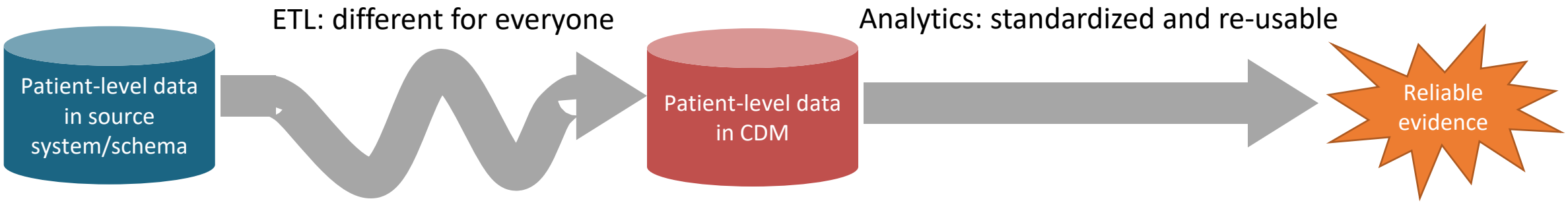
Advanced analytics with OHDSI tools

Why convert to the Common Data Model?

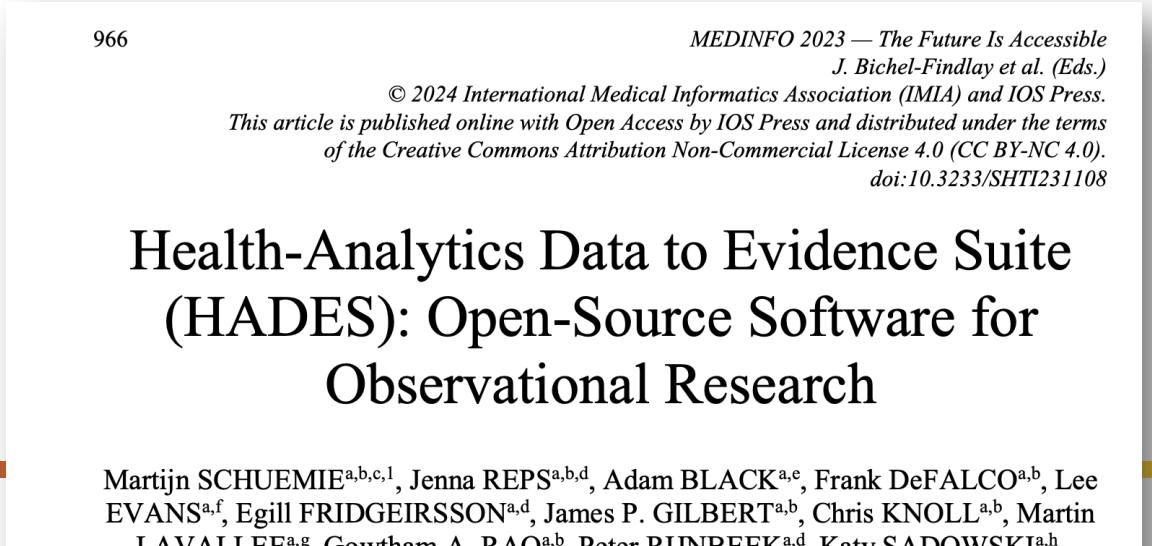
- Transforming data to the OMOP CDM is a large investment
- The benefits come from being able to use the same tools and analytics across many databases



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



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

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert your favorite drug>, how many patients experienced <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?
Population-level effect estimation	**Safety surveillance**	Does exposure to <insert your favorite drug> increase the risk of experiencing <insert an adverse event> within <time horizon following exposure start>?
	Comparative effectiveness	Does exposure to <insert your favorite drug> have a different risk of experiencing <insert any outcome (safety or benefit) > within <time horizon following exposure start>, relative to <insert your comparator treatment>?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert your favorite drug>, how many patients experienced <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?
Popula effect e		dverse any insert your comparator treatment>?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Standardizing the question itself helps clarify what your objective is

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert your favorite drug>, how many patients experienced <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?
Population-level effect estimation	**Safety surveillance**	Does exposure to <insert your favorite drug> increase the risk of experiencing <insert an adverse event> within <time horizon following exposure start>?
	Comparative effectiveness	Does exposure to <insert your favorite drug> have a different risk of experiencing <insert any outcome (safety or benefit) > within <time horizon following exposure start>, relative to <insert your comparator treatment>?
Patient level prediction	**Disease onset and progression**	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Analytic use case	Type	Structure
Clinical characterization	**Disease Natural History**	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert your favorite drug>, how many patients experienced <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?
Popula… effect e…		…dverse
		…any …insert
Patient level prediction	**progression**	…ability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Amongst patients who are new users of **GLP-1s**, how many patients experienced **Acute Myocardial Infarction** within **drug exposure**?

Analytic use case	Type	Structure
Clinical characterization	Disease Natural History	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease> and in which sequence?
	Outcome incidence	Amongst patients who are new users of <insert your favorite drug>, how many patients experienced <insert your favorite known adverse event from the drug profile> within <time horizon following exposure start>?
Population-level effect estimation	Safety surveillance	Does exposure to <insert your favorite drug> increase the risk of experiencing <insert an adverse event> within <time horizon following exposure start>?
	Comparative	Does exposure to <insert your favorite drug> have a different risk of experiencing <insert any
Patient level prediction	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

For a given patient who is a new user of **GLP-1s,** what is the probability that they will experience **an AMI** while **exposed to the drug**?

Analytic use case	Type	Structure
Clinical chara	Disease Natural History	Amongst patients who are diagnosed with <insert your favorite disease>, what are the patient's characteristics from their medical history?
	Treatment utilization	Amongst patients who have <insert your favorite disease>, which treatments were patients exposed to amongst <list of treatments for disease>, and in which sequence?
		time
Popula effect estimation		n adverse
	Comparative effectiveness	Does exposure to <insert your favorite drug> have a different risk of experiencing <insert any outcome (safety or benefit) > within <time horizon following exposure start>, relative to <insert your comparator treatment>?
Patient level prediction	Disease onset and progression	For a given patient who is diagnosed with <insert your favorite disease>, what is the probability that they will go on to have <another disease or related complication> within <time horizon from diagnosis>?
	Treatment response	For a given patient who is a new user of <insert your favorite chronically-used drug>, what is the probability that they will <insert desired effect> in <time window>?
	Treatment safety	For a given patient who is a new user of <insert your favorite drug>, what is the probability that they will experience <insert adverse event> within <time horizon following exposure>?

Does exposure to **GLP-1s** have a different risk of experiencing **AMI** while **exposed to drug**, relative to **DPP-4s**?

Cohorts of our examples

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

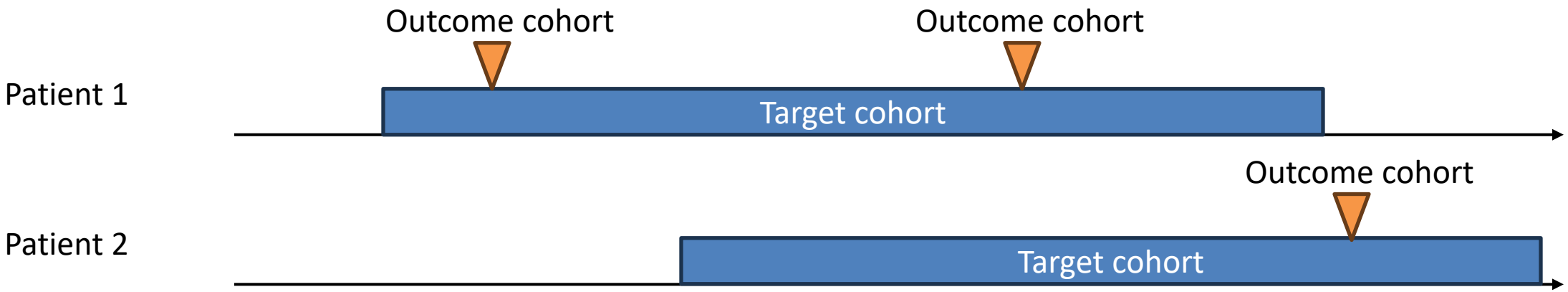
These same cohorts can be re-used to answer different questions

Patrick will discuss how to build these

Characterization: CohortIncidence package

Amongst patients who are new users of **GLP-1s,** how many patients experienced **Acute Myocardial Infarction** within **drug exposure**?

- Target: **GLP-1**
- Outcome: **AMI**



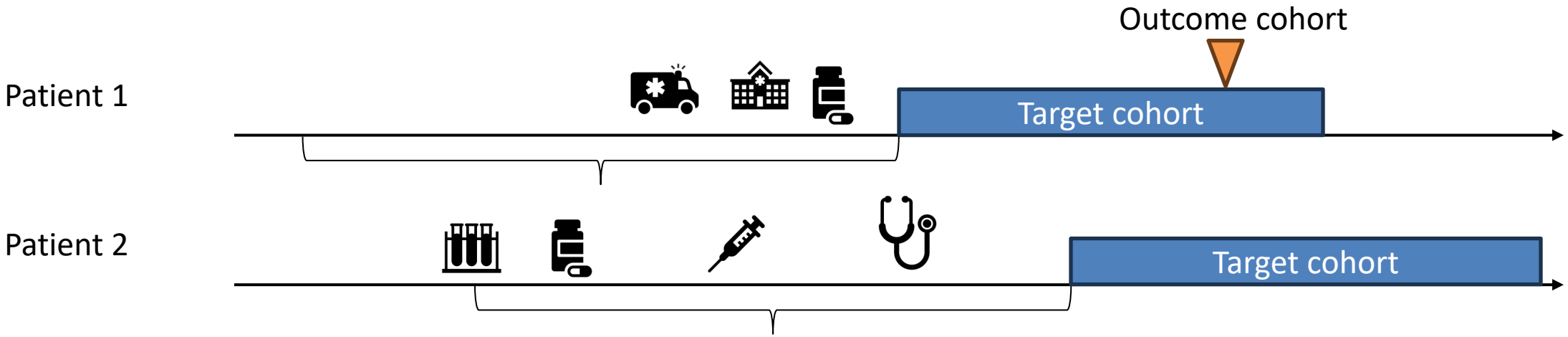
Computes the incidence rate of the Outcome cohort in some Target cohort

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

PatientLevelPrediction package

For a given patient who is a new user of **GLP-1s,** what is the probability that they will experience **an AMI** while **exposed to the drug**?

- Target: **GLP-1**, restricted to those with **T2DM** (and first use only)
- Outcome: **AMI**



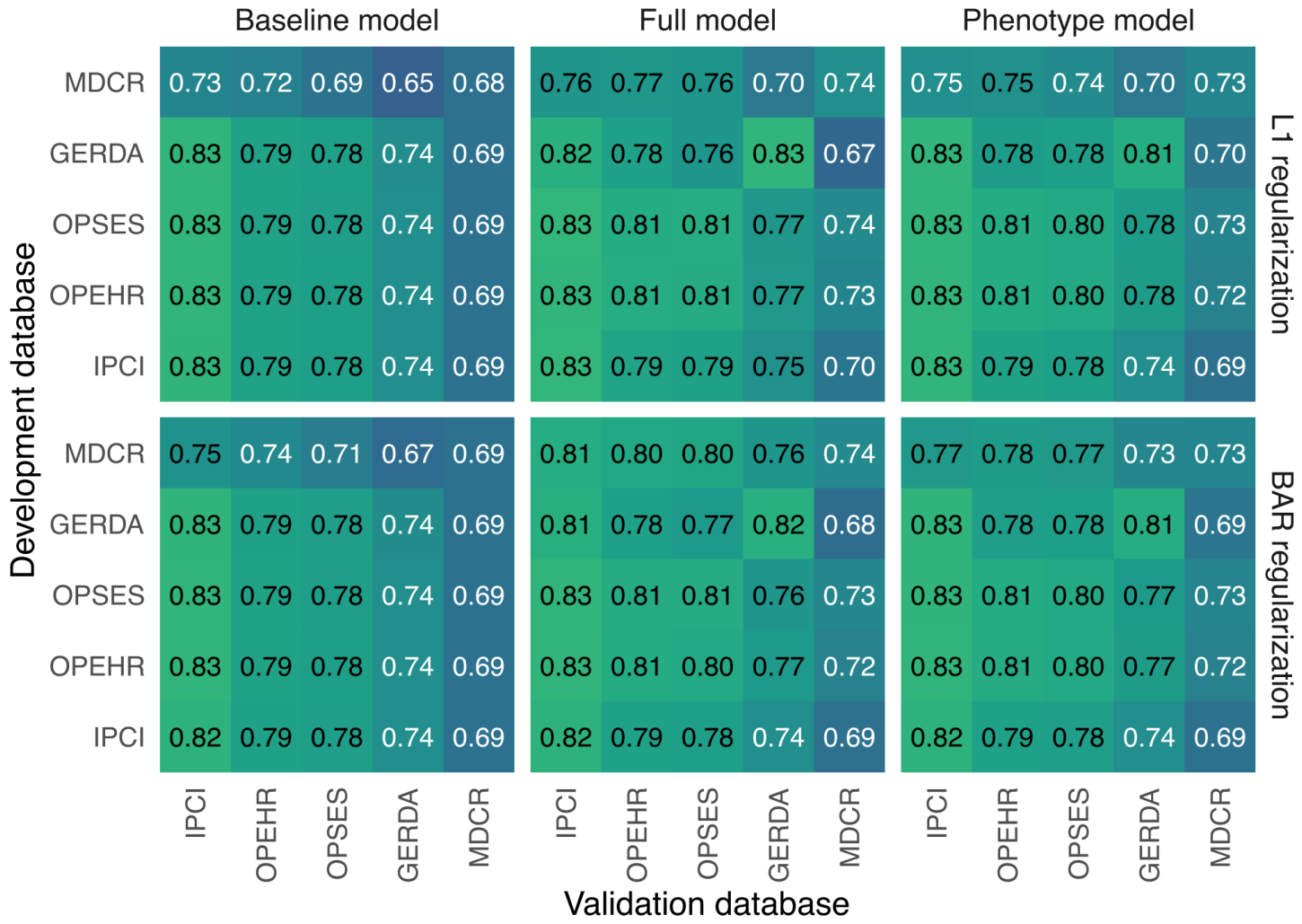
Builds a model to predict who in the Target will have the Outcome

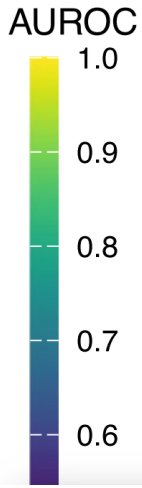
- Uses all observed data up to Target start
- Implements many machine learning / deep learning algorithms

Unique feature: external validation

- When using the PatientLevelPrediction package, models fit in one database can easily be validated in other databases
  - Portability of code
  - Standardized construction of features

Example of external validation





John *et al. BMC Medicine* (2024) 22:308
https://doi.org/10.1186/s12916-024-03530-9

BMC Medicine

RESEARCH ARTICLE Open Access

Development and validation of a patient-level model to predict dementia across a network of observational databases

Luis H. John[1*], Egill A. Fridgeirsson[1], Jan A. Kors[1], Jenna M. Reps[2], Ross D. Williams[1], Patrick B. Ryan[2] and Peter R. Rijnbeek[1]

Abstract

Background A prediction model can be a useful tool to quantify the risk of a patient developing dementia

CohortMethod package

Does exposure to **GLP-1s** have a different risk of experiencing **AMI** while **exposed to drug**, relative to **DPP-4s**?

- 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**

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

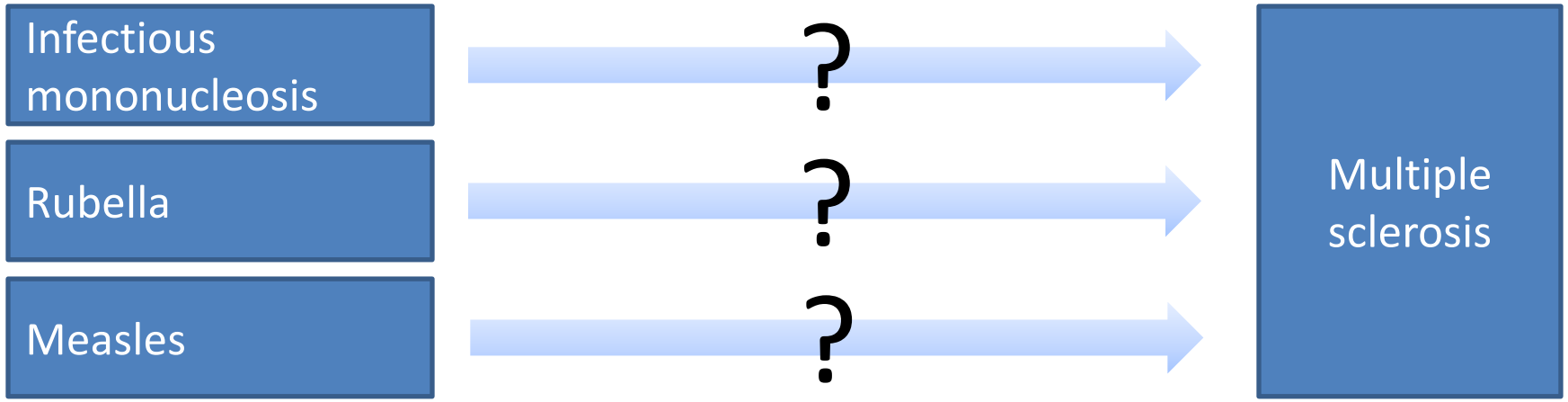
Unique feature: Large-scale propensity scores

- 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 machine learning

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



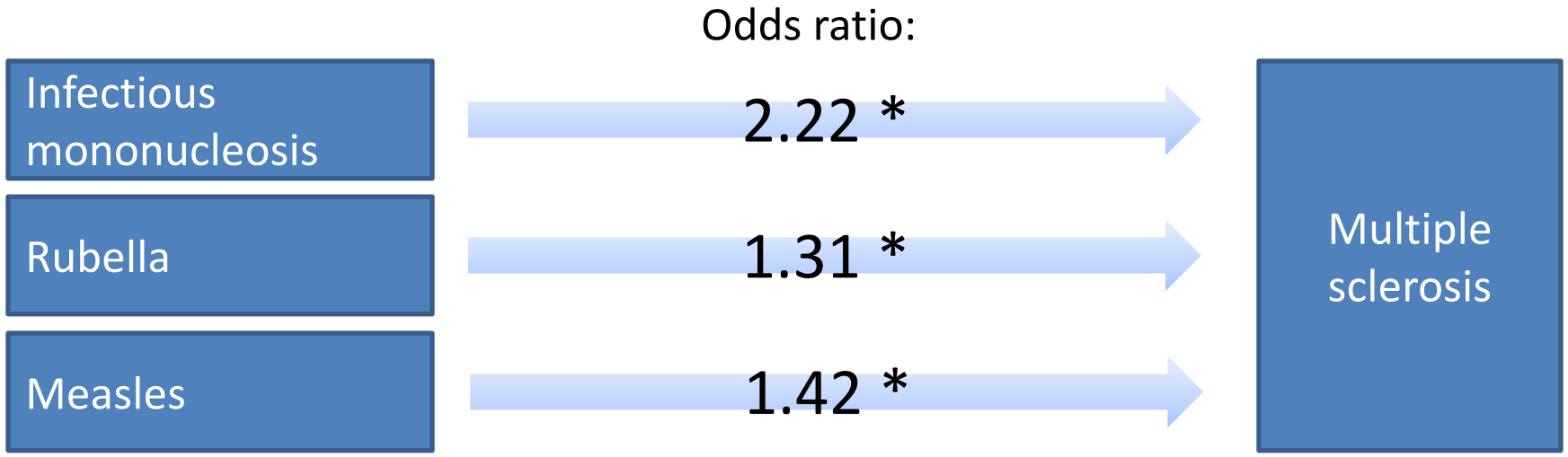
RESEARCH PAPER

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



* P < .05

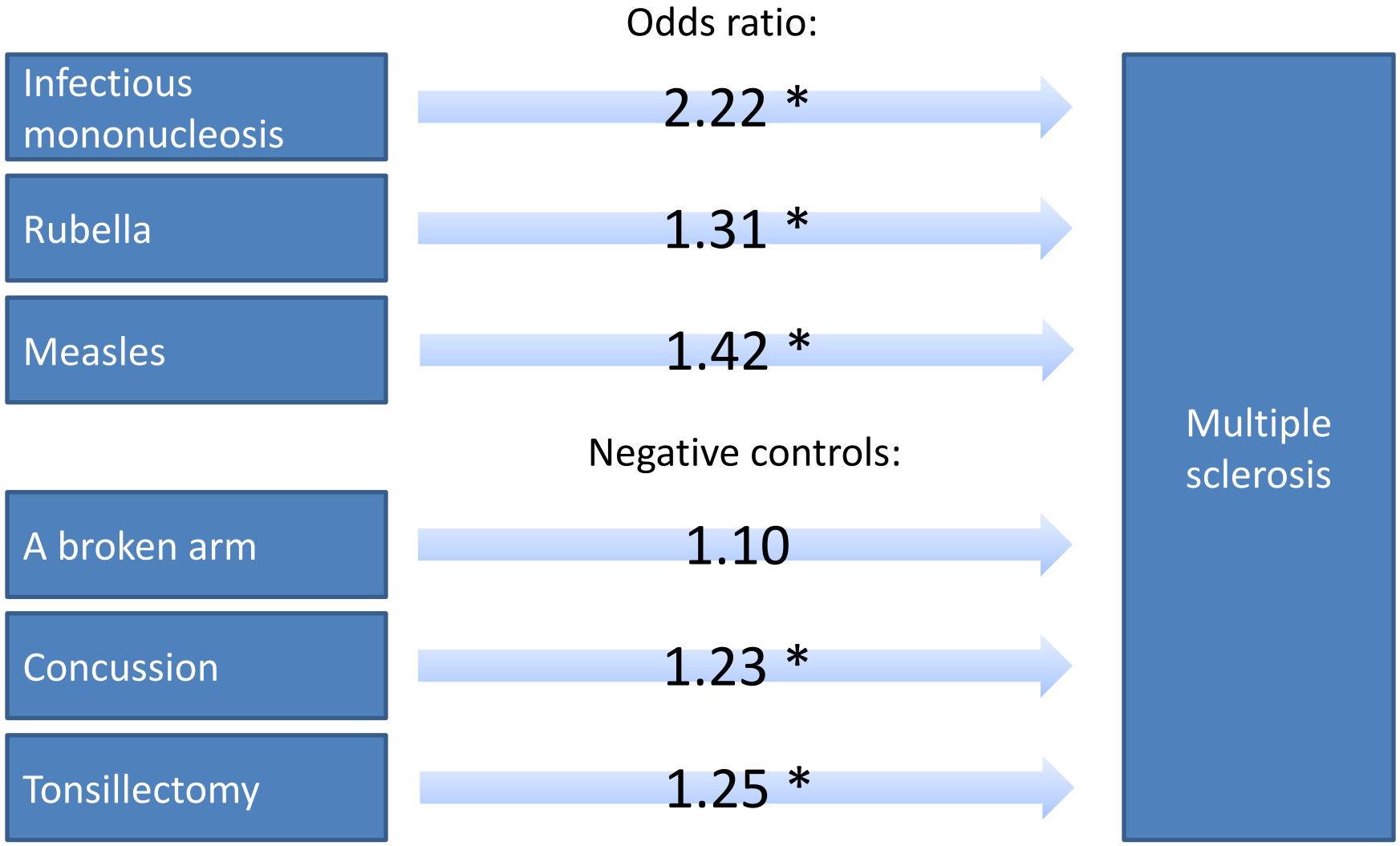
RESEARCH PAPER

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

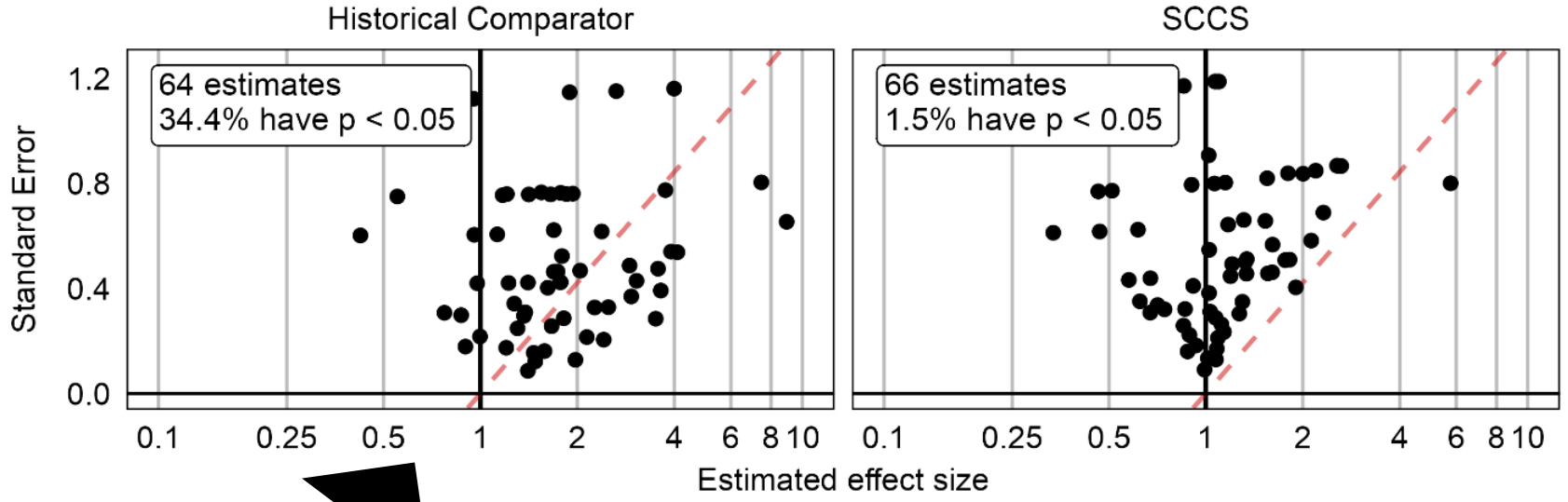


* P < .05

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



Need to execute estimation studies for 66 Target-Outcome combinations.

OHDSI tools readily allow for this (simply swapping out the outcome cohort for the negative controls)

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,

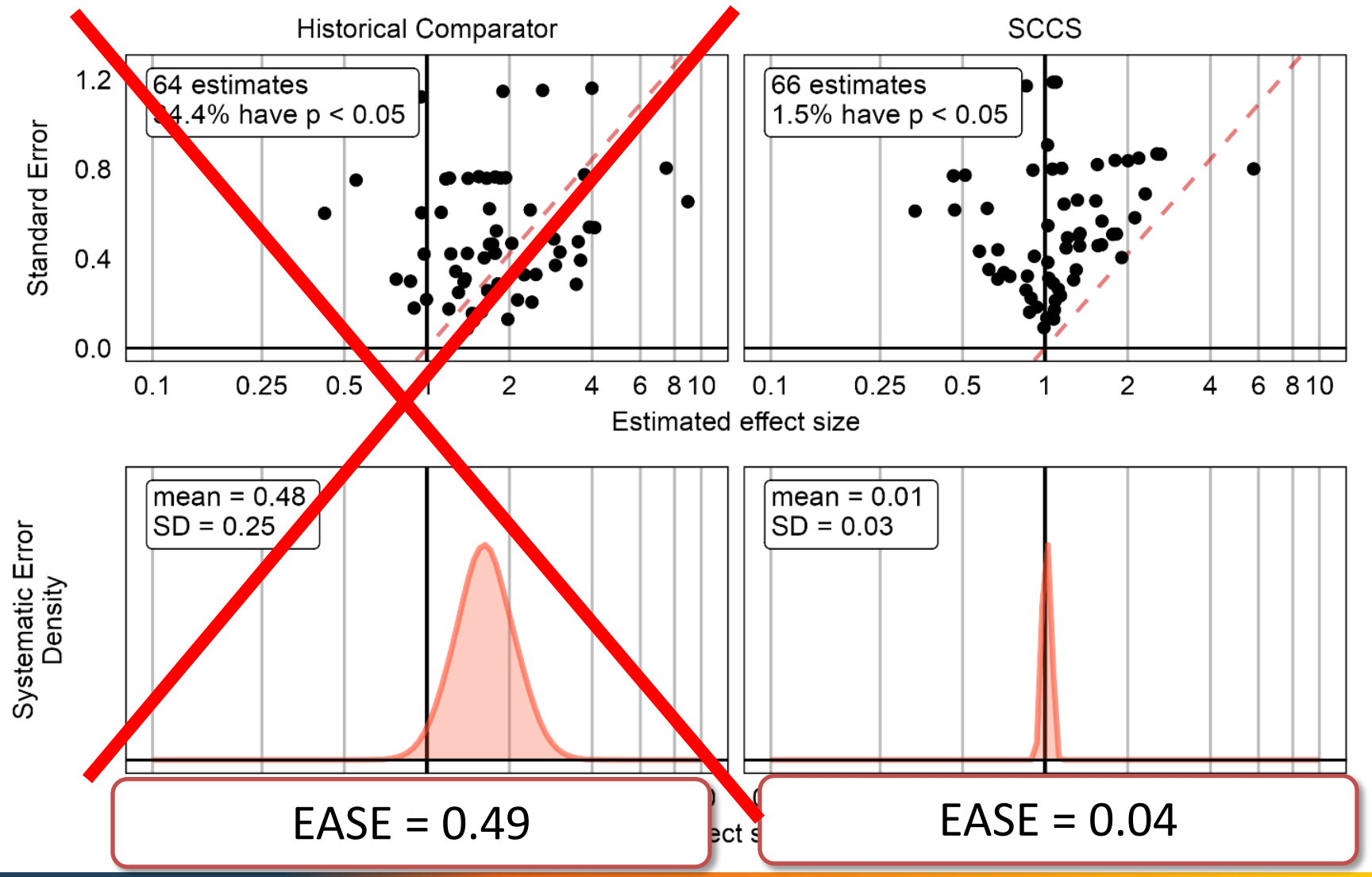
[3]Department of Human Genetics, University of California, Los Angeles,

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

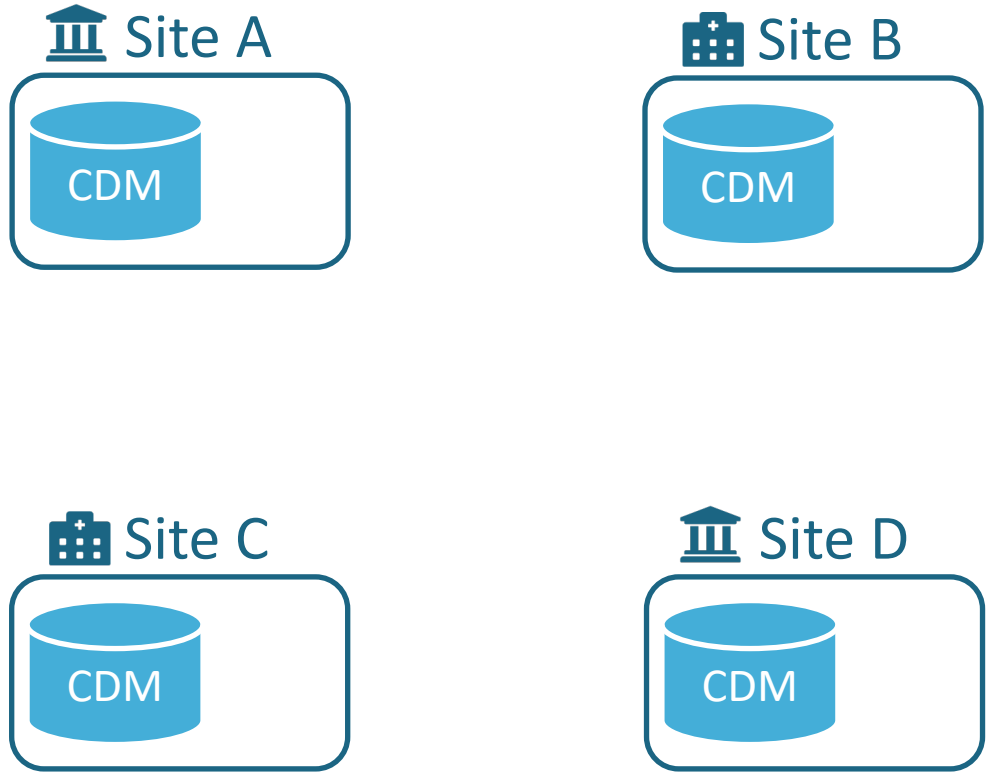


Distributed analyses

Using OHDSI tools

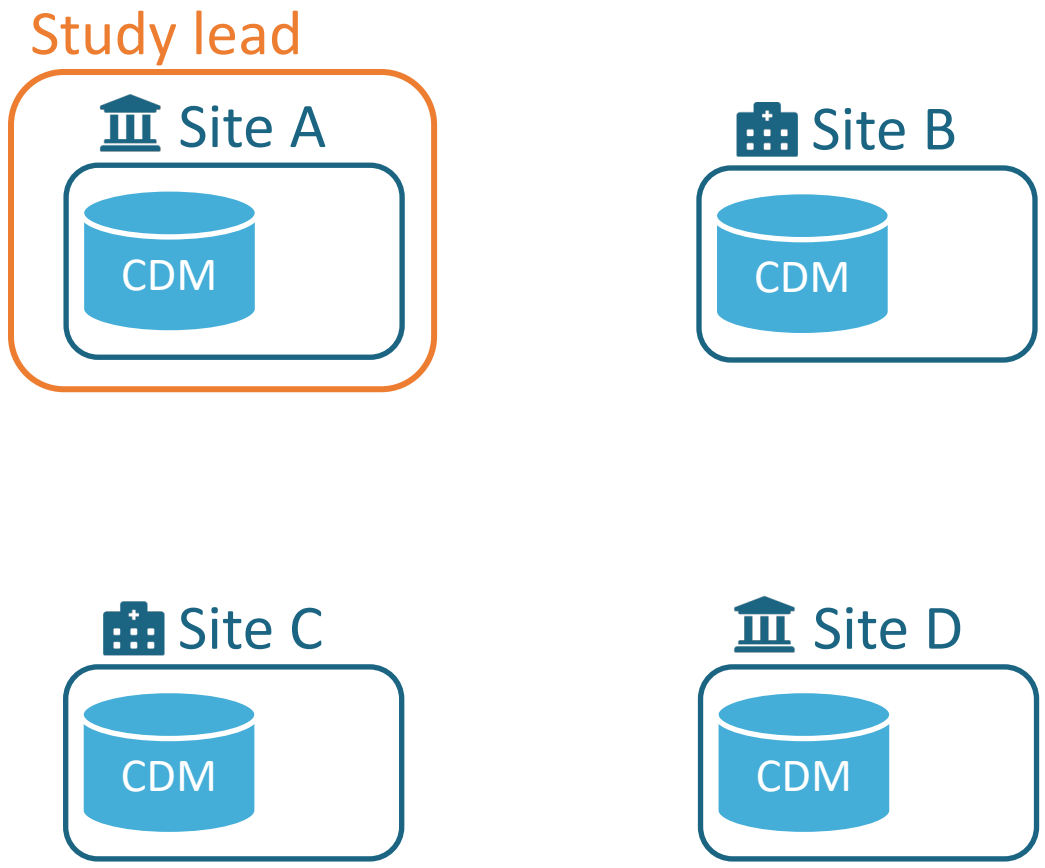
Distributed Research Network

- 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)



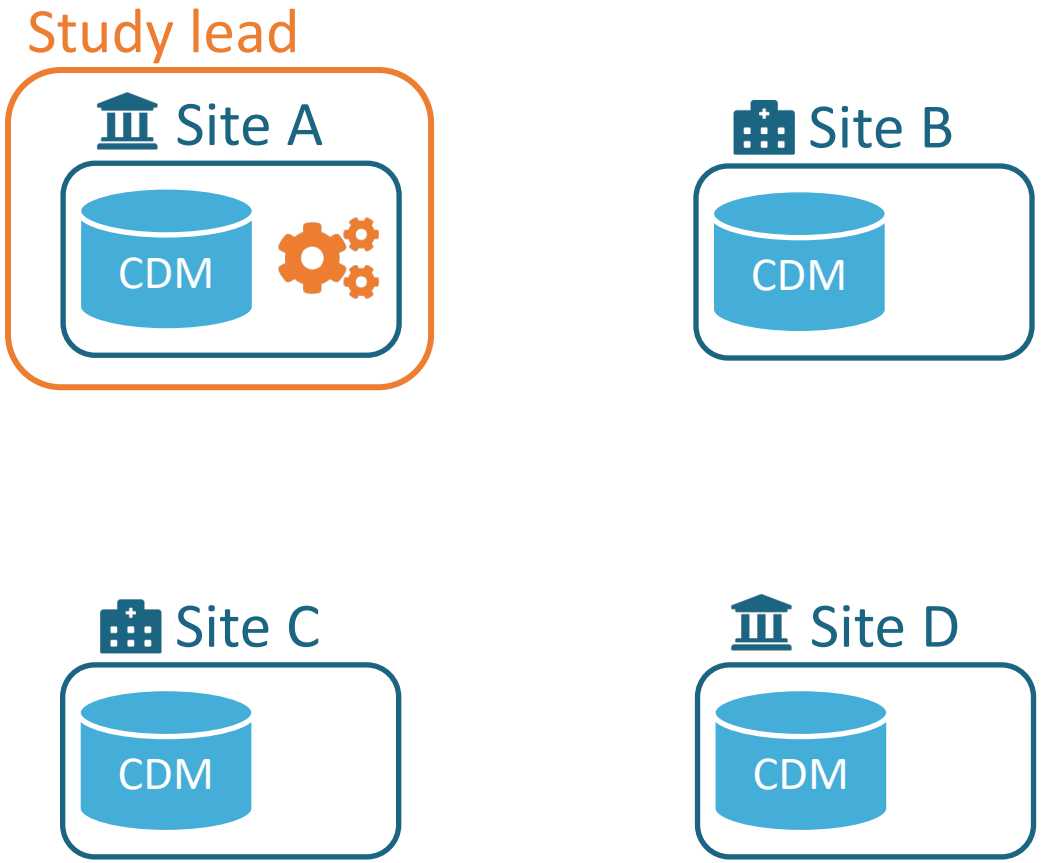
Distributed Research Network

- A site can lead a study



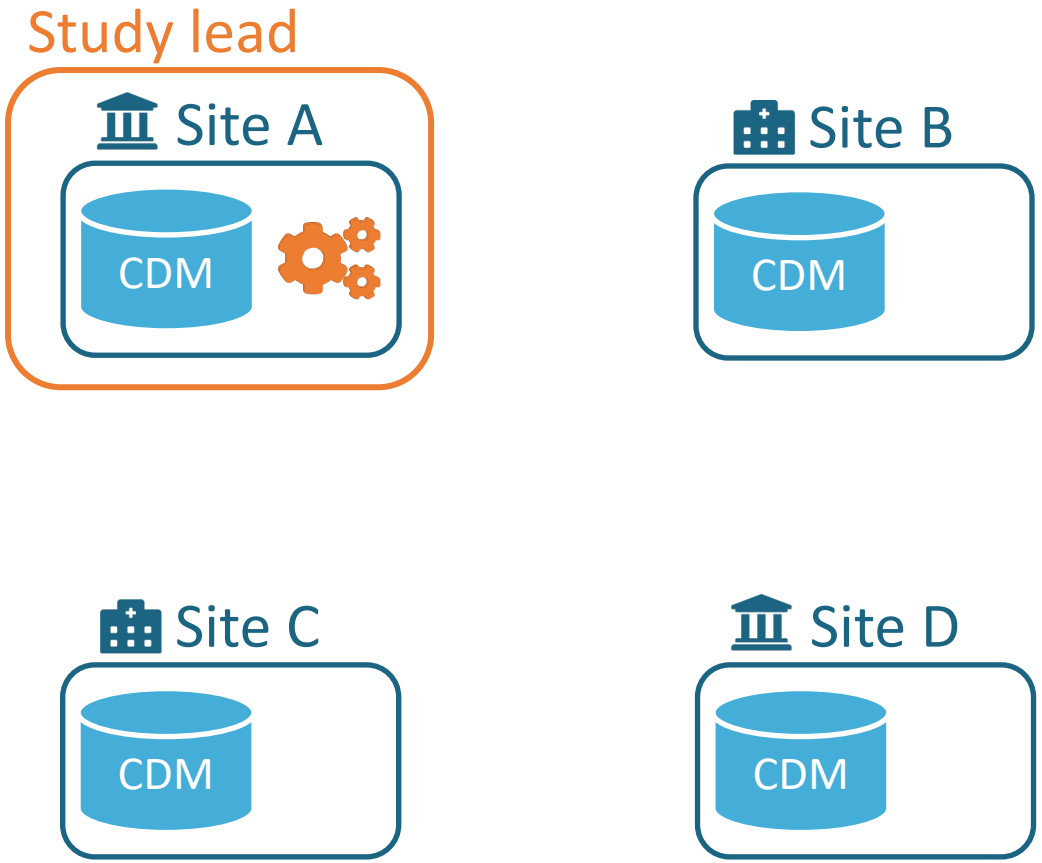
Distributed Research Network

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



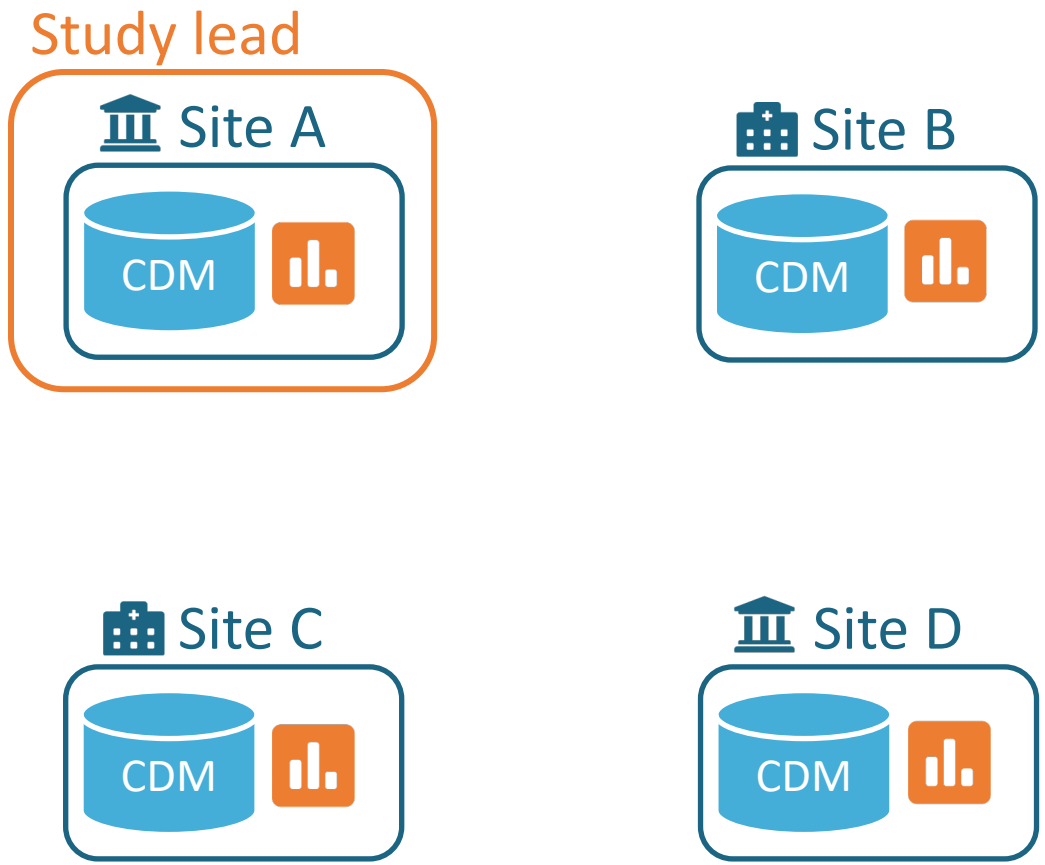
Distributed Research Network

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



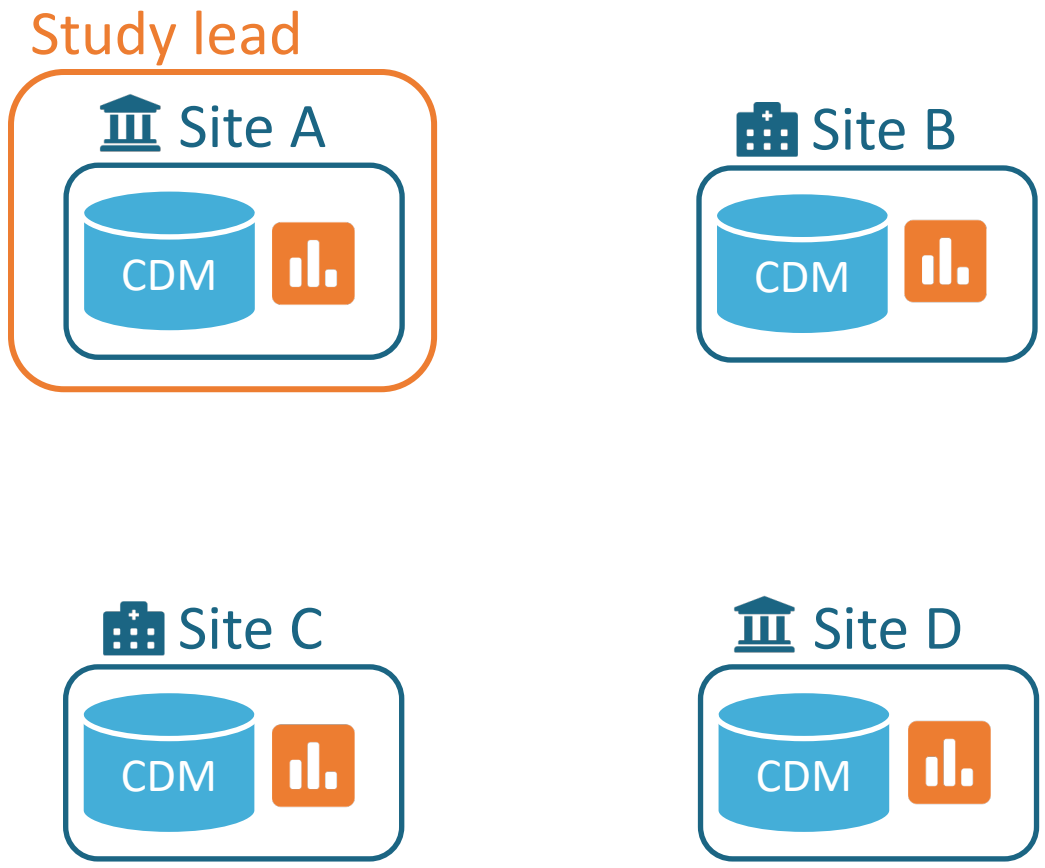
Distributed Research Network

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



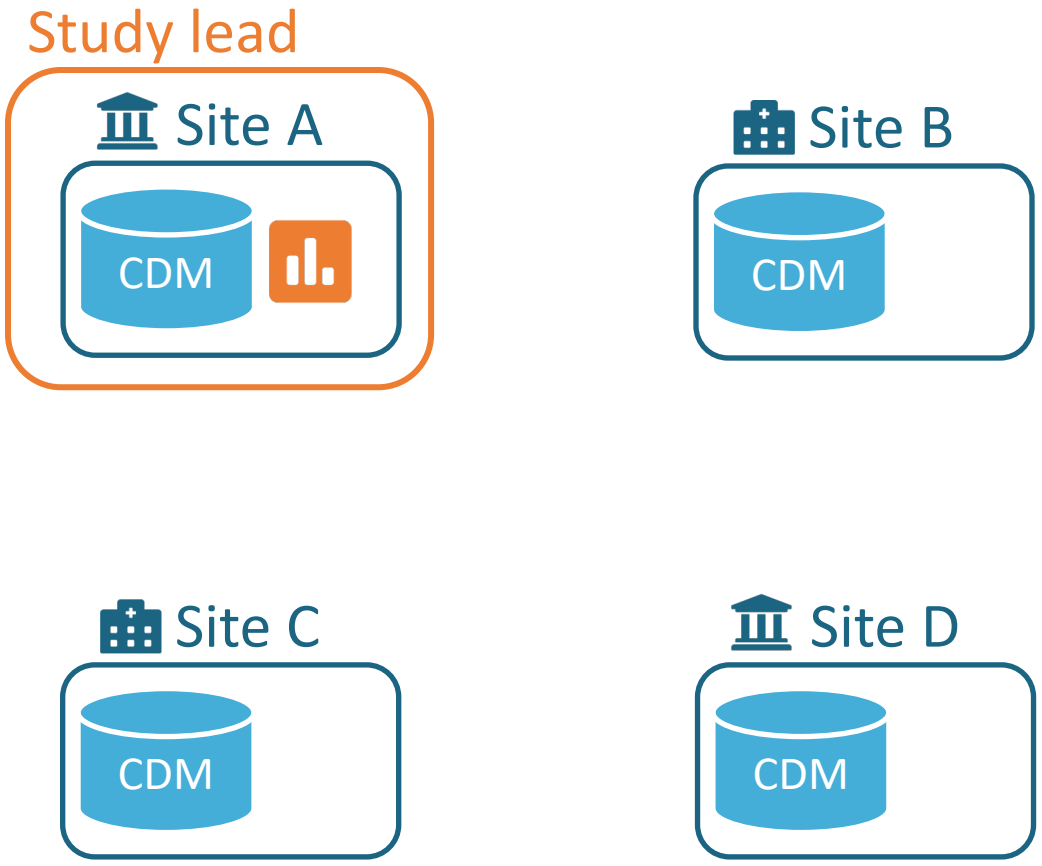
Distributed Research Network

- 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

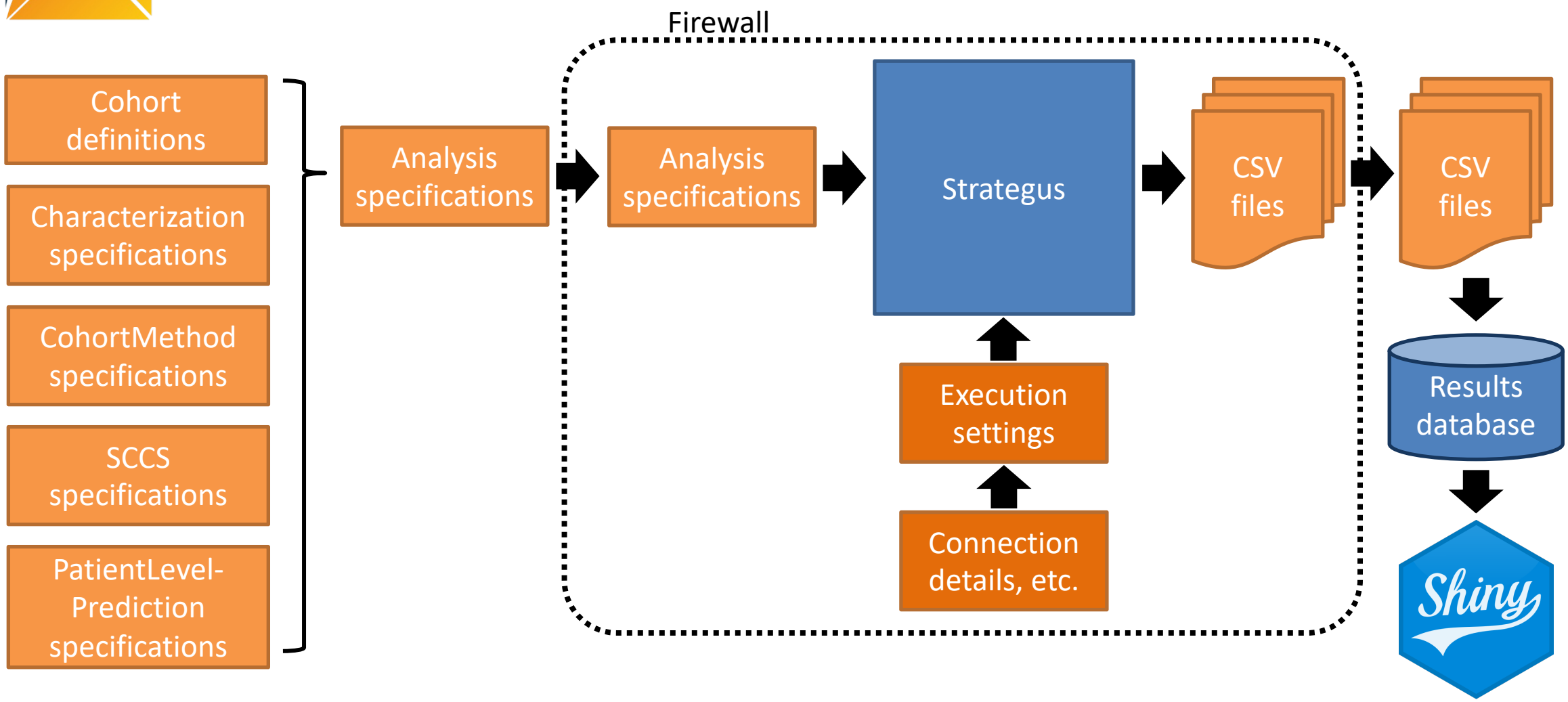


Distributed Research Network

- 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




OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

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
  - Splines for time in self-controlled case series
  - 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