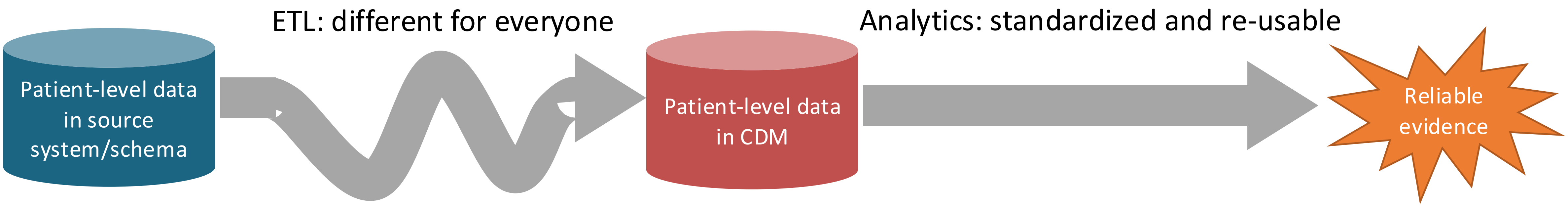
CohortDiagnostics and Population-Level Estimation

Phan Thanh-Phuc PhD
University Medical Center Ho Chi Minh City, Viet Nam

Why convert to the Common Data Model?

- Transforming data to the OMOP CDM is a large investment
- The benefits come from being able to use the same tools and analytics across many databases



Leading example

- Indication:
  - Type-2 diabetes mellitus (T2DM)
- Exposures:
  - GLP-1 agonists
  - DPP-4 inhibitors
- Outcomes:
  - Acute myocardial infarction
  - Diarrhea

OHDSI standardized analytics

- HADES is a set of open-source R package
- Developed and maintained by the community, for the community
- Can use cohort definitions created in ATLAS

ATLAS

966

MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231108

Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research

Martijn SCHUEMIE[a,b,c,1], Jenna REPS[a,b,d], Adam BLACK[a,e], Frank DeFALCO[a,b], Lee EVANS[a,f], Egill FRIDGEIRSSON[a,d], James P. GILBERT[a,b], Chris KNOLL[a,b], Martin

Cohorts of our example

Cohort: a group of people who satisfy some criteria for some period of time

- Indication cohorts:
  - Type-2 diabetes mellitus (**T2DM**) — People with T2DM, while having T2DM
- Exposures cohorts :
  - **GLP-1** agonists — People on GLP-1, while on the drug
  - **DPP-4** inhibitors — People on DPP-4, while on the drug
- Outcomes cohorts :
  - Acute myocardial infarction (**AMI**) — People with AMI, at the time of AMI
  - **Diarrhea** — People with Diarrhea, while having Diarrhea

These same cohorts can be re-used to answer different questions

What type of questions can we ask?

Clinical characterization:
What happened to them?

observation

Patient-level prediction:
What will happen to me?

inference

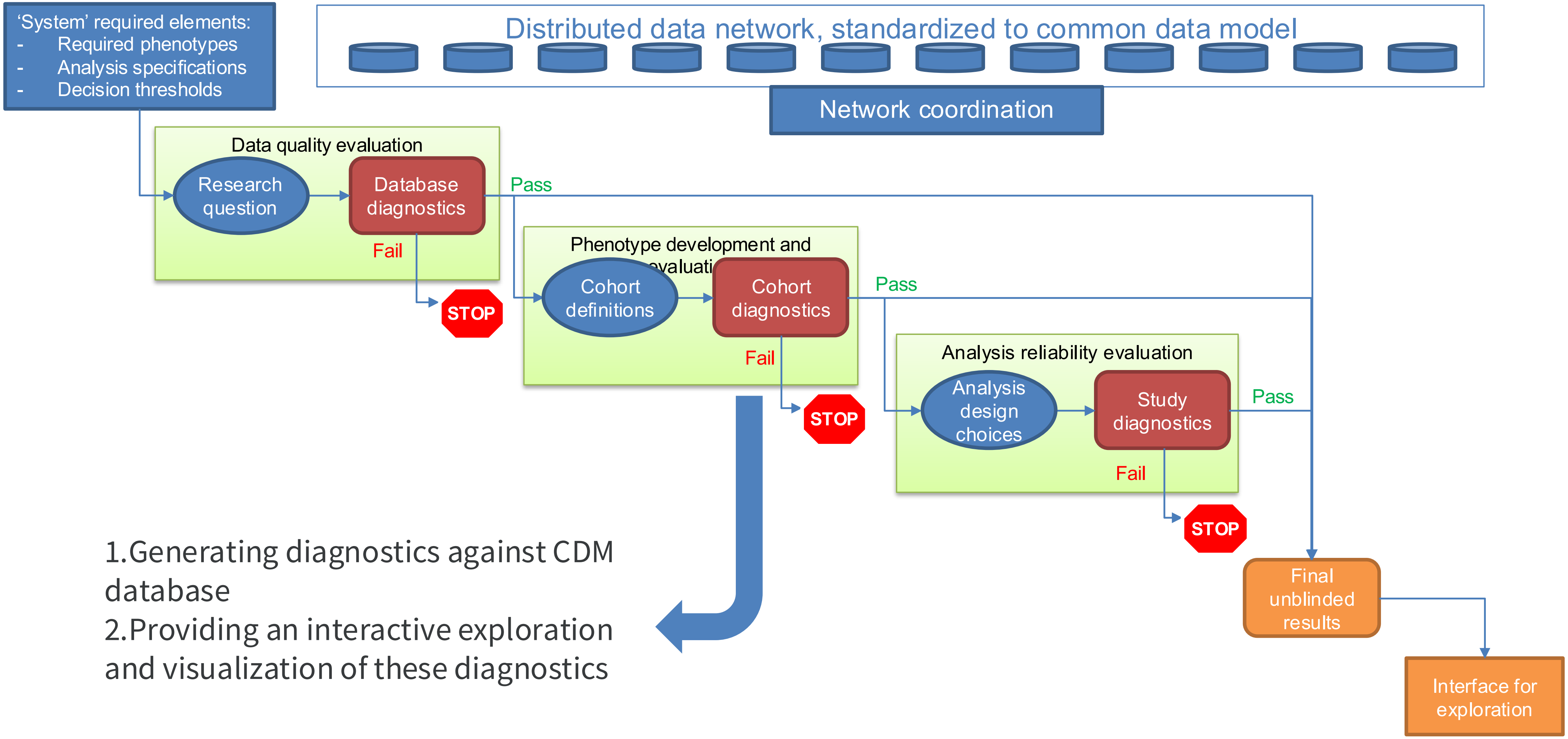
Population-level effect estimation:
What are the causal effects?

causal inference

Cohort Dianogstic

Using OHDSI tools

Engineering open science systems that build trust into the RWE generation and dissemination process



1. Generating diagnostics against CDM database
2. Providing an interactive exploration and visualization of these diagnostics

CohortDiagnostics utilities

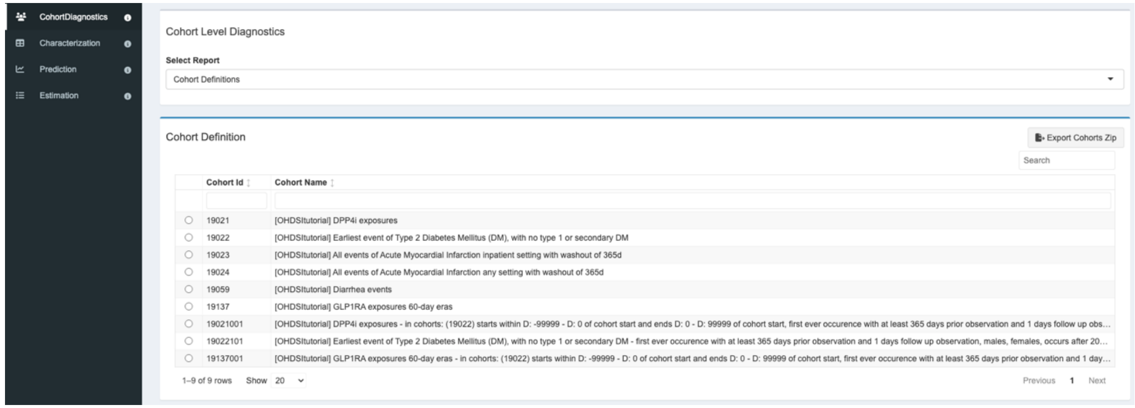
1. Enhancing Cohort Definition Confidence
2. Identifying Missing Concepts & Cohort Entry Events
3. Facilitating the Ideas Behind Comparative Analyses
4. Supporting Transparent Research

Features

1. Show cohort inclusion-rule attrition.
2. List all source codes used in a cohort definition.
3. Identify orphan codes missing from a concept set.
4. Compute cohort incidence by year, age, gender.
5. Break down index events by triggering concepts.
6. Measure cohort overlap.
7. Characterize cohorts and compare (including temporal comparisons).
8. Inspect patient profiles from a random cohort sample.

Example questions

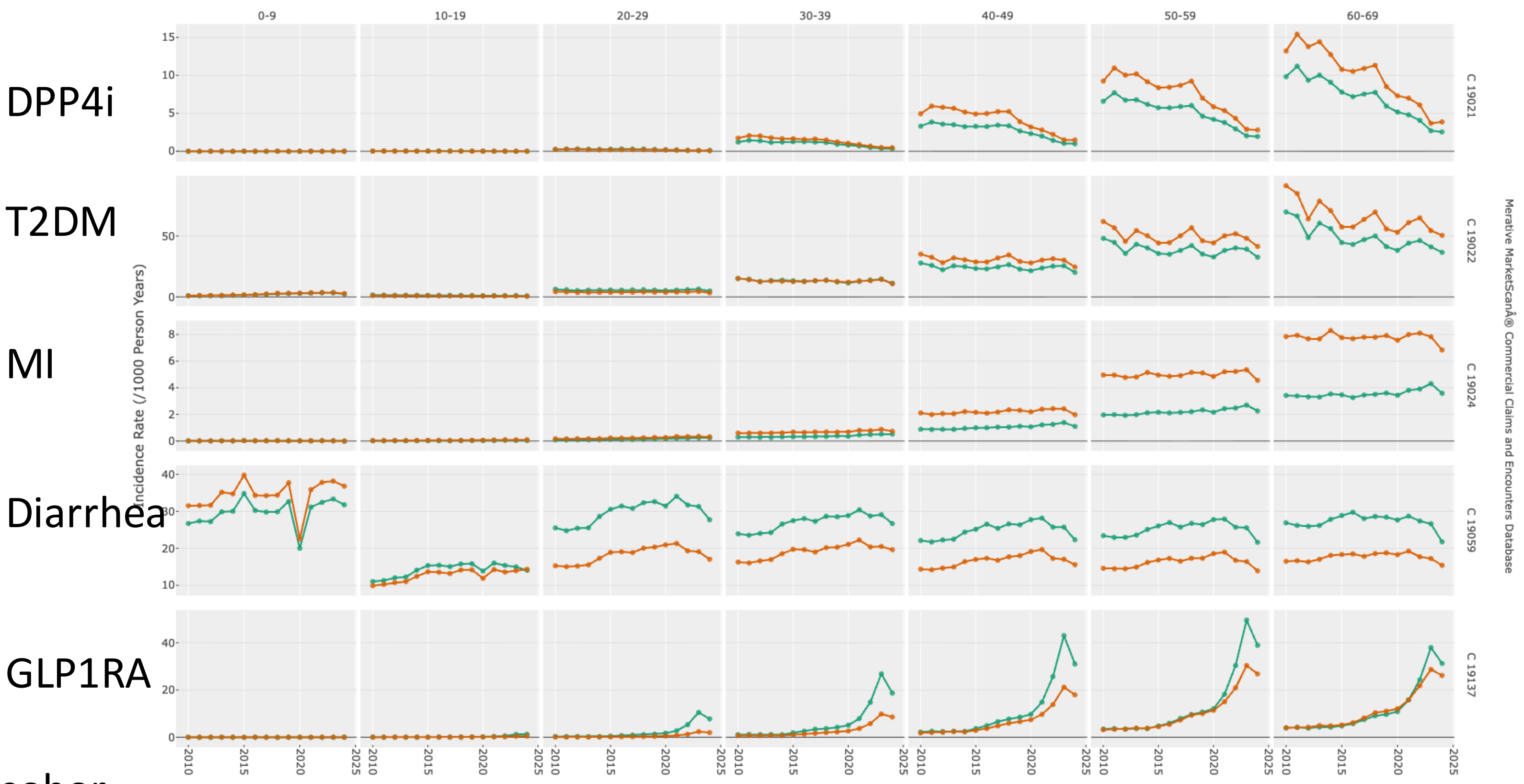
- How did the rate of AMI in patients with T2DM change over time?
- What other drugs to DPP-4 users use?



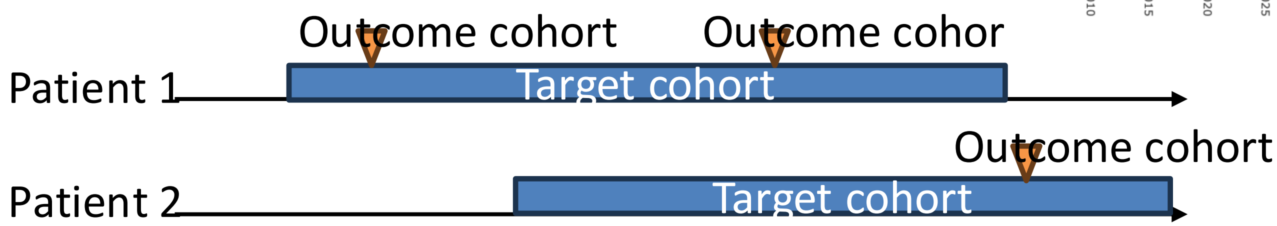
Cohort Incidence

Computes the incidence rate of the Outcome cohort in some Target cohort

- Standardized computation of incidence rates
- Default: overall and stratified by age, sex, and calendar time



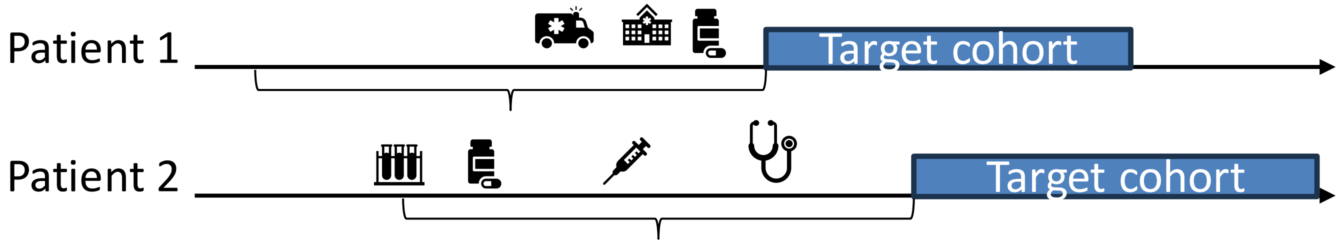
- How did the rate of AMI in patients with T2DM change over time?
  - Target: **T2DM**
  - Outcome: **AMI**



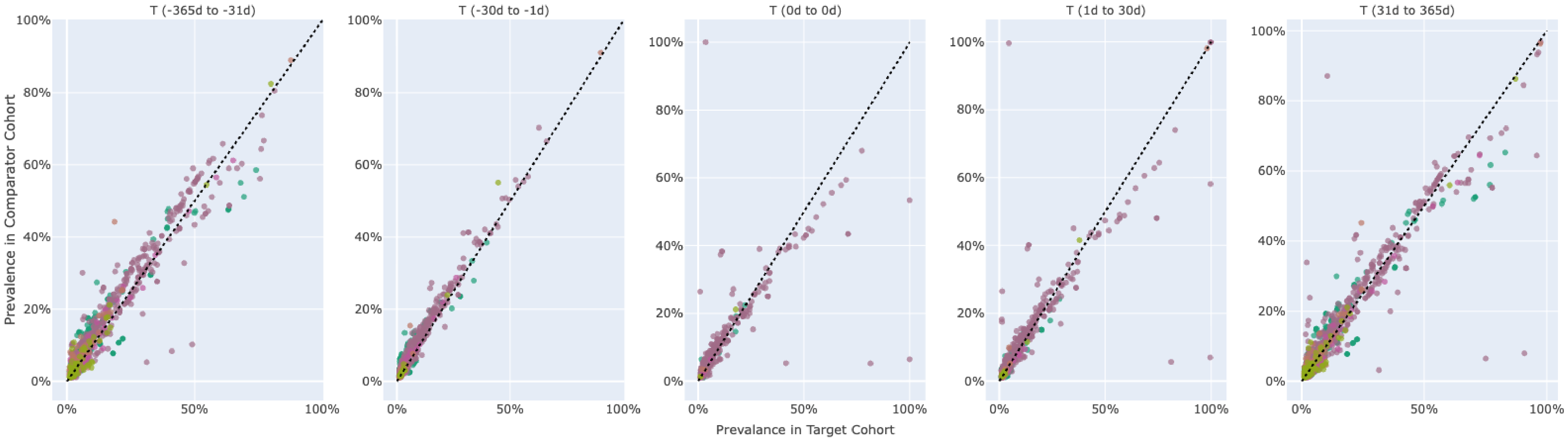
Cohort Characterization

Counts all observed events (concepts) relative to Target cohort start, etc.

- Additional analyses include time-to-event, risk factors, case series



- What other drugs to DPP-4 users use?
  - Target: **T2DM**



R setup

- Follow our R HADES setup guide for getting an R environment set up
- Almost all code blocks can be copy pasted

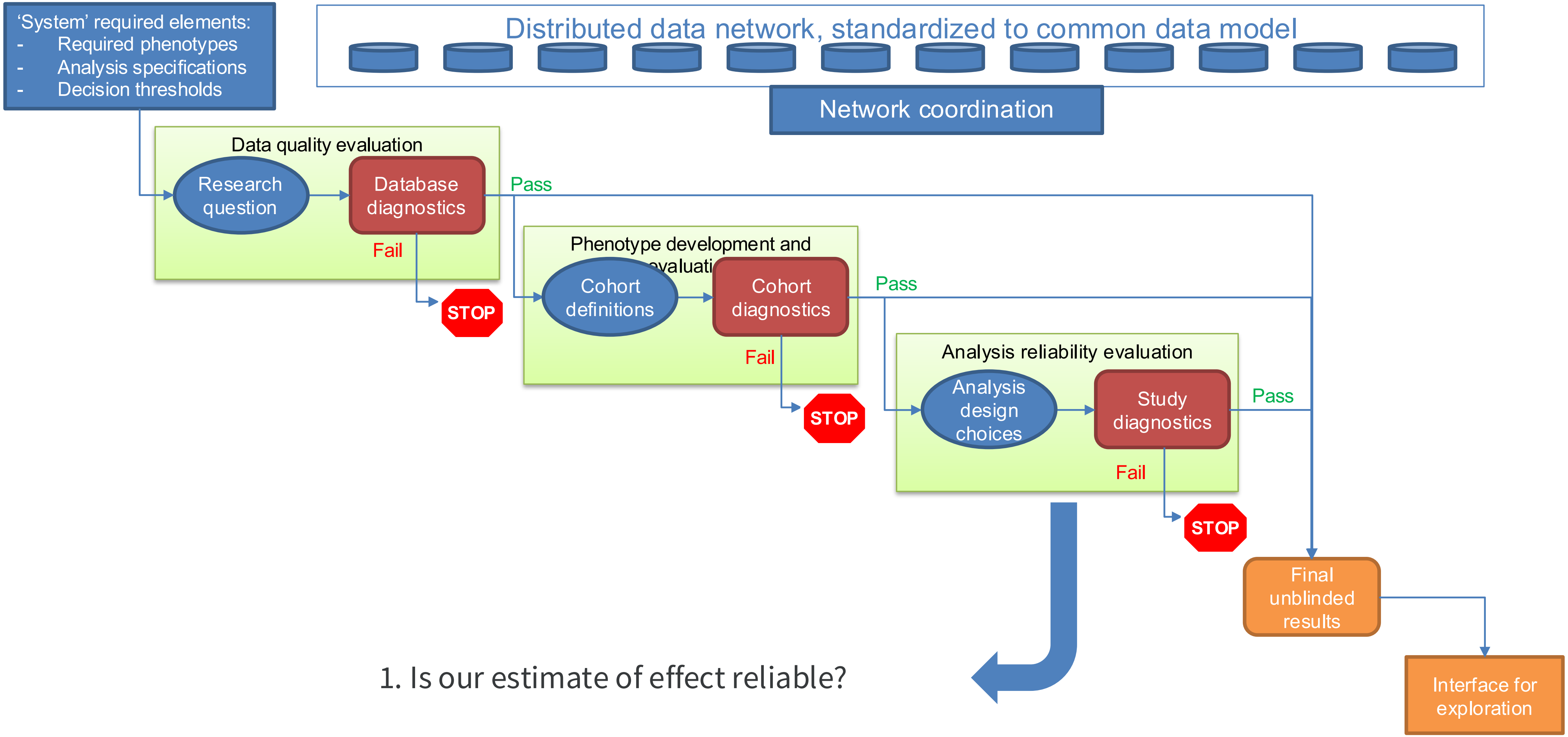
https://ohdsi.github.io/CohortDiagnostics/

- Download the Rproject from Github

Causal effect estimation

Using OHDSI tools

Engineering open science systems that build trust into the RWE generation and dissemination process



Example causal effect estimation questions

- Does exposure to GLP-1 antagonists decrease the risk of AMI?
- Does exposure to GLP-1 antagonists decrease the risk of AMI compared to DPP-4 inhibitors?

Can be answered using

- SelfControlledCaseSeries package
- CohortMethod package

CohortMethod package

Computes the hazard of the Outcome cohort in the Target cohort compared to the Comparator

- Does exposure to GLP-1 antagonists decrease the risk of AMI compared to DPP-4 inhibitors?
  - Target: **GLP-1**, restricted to those with **T2DM** (and first use only)
  - Comparator: **DPP-4**, restricted to those with **T2DM** (and first use only)
  - Outcome: **AMI**

Unique feature: Large-scale propensity scores

- Treatment assignment is often non-random, which can cause confounding
  - E.g. GLP-1 may be prescribed more often to obese, who already have a higher risk of AMI
- Propensity scores are an establish way to address this
  - Fit a model to predict treatment assignment, and use to compute probability (propensity score)
  - Match subjects in Target to Comparator with similar propensity scores
- Traditionally, expert pick a few variables to use in the prediction model
- Large-scale propensity scores include all baseline covariates, and uses regularized regression (LASSO)

Demonstrating large-scale propensity scores

- Comparing paracetamol to ibuprofen
- CPRD database
- Propensity score matching
  - 37 'publication covariates'
  - 'Large-scale covariates' + LASSO

Large-scale covariates:

- Demographics
- Conditions
- Drugs
- Lab values
- Procedures
- …

Typically between 10,000 and 100,000 variables

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

CrossMark

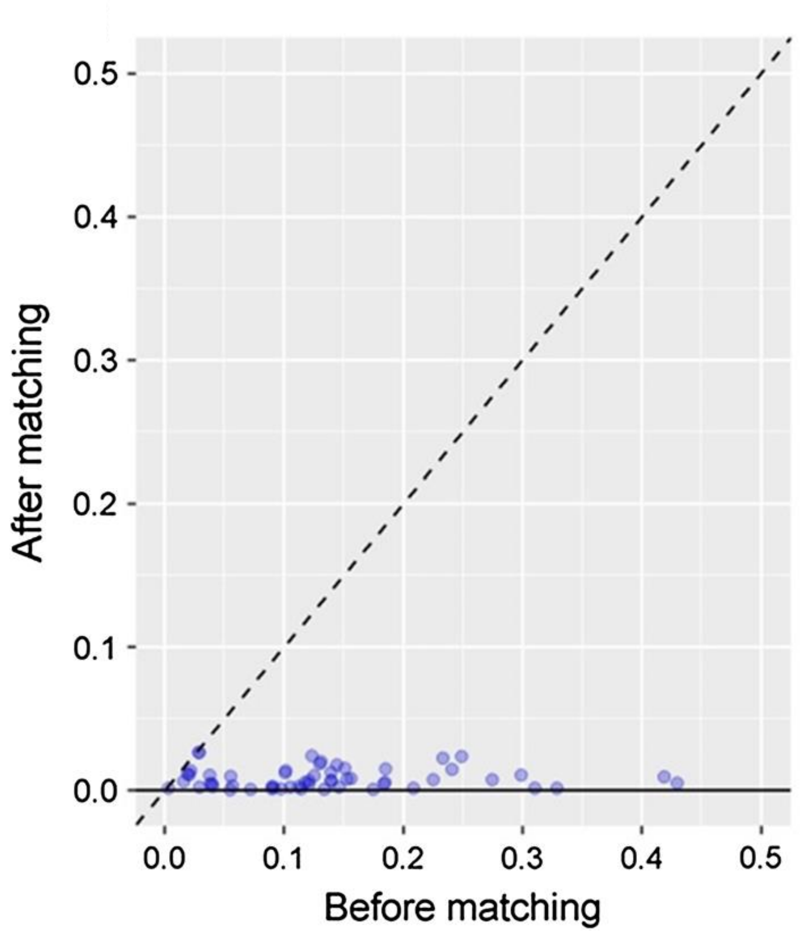
ORIGINAL RESEARCH ARTICLE

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

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

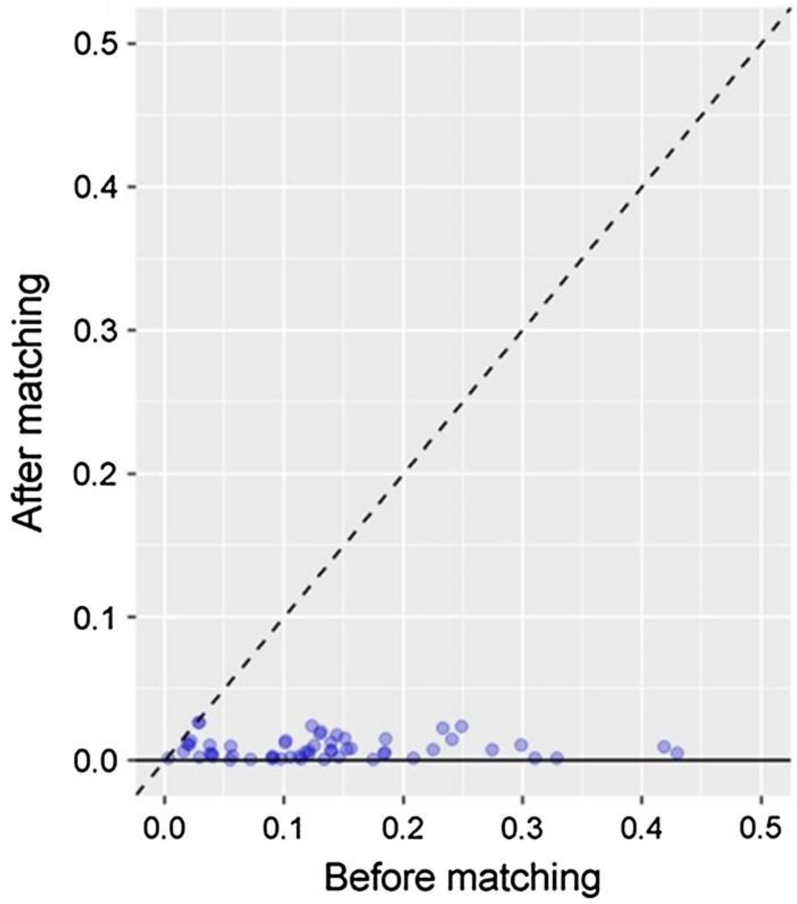
Covariate balance: standardized difference of means

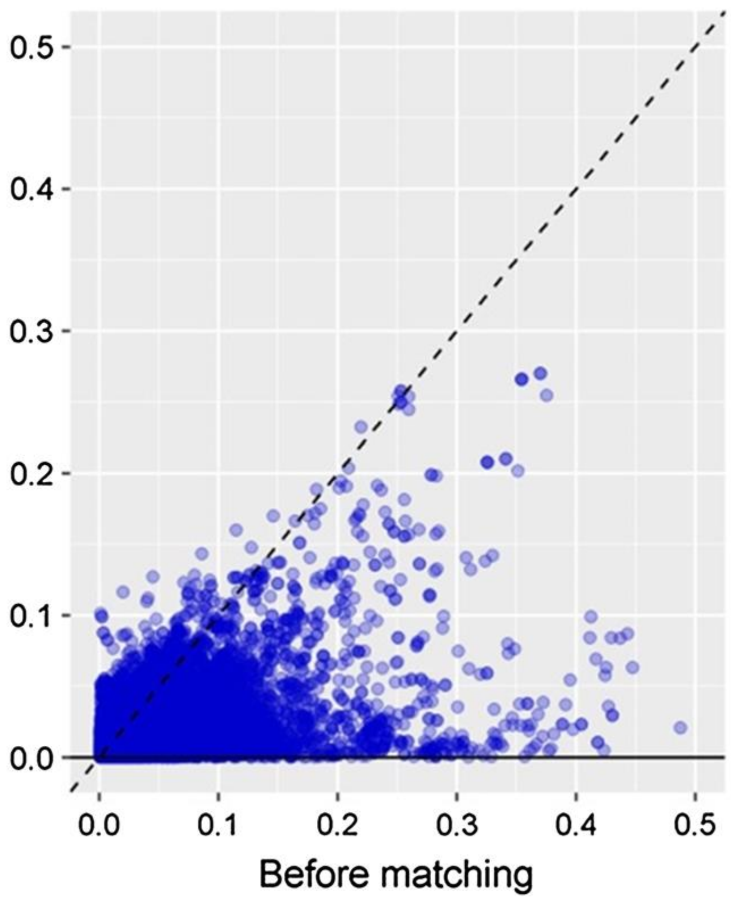
Shown: Publication covariates
PS: Publication covariates



Covariate balance: standardized difference of means

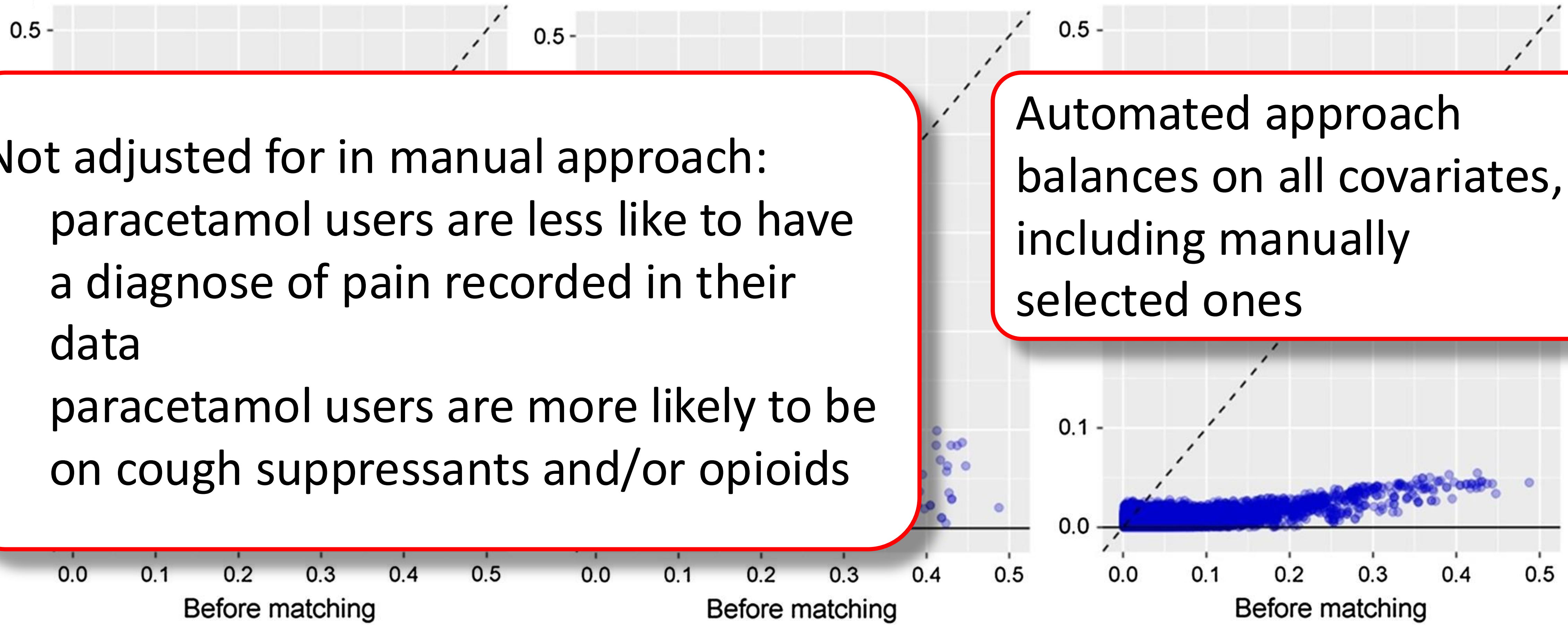
Shown:	Publication covariates	Shown:	Large-scale covariates
PS:	Publication covariates	PS:	Publication covariates





Covariate balance: standardized difference of means

Shown:	Publication covariates	Shown:	Large-scale covariates	Shown:	Large-scale covariates
PS:	Publication covariates	PS:	Publication covariates	PS:	Large-scale covariates



Not adjusted for in manual approach:

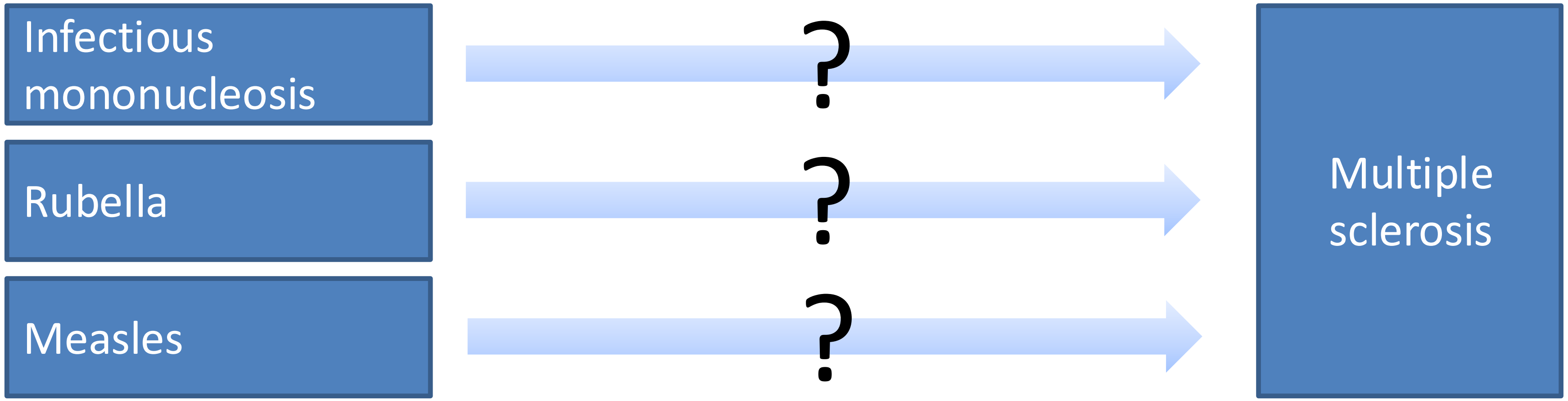
- paracetamol users are less like to have a diagnose of pain recorded in their data
- paracetamol users are more likely to be on cough suppressants and/or opioids

Automated approach balances on all covariates, including manually selected ones

Unique feature: objective diagnostics

- Whether study results are reliable depends on whether certain assumptions have been met
  - E.g. we assume our PS adjustment makes our treatment groups comparable
- Most of these assumptions are testable through diagnostics
  - E.g. we can test whether our PS adjustment achieved balance by computing the standardized difference of means (SDM)
- By 'objective' diagnostics we mean diagnostics that are evaluated while blinded to the results of the study
  - E.g. Pre-specify that we will not look at results where max(|SDM|) > 0.1
  - Unique: negative controls

Example of a negative control



RESEARCH PAPER

Multiple Sclerosis 2008; **14**: 307–313

Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control

Odds ratio:

Infectious mononucleosis → 2.22 * → Multiple sclerosis

Rubella → 1.31 * → Multiple sclerosis

Measles → 1.42 * → Multiple sclerosis

* $P < .05$

RESEARCH PAPER *Multiple Sclerosis* 2008; **14**: 307–313

Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control

Odds ratio:

Infectious mononucleosis → 2.22 * → Multiple sclerosis

Rubella → 1.31 * → Multiple sclerosis

Measles → 1.42 * → Multiple sclerosis

Negative controls:

A broken arm → 1.10 → Multiple sclerosis

Concussion → 1.23 * → Multiple sclerosis

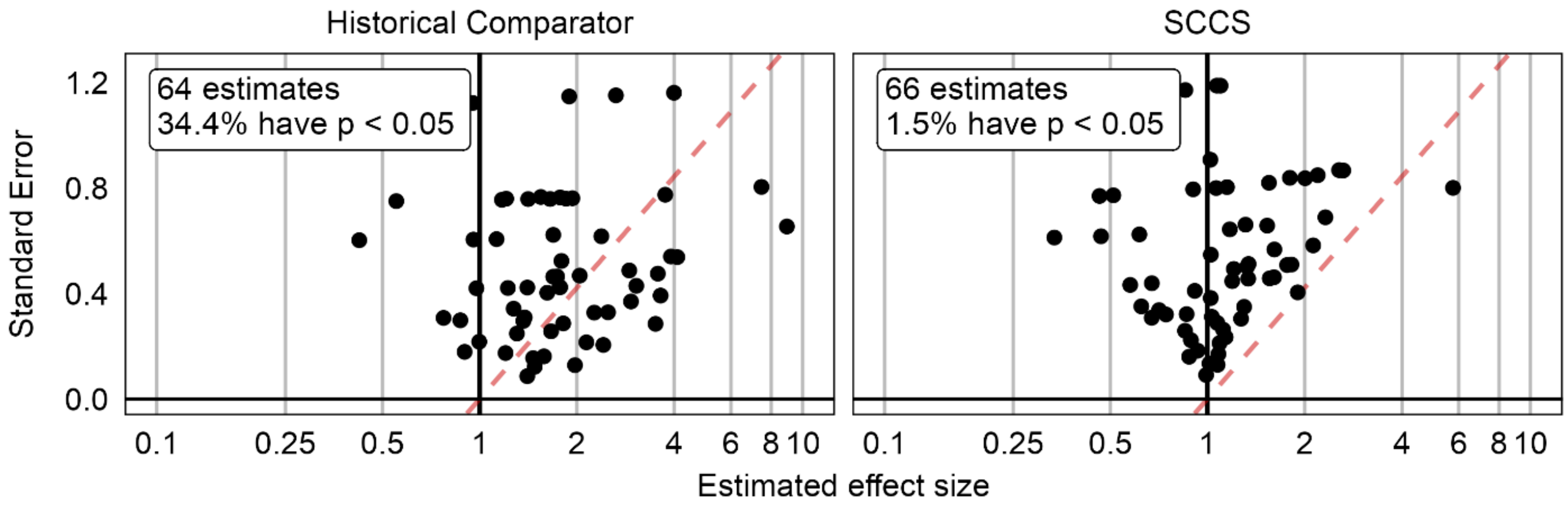
Tonsillectomy → 1.25 * → Multiple sclerosis

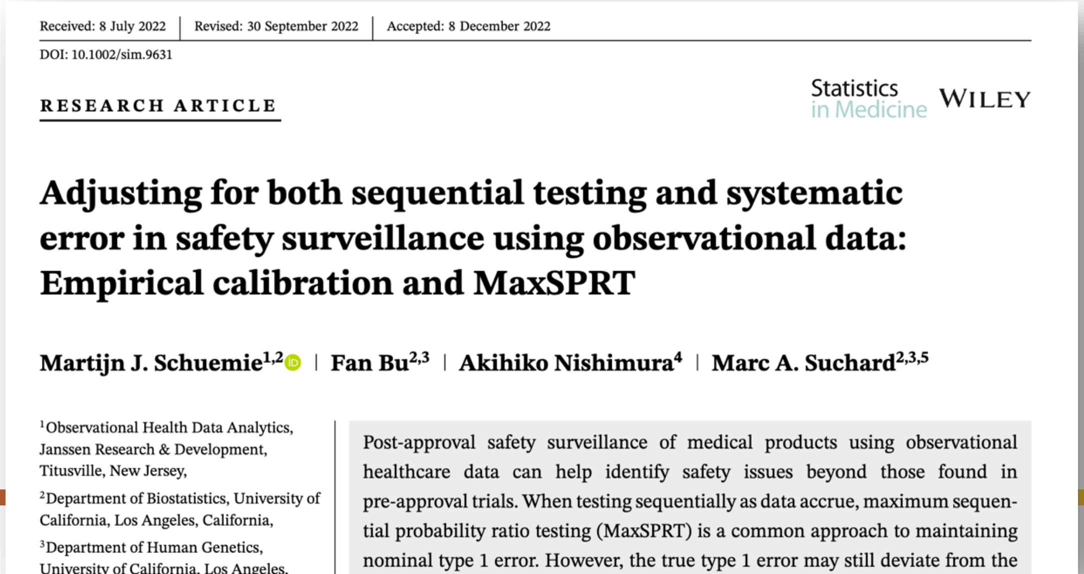
* P < .05

How to interpret negative control findings?

- Unique: use a sample (n > 50) of negative controls to understand distribution of bias
- Systematic error distribution can be used as
  - Diagnostic: if too much systematic error, we stop
  - Calibration: can adjust p-values and confidence intervals to take into account possible systematic error

Quantifying systematic error





Received: 8 July 2022 | Revised: 30 September 2022 | Accepted: 8 December 2022

DOI: 10.1002/sim.9631

RESEARCH ARTICLE

Statistics in Medicine WILEY

Adjusting for both sequential testing and systematic error in safety surveillance using observational data: Empirical calibration and MaxSPRT

Martijn J. Schuemie[1,2] | Fan Bu[2,3] | Akihiko Nishimura[4] | Marc A. Suchard[2,3,5]

[1]Observational Health Data Analytics, Janssen Research & Development, Titusville, New Jersey,

[2]Department of Biostatistics, University of California, Los Angeles, California,

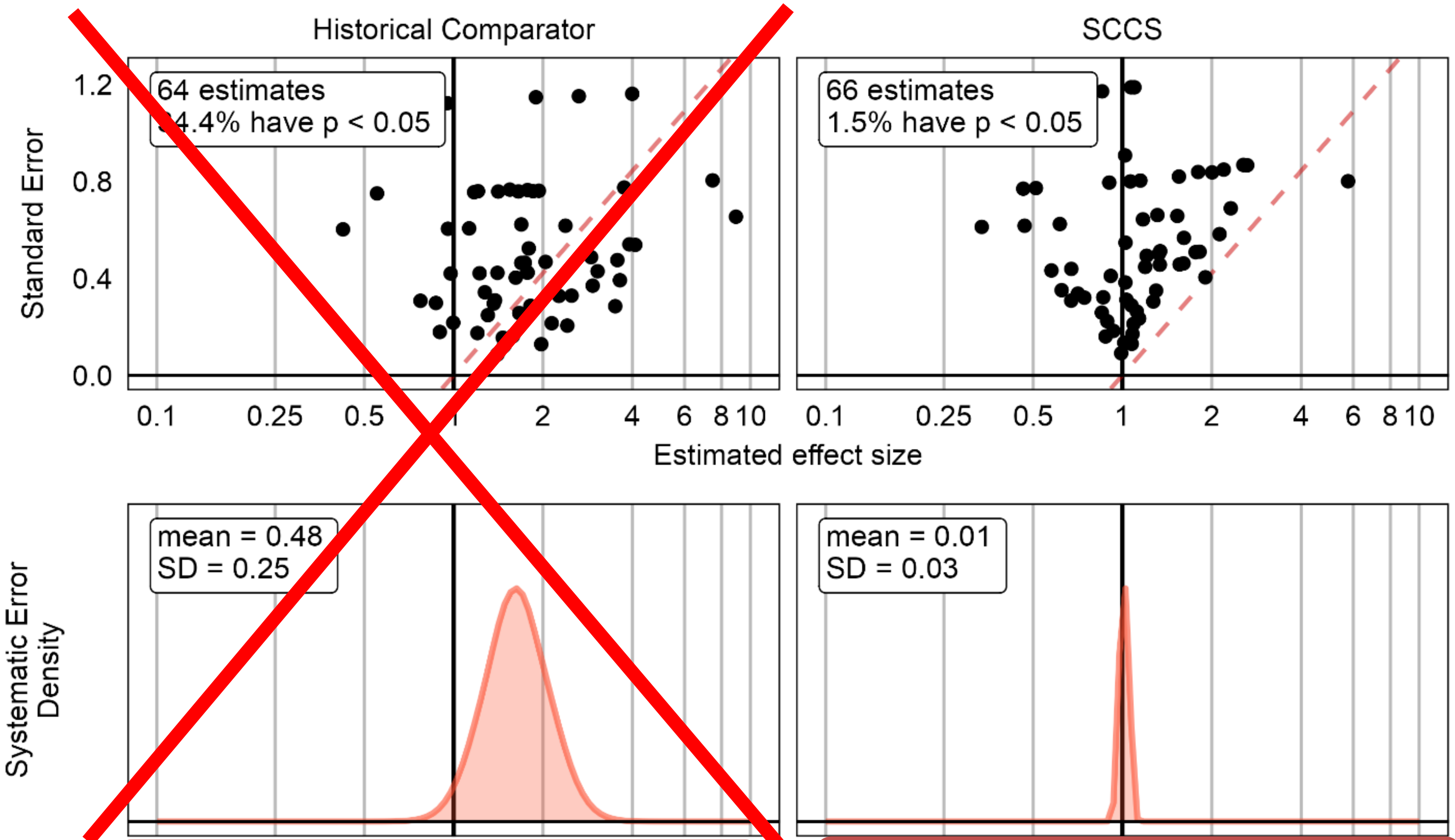
[3]Department of Human Genetics, University of California, Los Angeles,

Post-approval safety surveillance of medical products using observational healthcare data can help identify safety issues beyond those found in pre-approval trials. When testing sequentially as data accrue, maximum sequential probability ratio testing (MaxSPRT) is a common approach to maintaining nominal type 1 error. However, the true type 1 error may still deviate from the

Quantifying systematic error

Expected Absolute Systematic Error (**EASE**) summarizes this distribution

We use a **prespecified** EASE threshold (EASE < 0.25) for go – no go decisions for our studies



EASE = 0.49

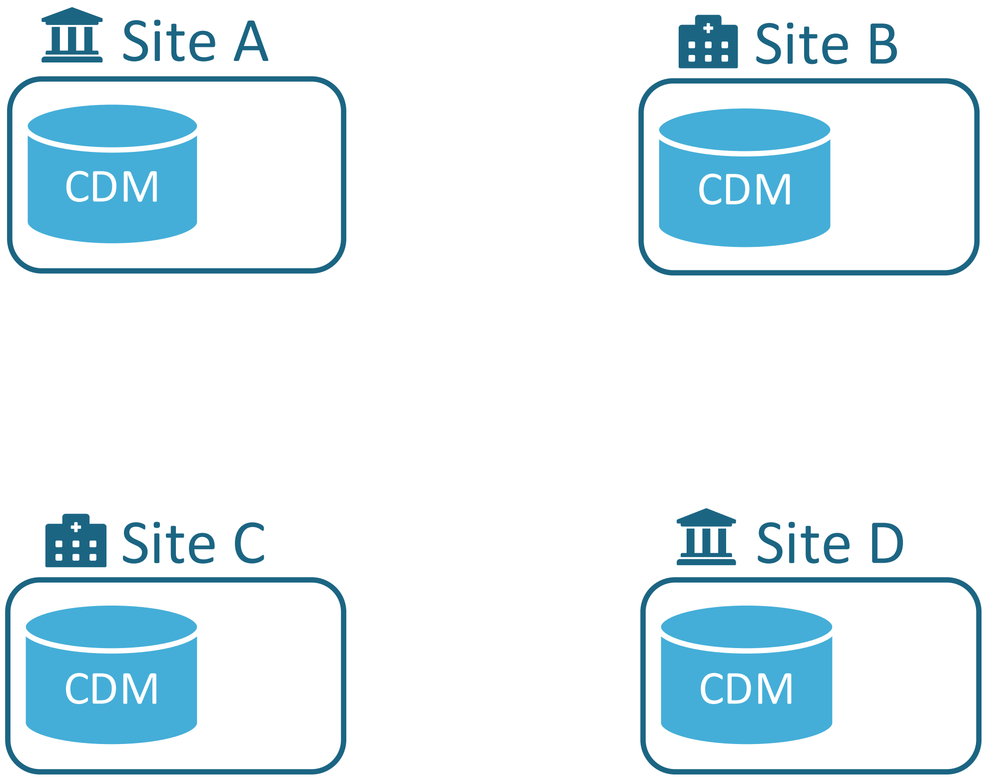
EASE = 0.04

Distributed analyses

Using OHDSI tools

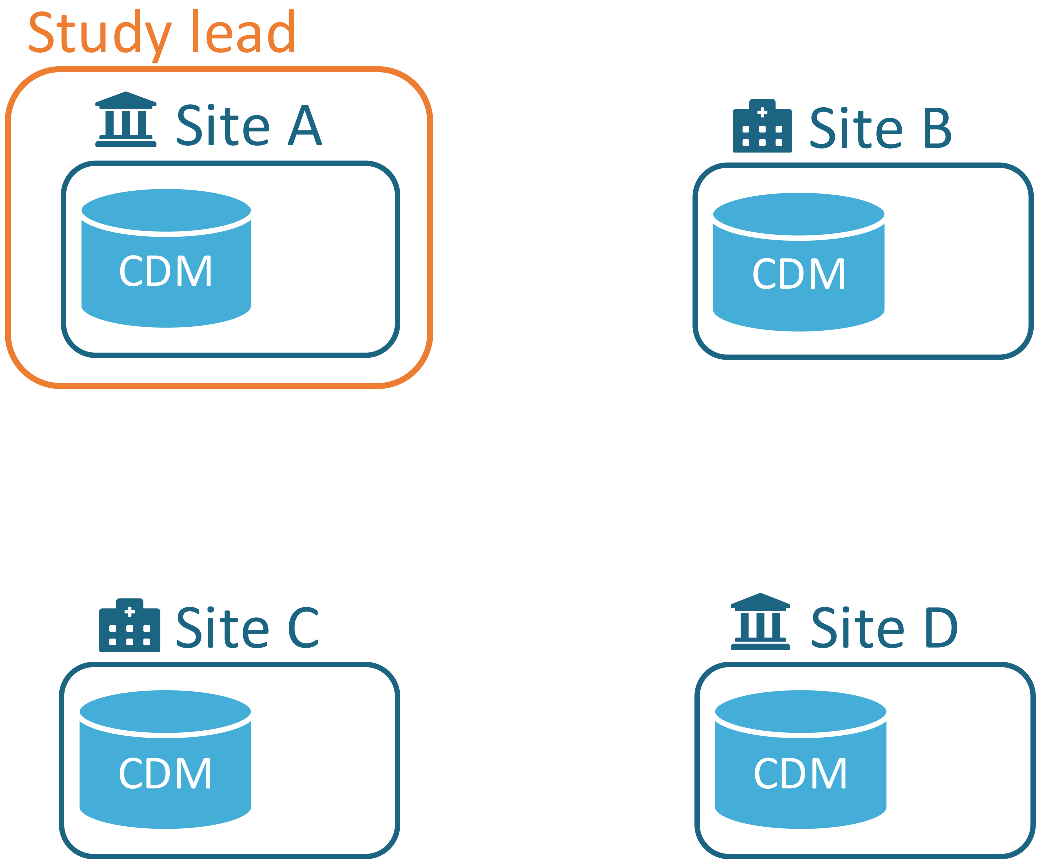
Distributed Research Network

- Multiple sites with data
  - Hospital EHRs (Electronic Health Records)
  - Administrative Claims
- Patient-level data cannot be shared
- Each site uses the Common Data Model (CDM)



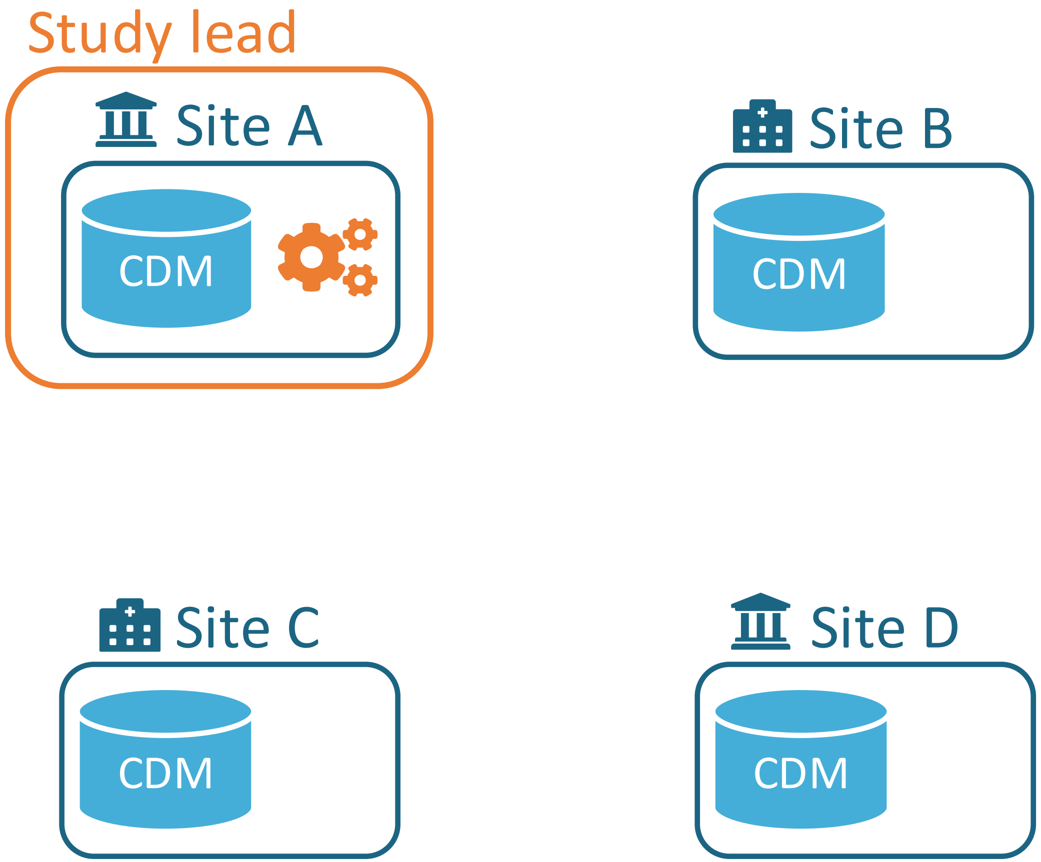
Distributed Research Network

- A site can lead a study



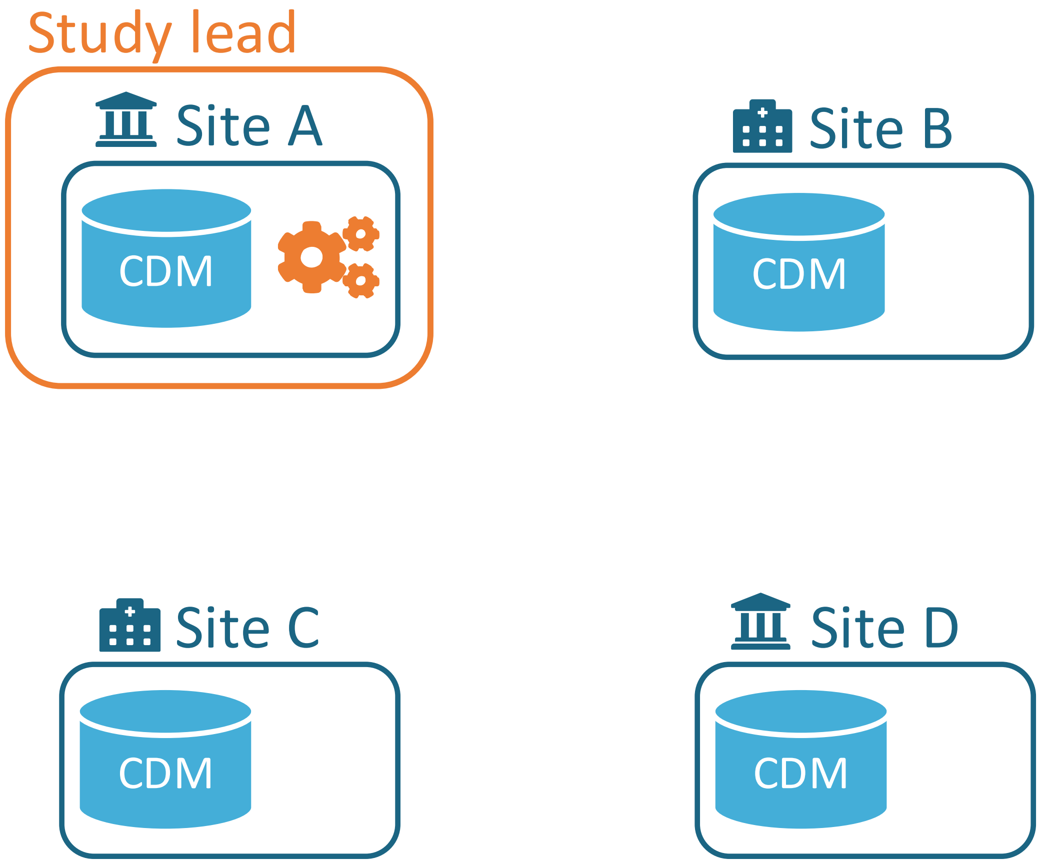
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally



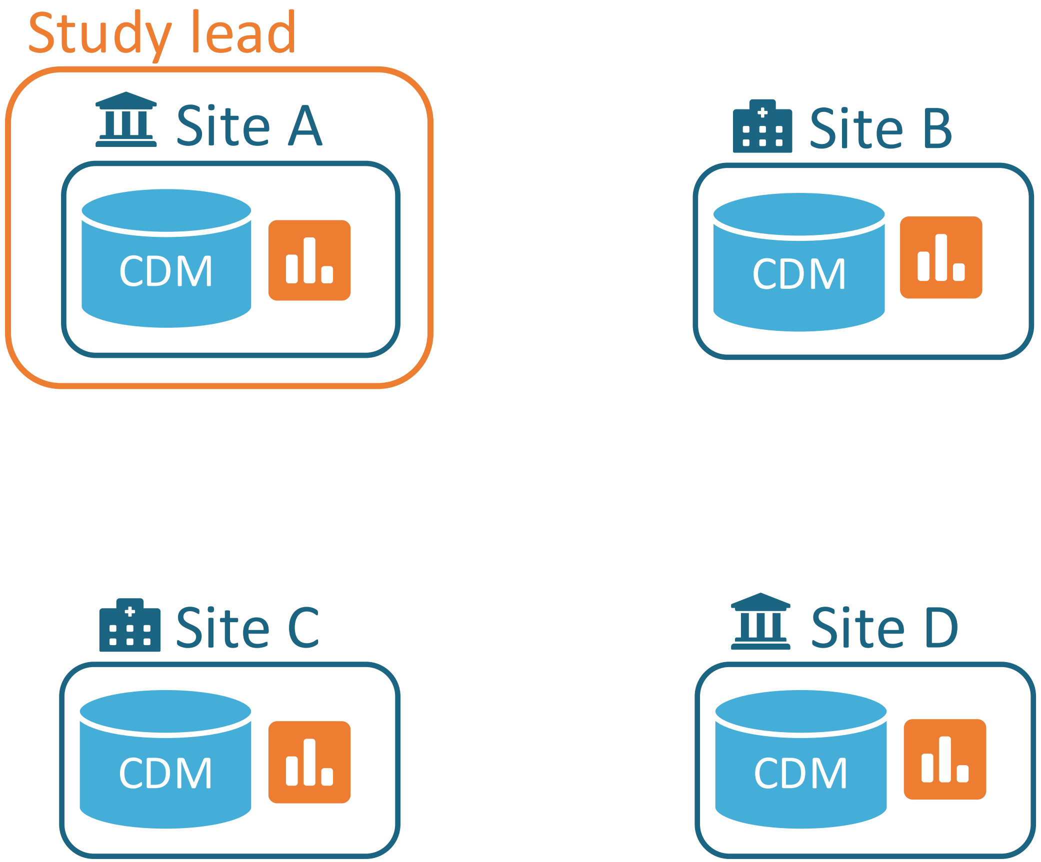
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants



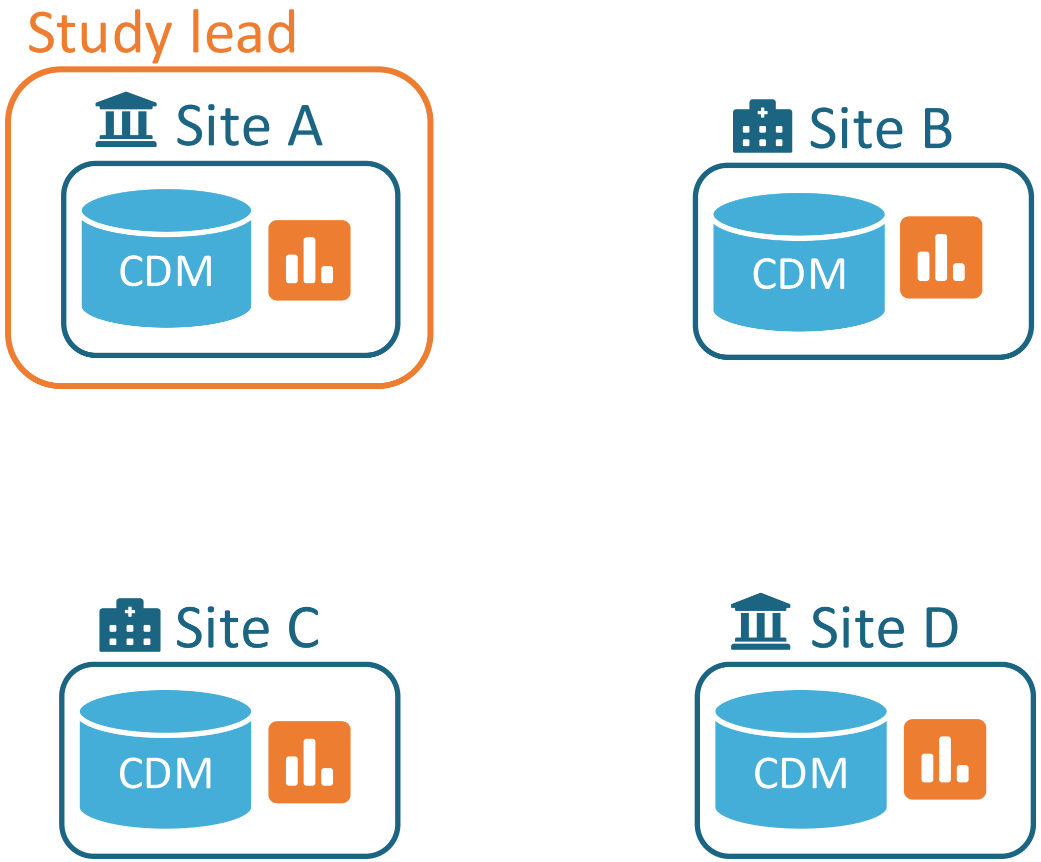
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)



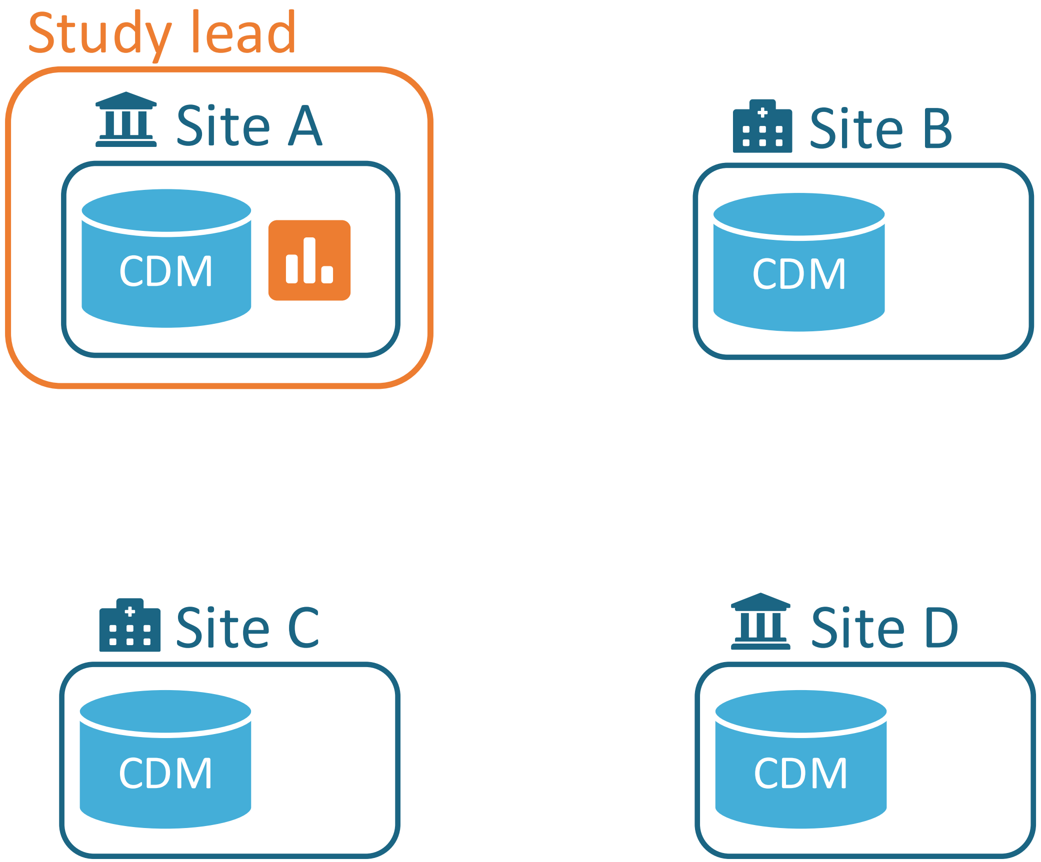
Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)
- Results are sent back to lead site

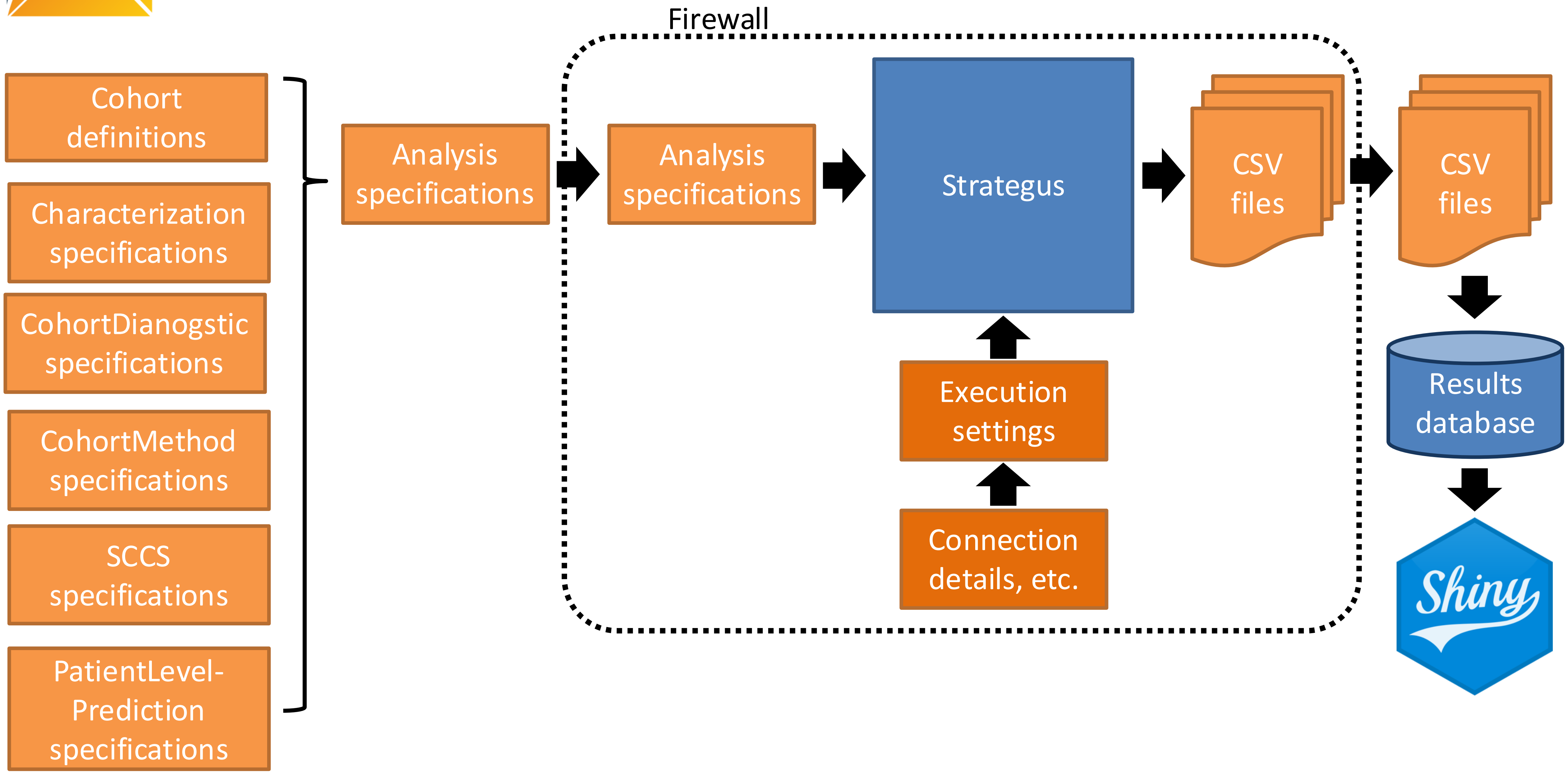


Distributed Research Network

- A site can lead a study
- Analysis code is developed locally
- Code is distributed to study participants
- Results are generated (aggregated statistics)
- Results are sent back to lead site
- Evidence is synthesized



Strategus for study execution



Summary

Unique features of HADES analytics

- Re-use of cohort definitions
- Standardization of analytics in open-source software
  - Many opportunities for testing, review, fixing bugs, etc.
  - Making it hard to do the wrong thing (opinionated)
- Advanced methods to reduce bias
  - Large-scale propensity scores in cohort method
- Objective study diagnostics to improve reliability of evidence
  - Including negative controls
- Designed to run across a network of databases
  - Without sharing patient-level data