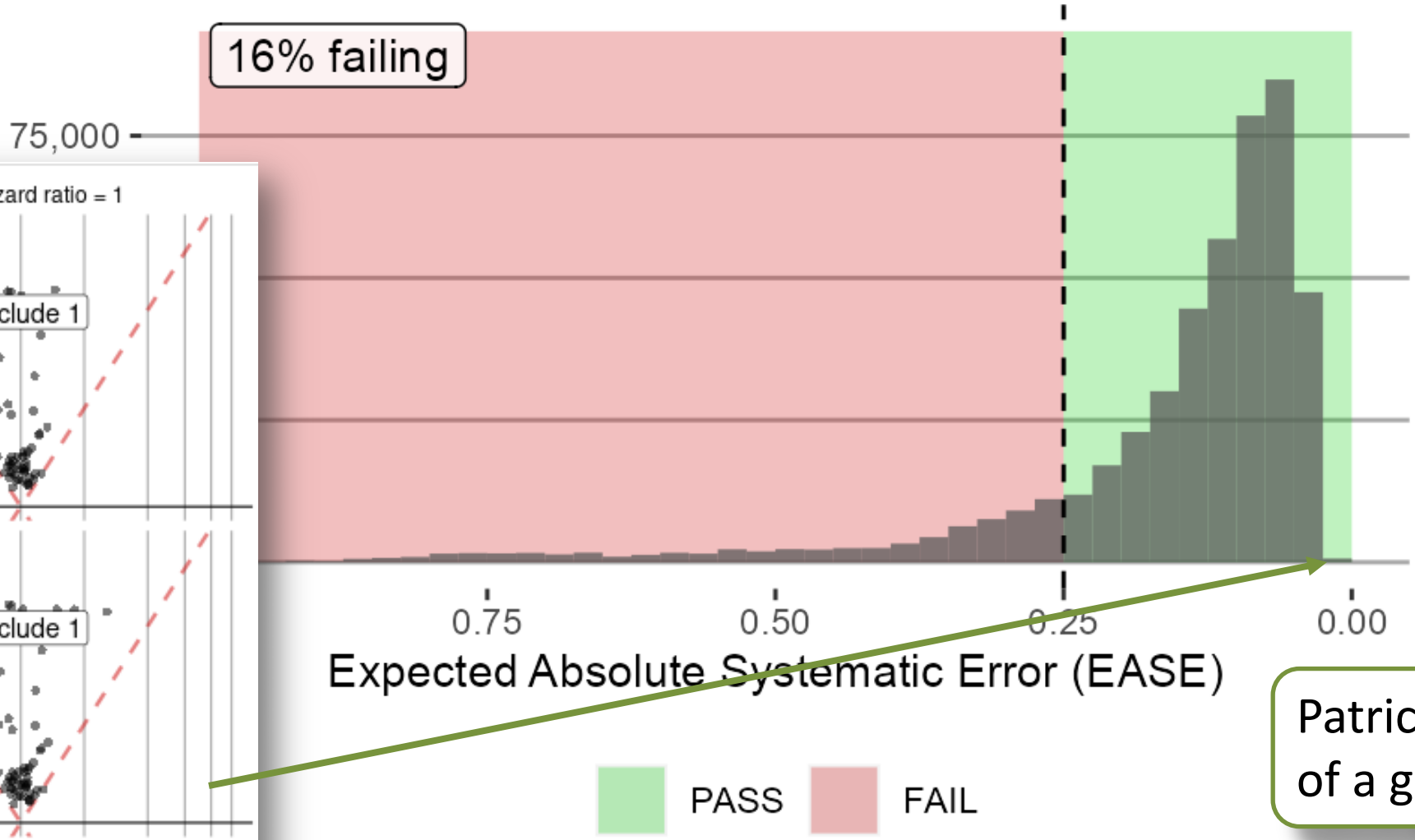
Our evaluation uses calibrated estimates, which already incorporates the systematic error observed for the original set of negative controls.

Systematic error



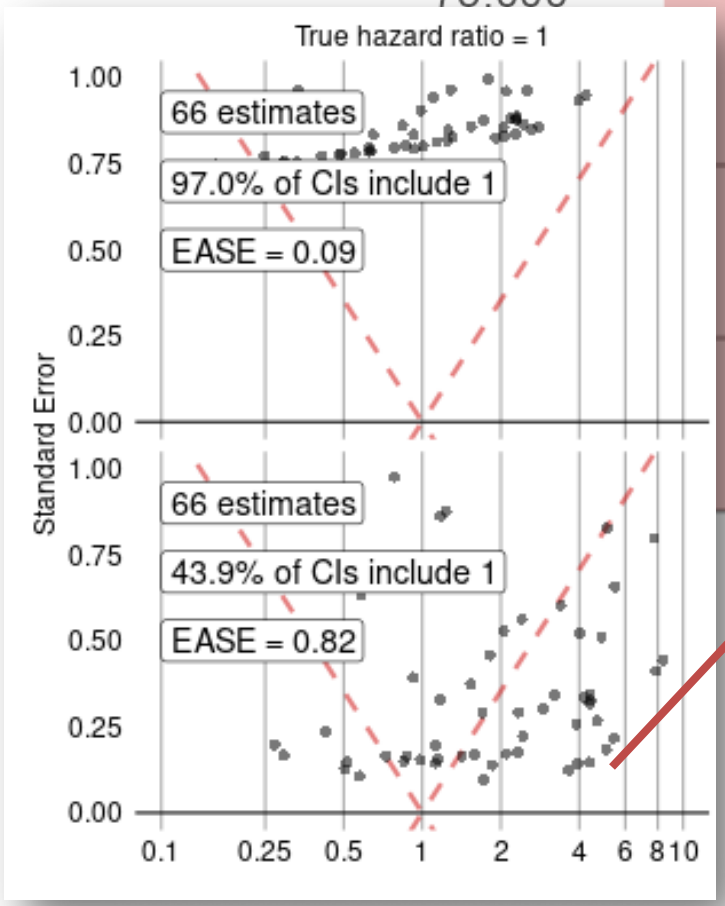
Systematic error

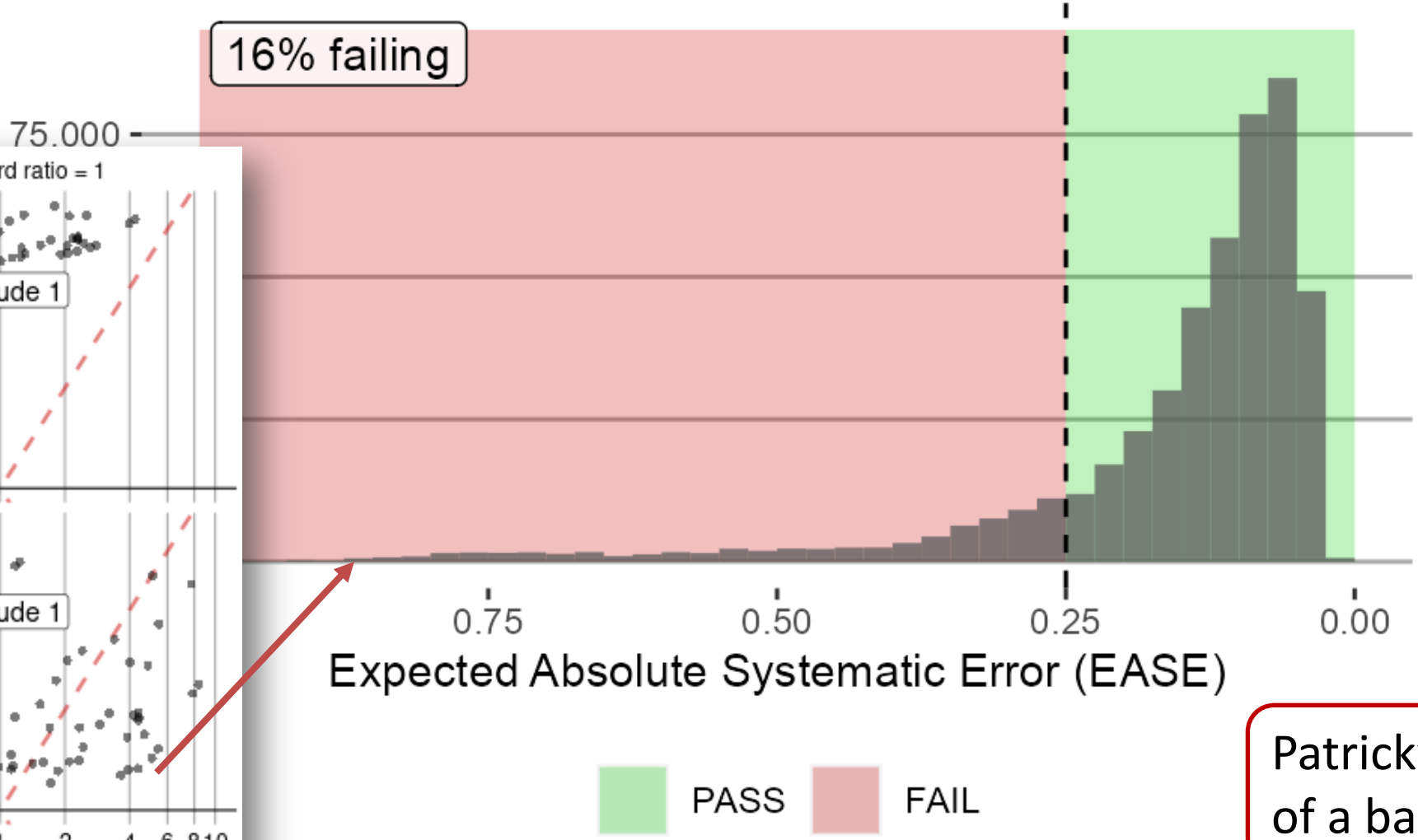




Patrick's example of a good study

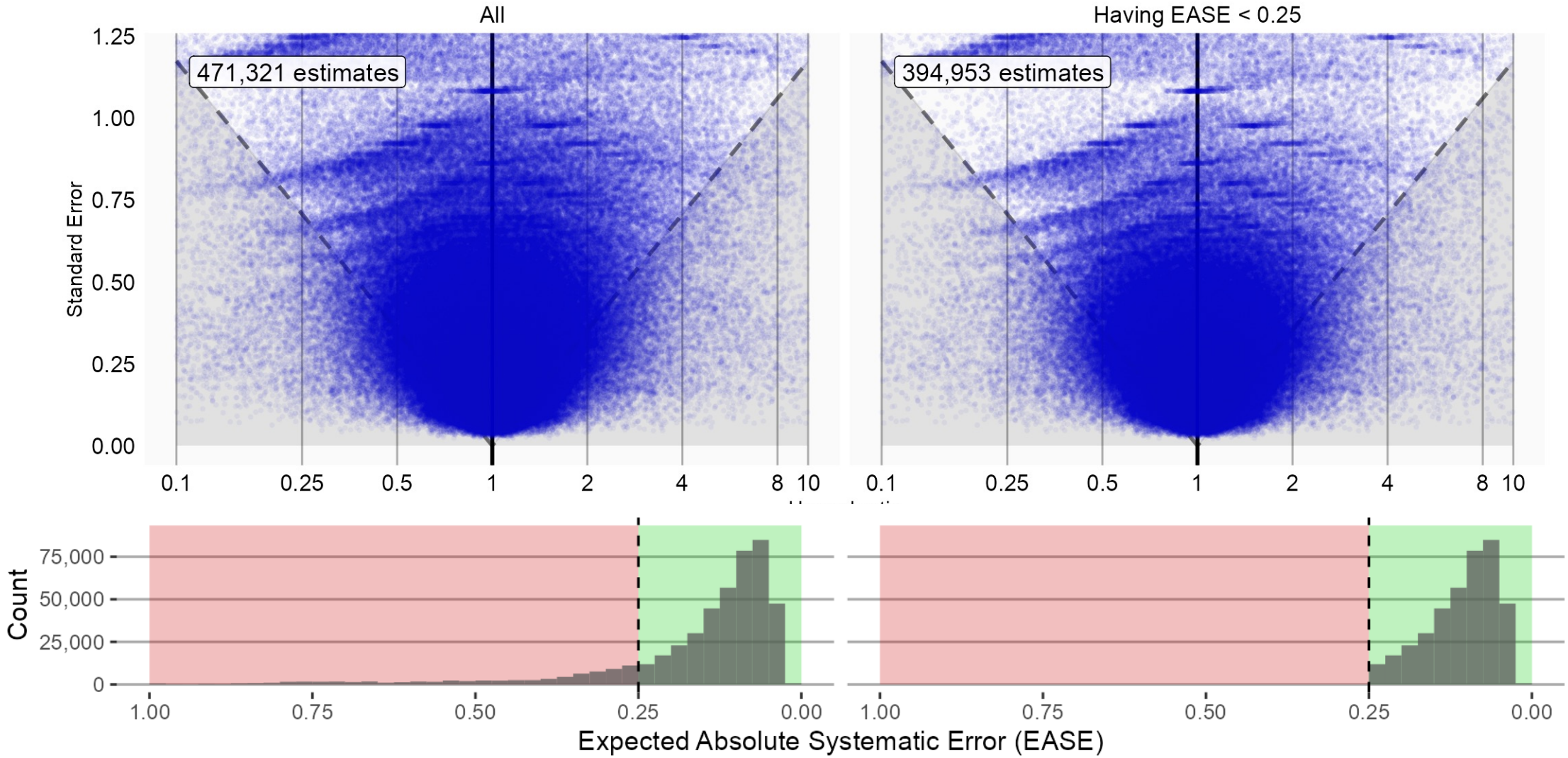
Systematic error



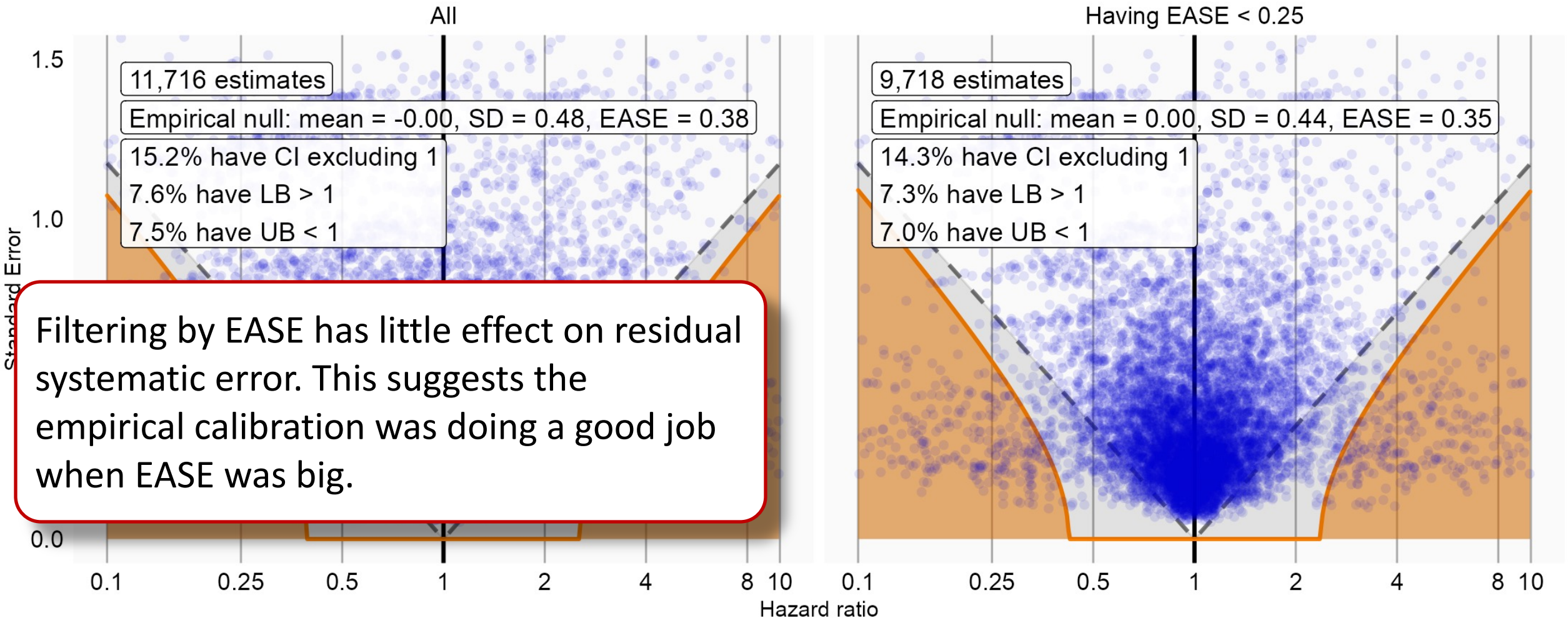


Patrick's example of a bad study

Systematic error



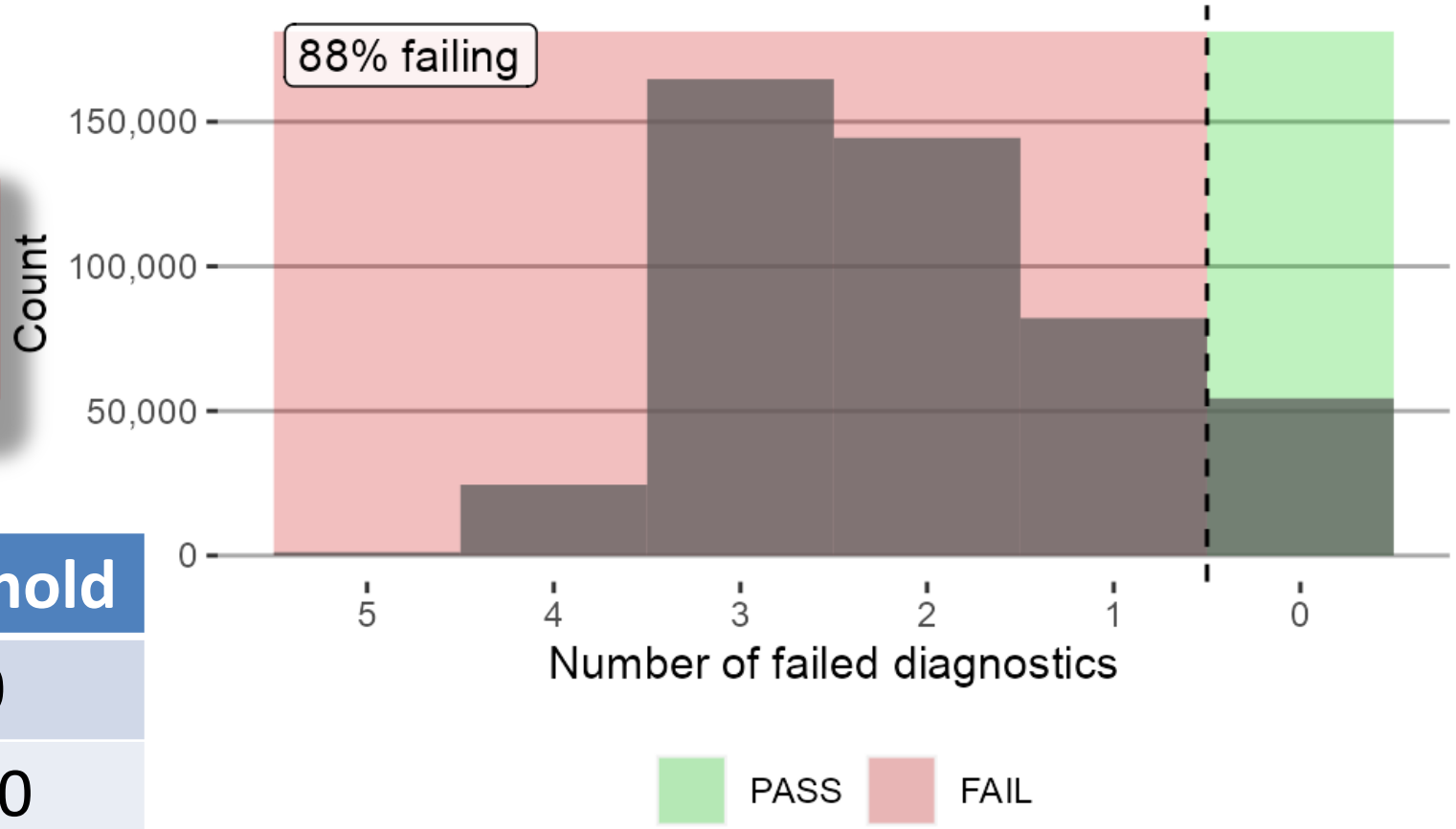
Systematic error



Filtering by EASE has little effect on residual systematic error. This suggests the empirical calibration was doing a good job when EASE was big.

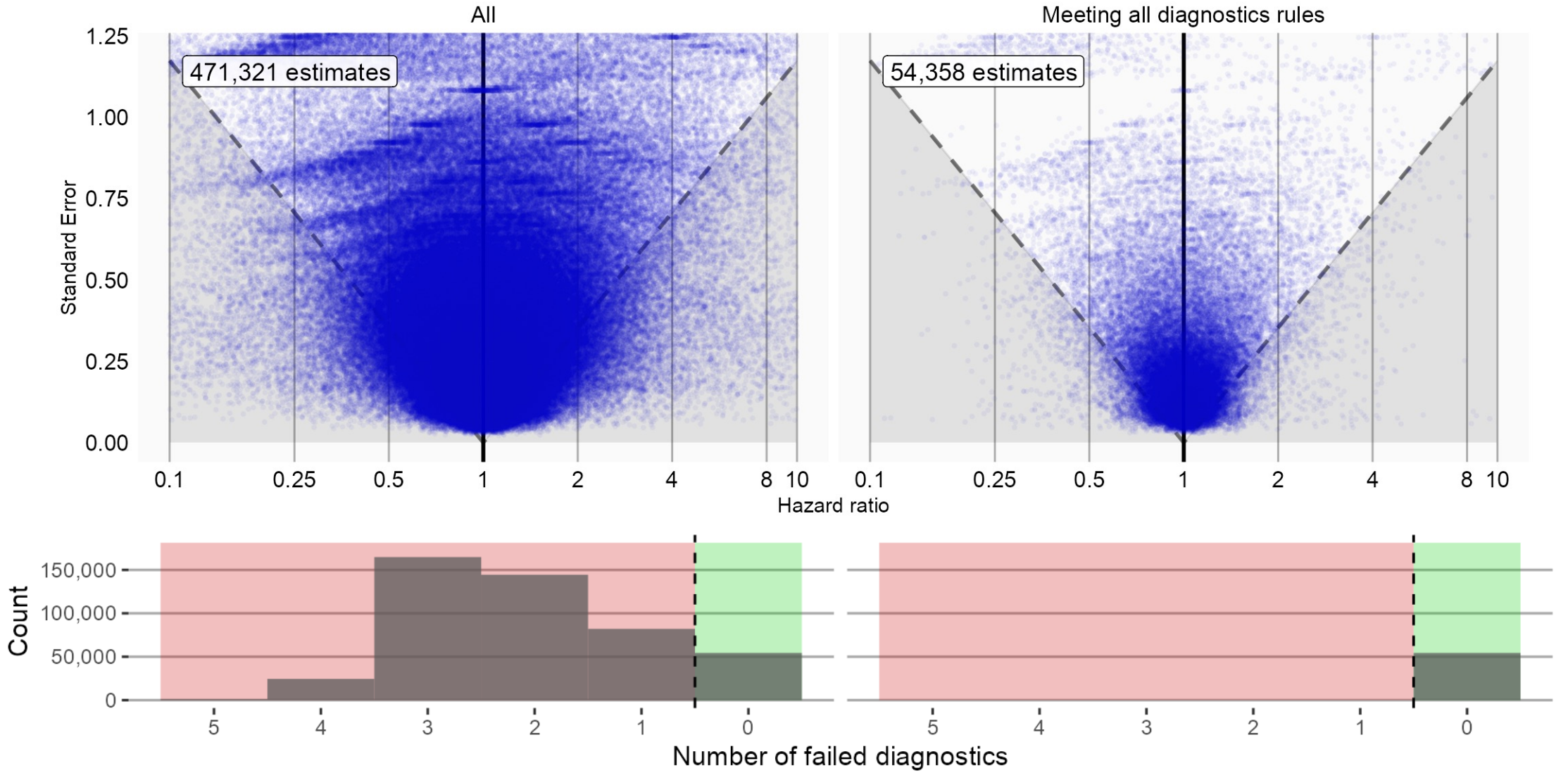
Combining all diagnostics

Majority of estimates fail 2 or more diagnostics. Need to understand how they interact.

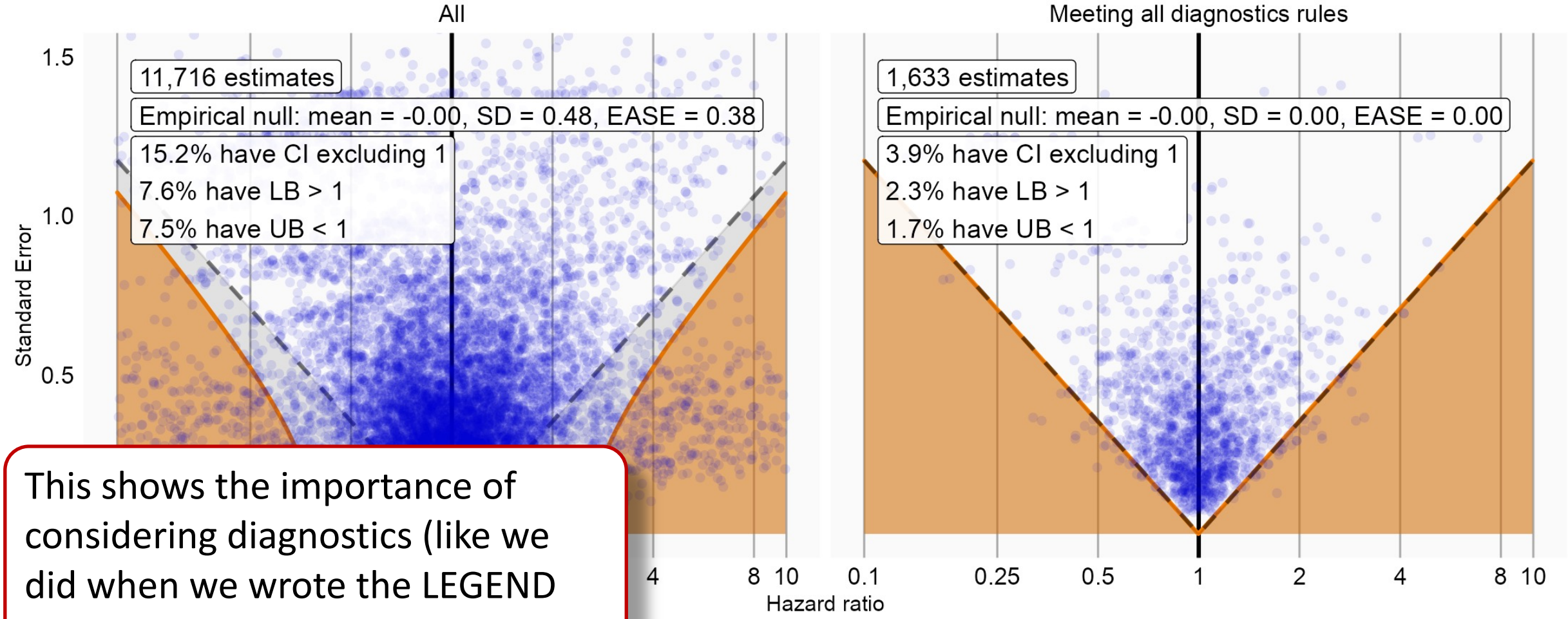


Diagnostic	Threshold
Statistical power (MDRR)	10
Equipoise	0.50
Covariate balance (SDM)	0.10
Generalizability (SDM)	0.25
Systematic error (EASE)	0.25

Combining all diagnostics



Combining all diagnostics



This shows the importance of considering diagnostics (like we did when we wrote the LEGEND Hypertension papers)

Can we do better?

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

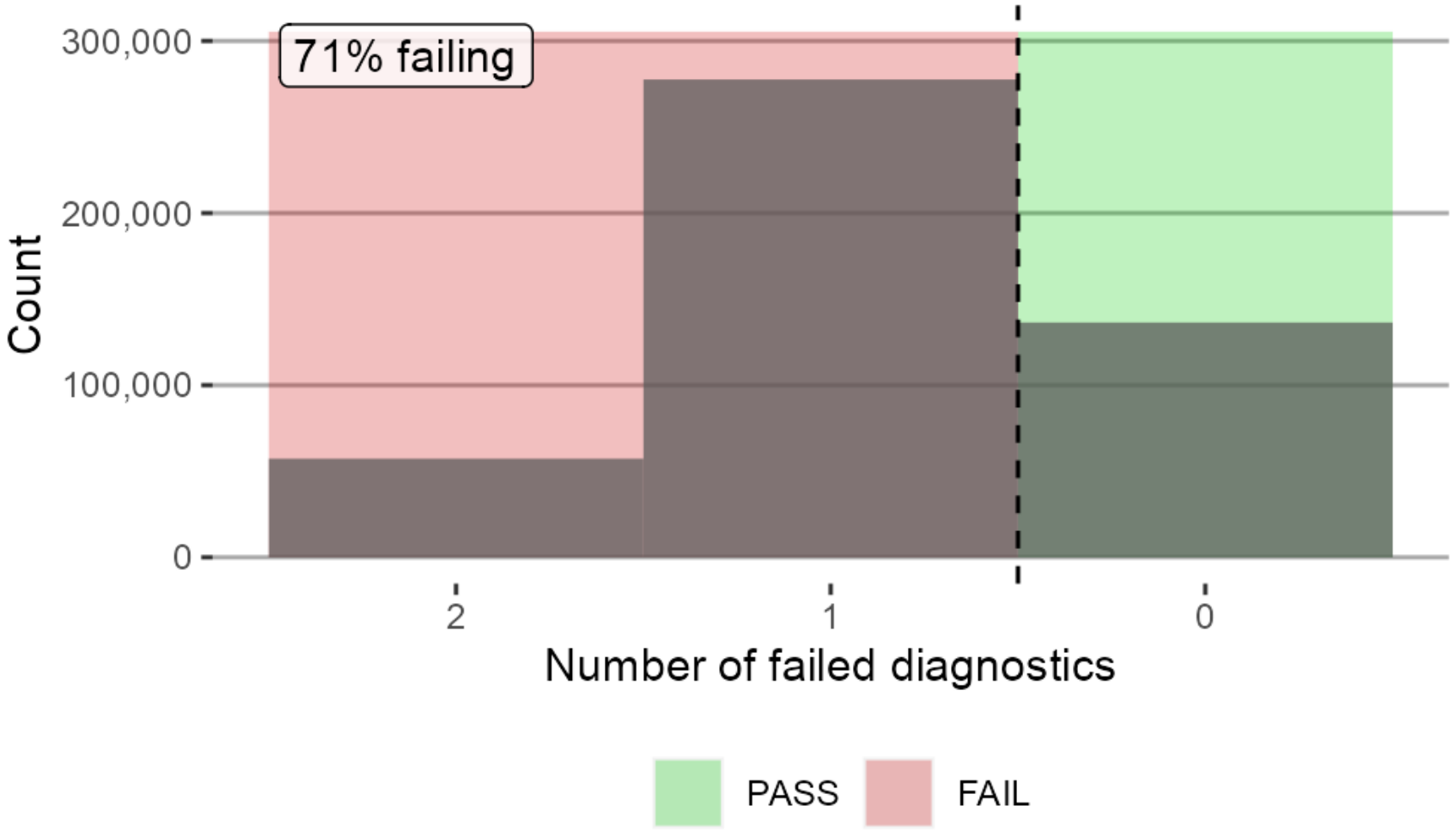
Rules as an optimization problem

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

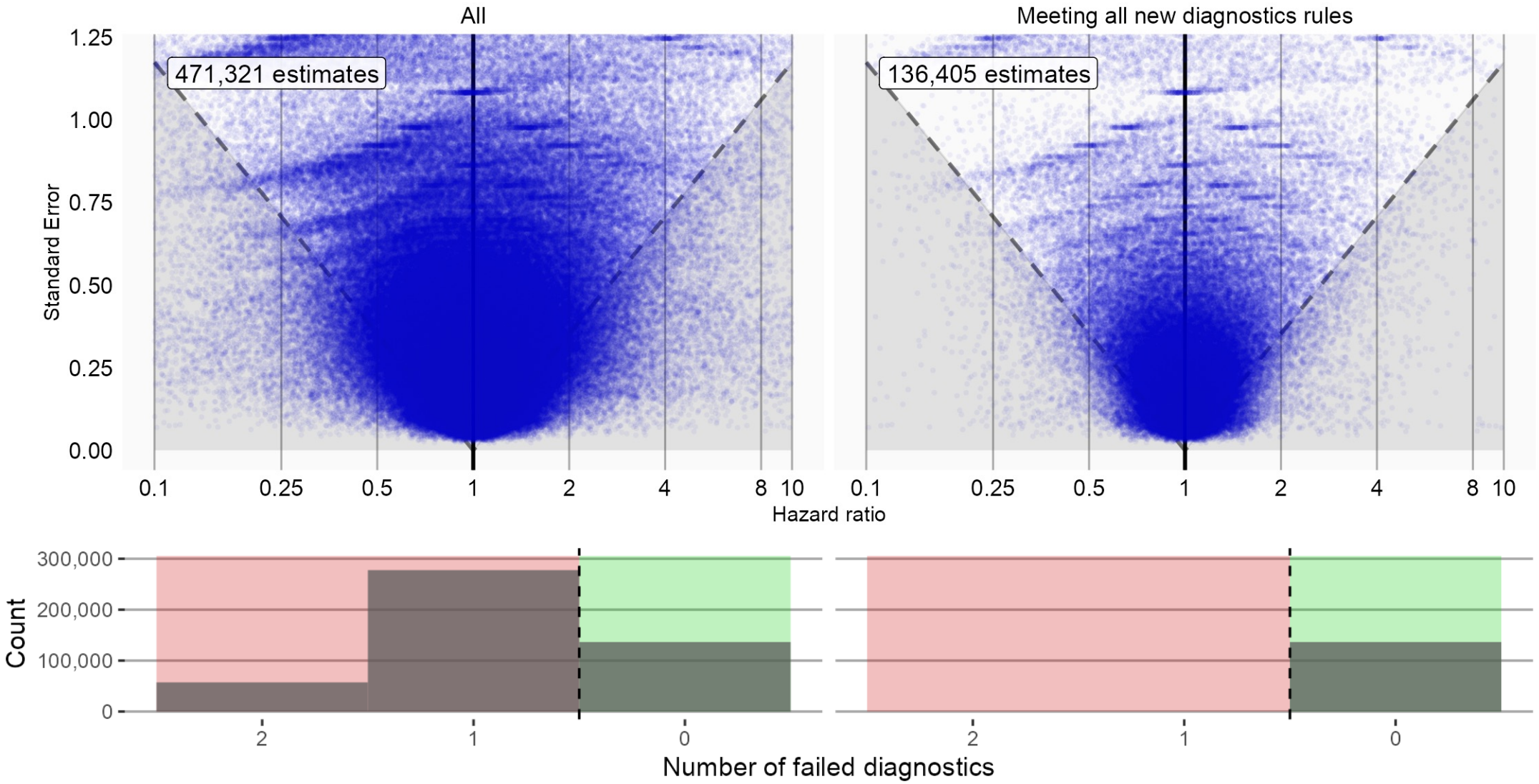
‘Optimal’ thresholds

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 remaining	12%	29%

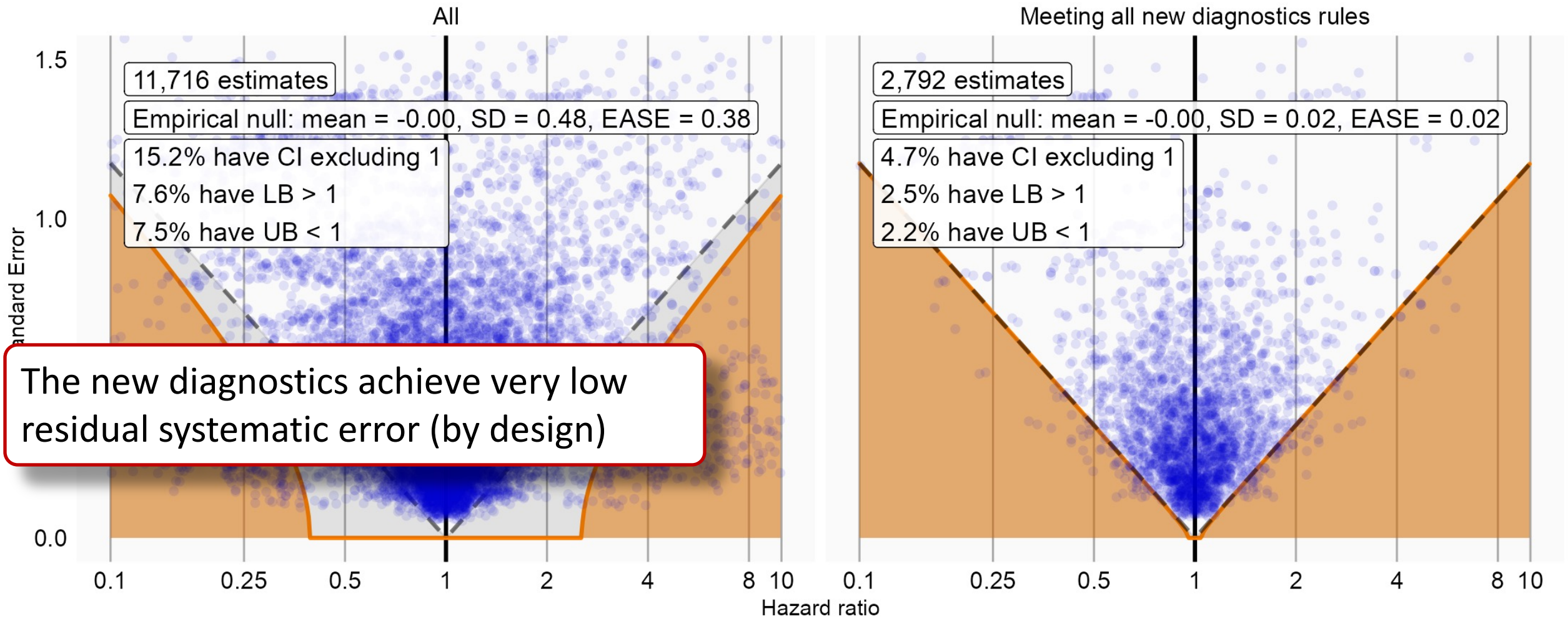
Using data-driven rule set



Using data-driven rule set



Using data-driven rule set



The new diagnostics achieve very low residual systematic error (by design)

Deriving multi-objective decision thresholds empirically

- Diagnostics may reflect different objectives:
  - improving interpretability (MDRR)
  - Reducing systematic error (equipoise, covariate balance, EASE)
  - Ensuring generalizability
- Optimization allows for specifying constraints across all objectives as desired.
- Example: if you want to ensure high generalizability, set max SDM < 0.25

‘Optimal’ thresholds when requiring generalizability

Diagnostic	Literature-derived threshold	Data-driven threshold	Requiring generalizability
Statistical power (MDRR)	10	-	-
Equipoise	0.50	0.50	0.25
Covariate balance (SDM)	0.10	0.50	0.15
Generalizability (SDM)	0.25	-	**0.25***
Systematic error (EASE)	0.25	-	-
Fraction remaining	12%	29%	23%
Likely null set EASE	**<0.05***	**<0.05***	**<0.05***

* Specified constraint

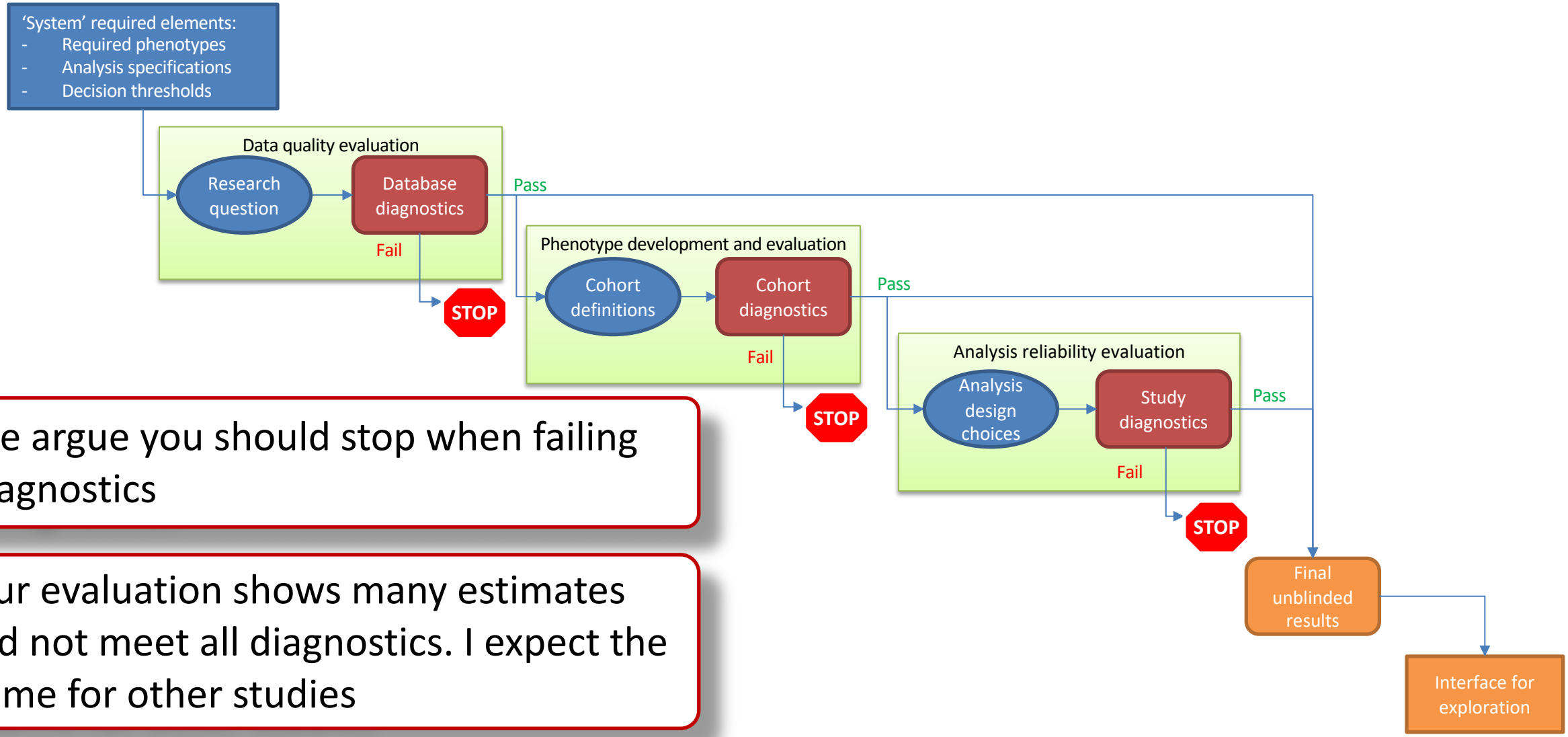
Wrapping up the evaluation of diagnostics

- We've shown some of the 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.

Pre-specification of diagnostic rules

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

Pre-specification of diagnostic rules

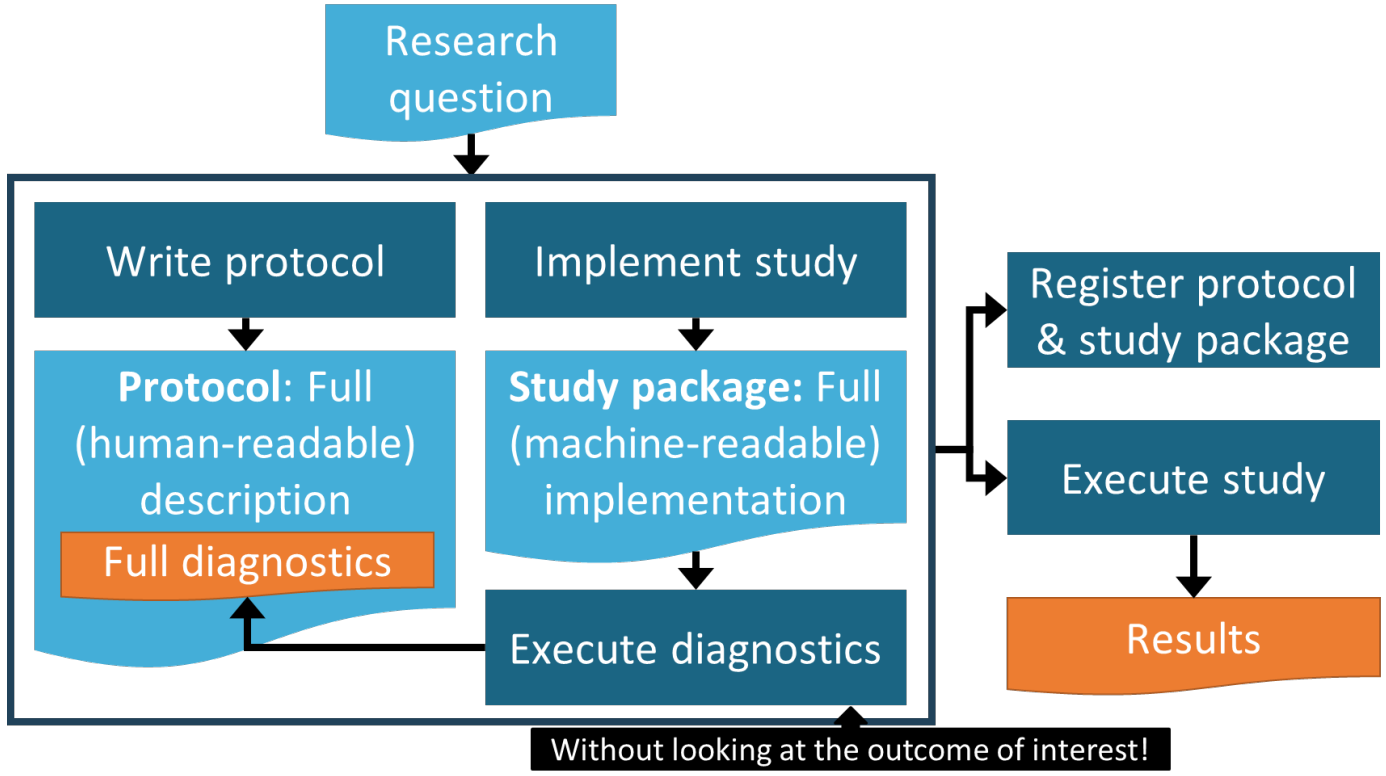


We argue you should stop when failing diagnostics

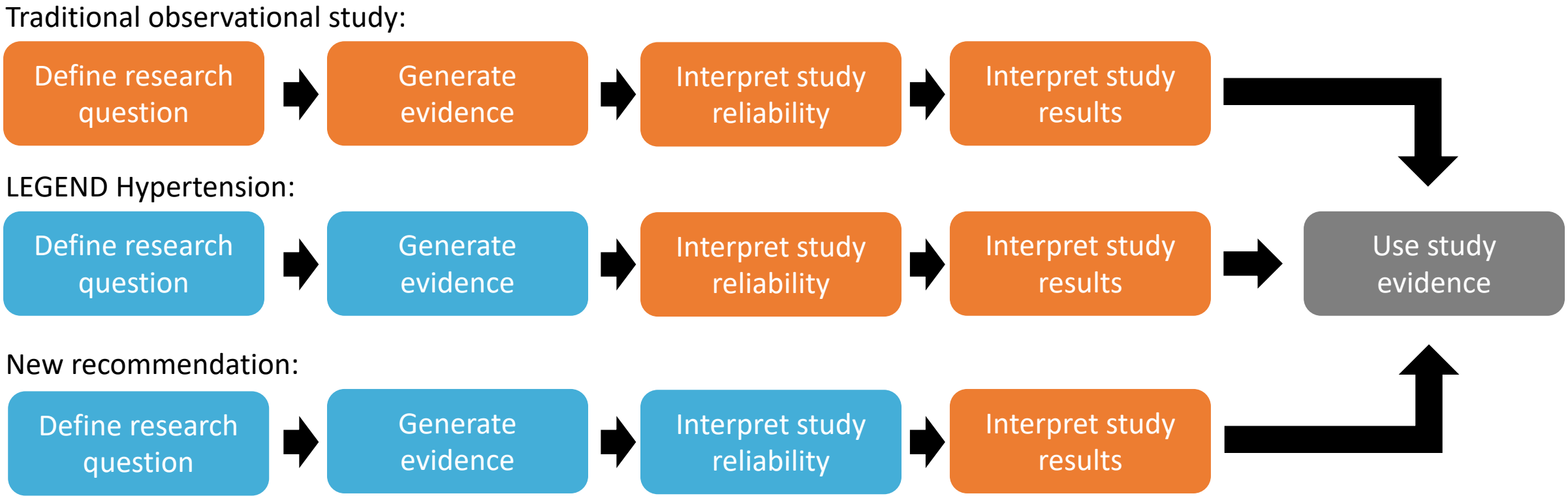
Our evaluation shows many estimates did not meet all diagnostics. I expect the same for other studies

Avoiding investigator bias when interpreting diagnostics

- Diagnostics need to be evaluated prior to looking at the study results
- Protocol can contain **diagnostics results**, or
- Protocol can contain **prespecified diagnostics rules** (So long as they are not modified post-hoc)

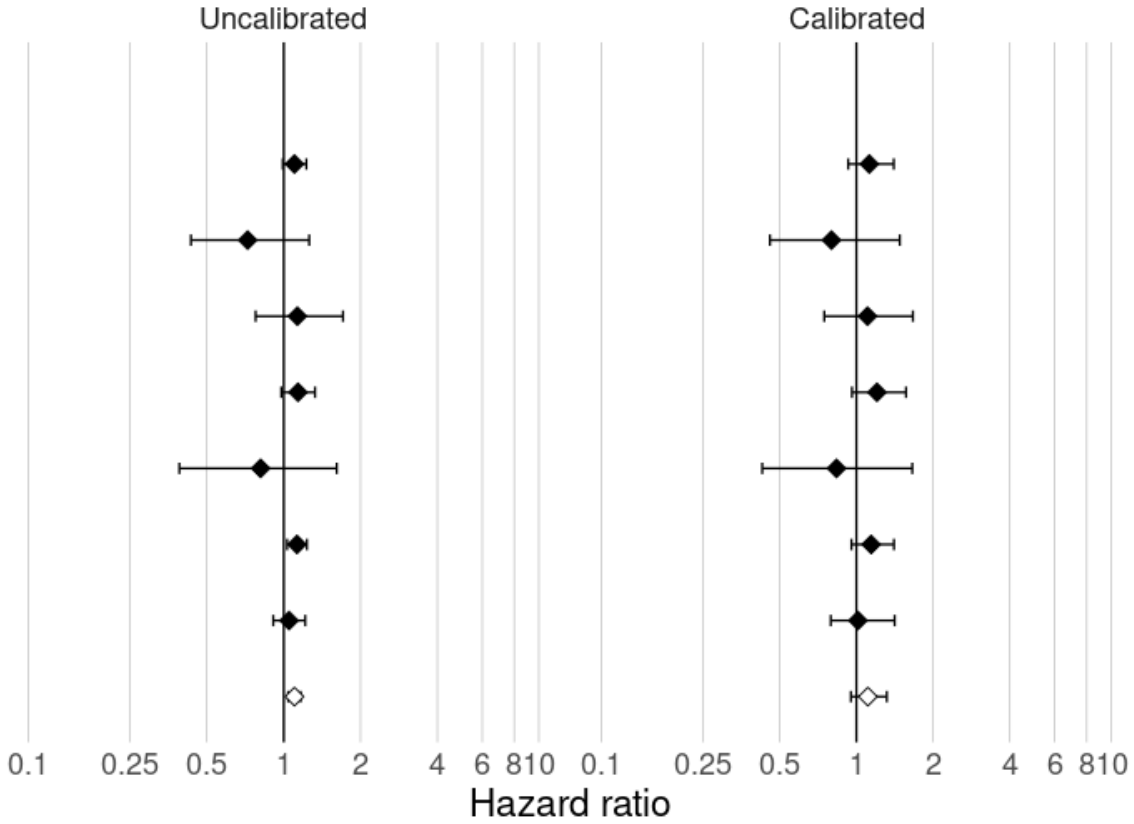


Pre-specification of a systematic approach



Interpreting results from multiple databases

Source	HR (95% CI)	Calibrated HR (95% CI)
CCAE	1.10 (0.99-1.23)	1.12 (0.93-1.40)
CUMC	0.72 (0.43-1.26)	0.80 (0.46-1.48)
MDCD	1.13 (0.78-1.71)	1.11 (0.75-1.67)
MDCR	1.14 (0.98-1.33)	1.20 (0.96-1.57)
NHIS_NSC	0.81 (0.39-1.61)	0.84 (0.43-1.66)
Optum	1.13 (1.03-1.23)	1.14 (0.96-1.40)
Panther	1.05 (0.91-1.21)	1.01 (0.79-1.41)
Summary (I^2 = 0.00)	1.10 (1.04-1.16)	1.11 (0.95-1.32)

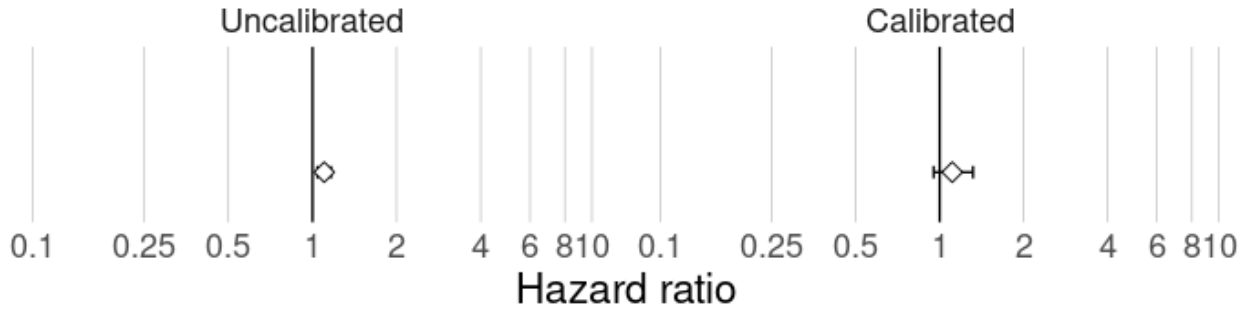


LEGEND Hypertension. ACEs vs ARBs for acute MI using stratification

https://data.ohdsi.org/LegendBasicViewer/

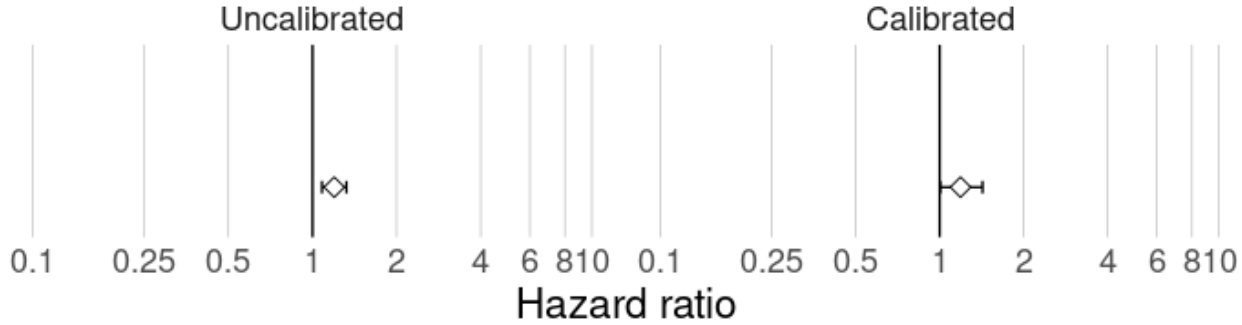
Interpreting results from multiple analyses

Source	HR (95% CI)	Calibrated HR (95% CI)
Summary (I^2 = 0.00)	1.10 (1.04-1.16)	1.11 (0.95-1.32)



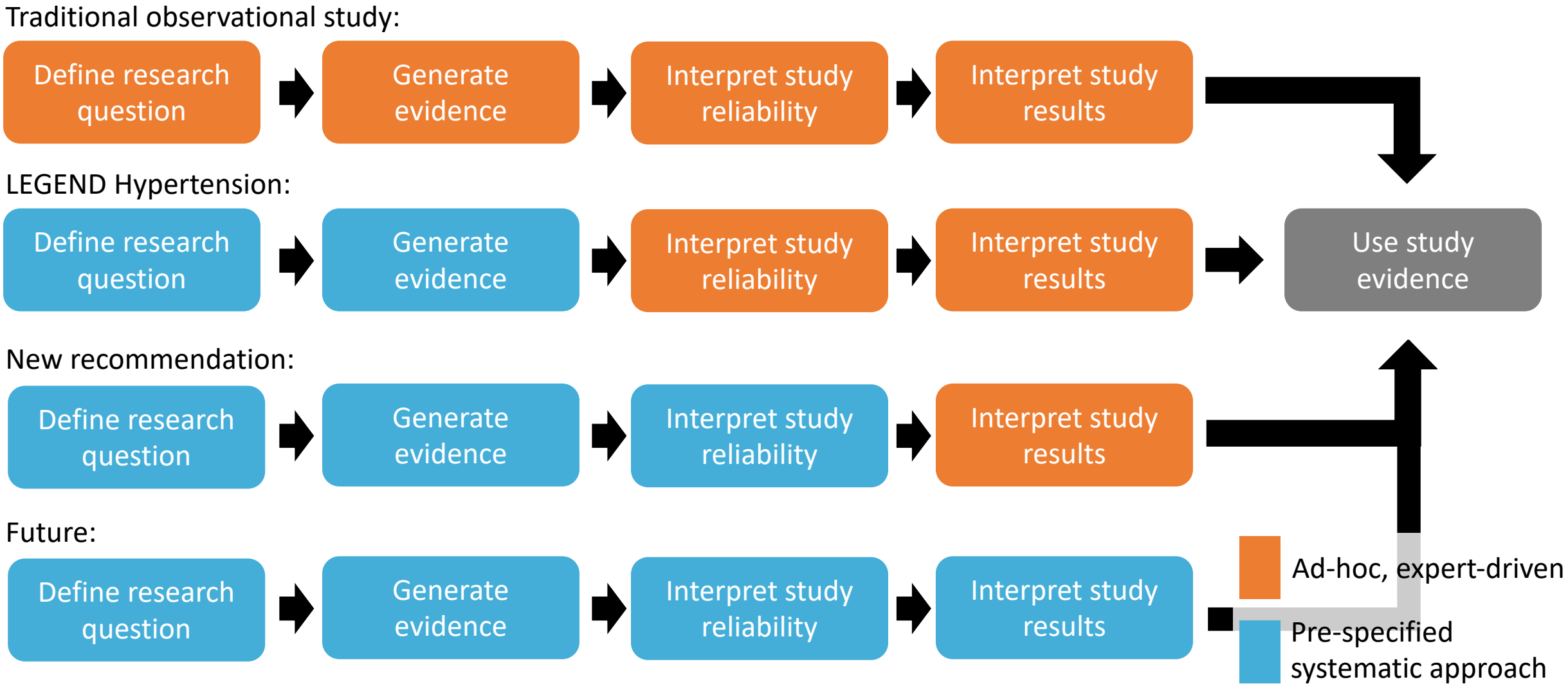
LEGEND Hypertension. ACEs vs ARBs for acute MI using **stratification**

Source	HR (95% CI)	Calibrated HR (95% CI)
Summary (I^2 = 0.00)	1.20 (1.08-1.32)	1.19 (1.01-1.42)



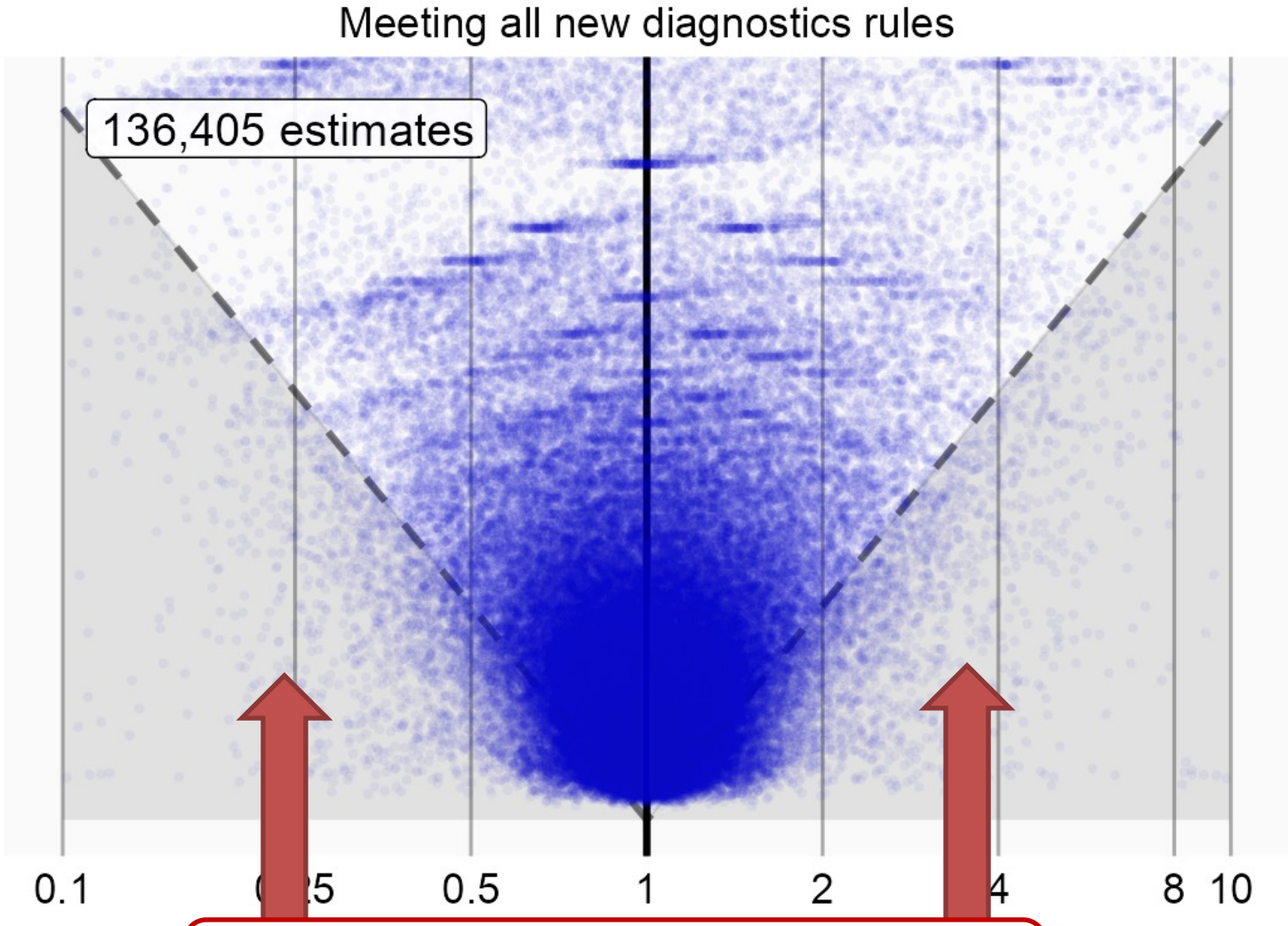
LEGEND Hypertension. ACEs vs ARBs for acute MI using **matching**

Pre-specification of a systematic approach

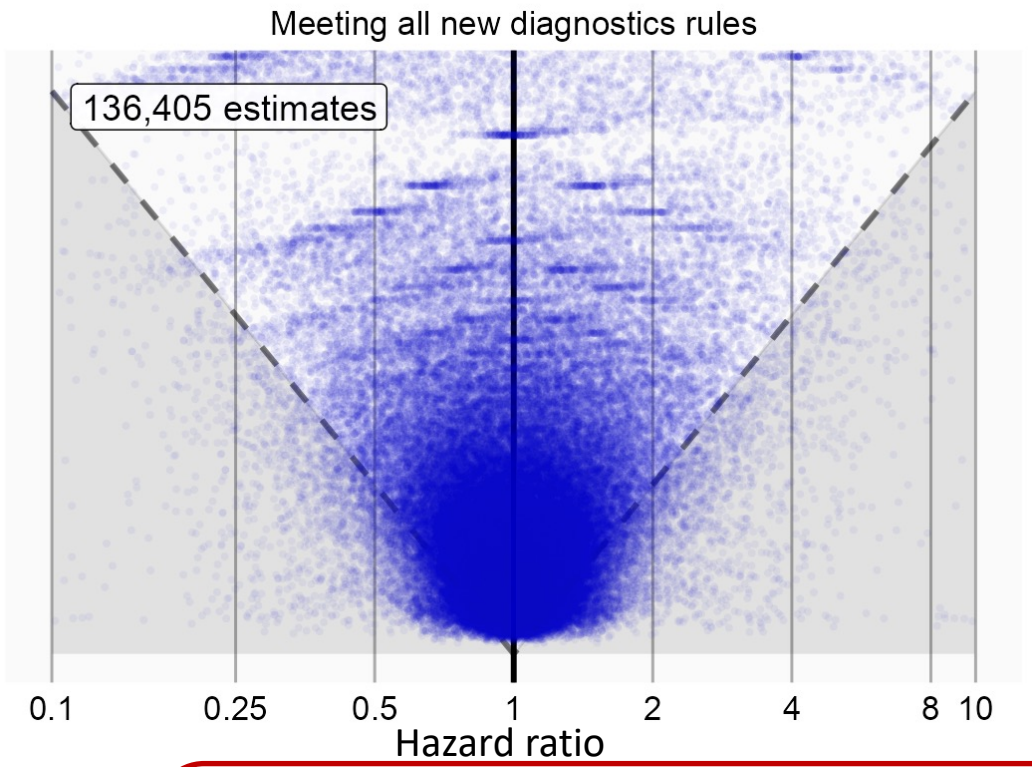


Discovery in causal inference

Effect discovery



Adjustment for multiple testing - estimation



What is the effect of ACEs on AMI compared to ARBs?

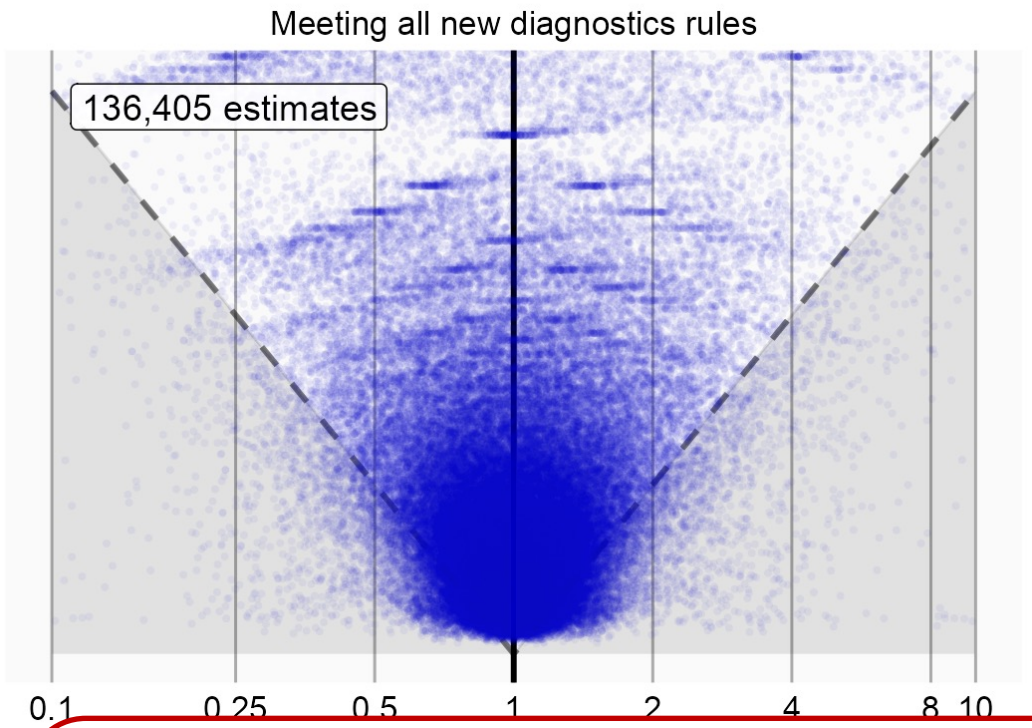
Using alpha = 0.05

Hazard ratio = 1.19 (95% CI: 1.01-1.42)

1 hypothesis, 1 result, 1 test to correct so the doctor has the desired alpha

Note that in literature you'd need to adjust for publication bias and p-hacking

Adjustment for multiple testing - discovery



What effects achieve statistical significance?

Using family-wise alpha = 0.05

List of effects

To answer this question, we have to consider all results. To achieve desired alpha, we must adjust for the total number of estimates (n = 136,405).

Interestingly, because we failed diagnostics for many, we have fewer tests to adjust for!

Discovery in effect surveillance

Assume a surveillance system monitoring

- Multiple treatments
- Multiple outcomes
- Multiple time-at-risks
- Multiple methods
- Multiple databases
- Multiple looks over time

Check out Fan's talk at 1pm!

How best to adjust for multiple testing?

What are the overall operating characteristics we're like to see?

Conclusions

Conclusions

- Diagnostics are important to ensure reliability of results
  - We've been saying that for a while
  - We now have some empirical evidence demonstrating this
- As OHDSI we need to become more rigorous in applying diagnostics
  - Make interpretation of diagnostics a systematic process
  - Either evaluate diagnostics beforehand, or pre-specify diagnostics rules beforehand (just don't look at results that don't pass diagnostics)
- Many studies will fail diagnostics
  - Less 'evidence' is better, when we can trust what remains
  - Disseminate failures. Argues for LEGEND-like studies, where failures are part of result set

Looking forward

- Many opportunities to propose and improve diagnostics
  - Better balance metric? Equipoise? Generalizability?
  - Go / no go rules for data diagnostics? Cohort diagnostics?
- Improving interpretation of results
  - Synthesizing results from multiple databases and multiple analyses
  - Designing a discovery and surveillance system, deciding on what operating characteristics really matter

Thank you!

Joint research with

- Marc Suchard
- Yong Chen
- George Hripcsak
- Others who’ve joined the PLE Workgroup call