



# An Open Science Approach to Medical Evidence Generation: Introducing Observational Health Data Sciences and Informatics

Jon Duke, MD MS  
Regenstrief Institute  
Academy Heath  
June 14 2015



# What is OHDSI?

- The Observational Health Data Sciences and Informatics (OHDSI) program is a multi-stakeholder, interdisciplinary collaborative
- The goal of OHDSI is to bring out the value of observational health data through large-scale analytics and evidence generation
- All our software and other products are released as open-source



# OHDSI: a global community



## OHDSI Collaborators:

- >140 researchers in academia, industry and government
- >10 countries

## OHDSI Data Network:

- >50 databases standardized to OMOP common data model
- >680 million patients



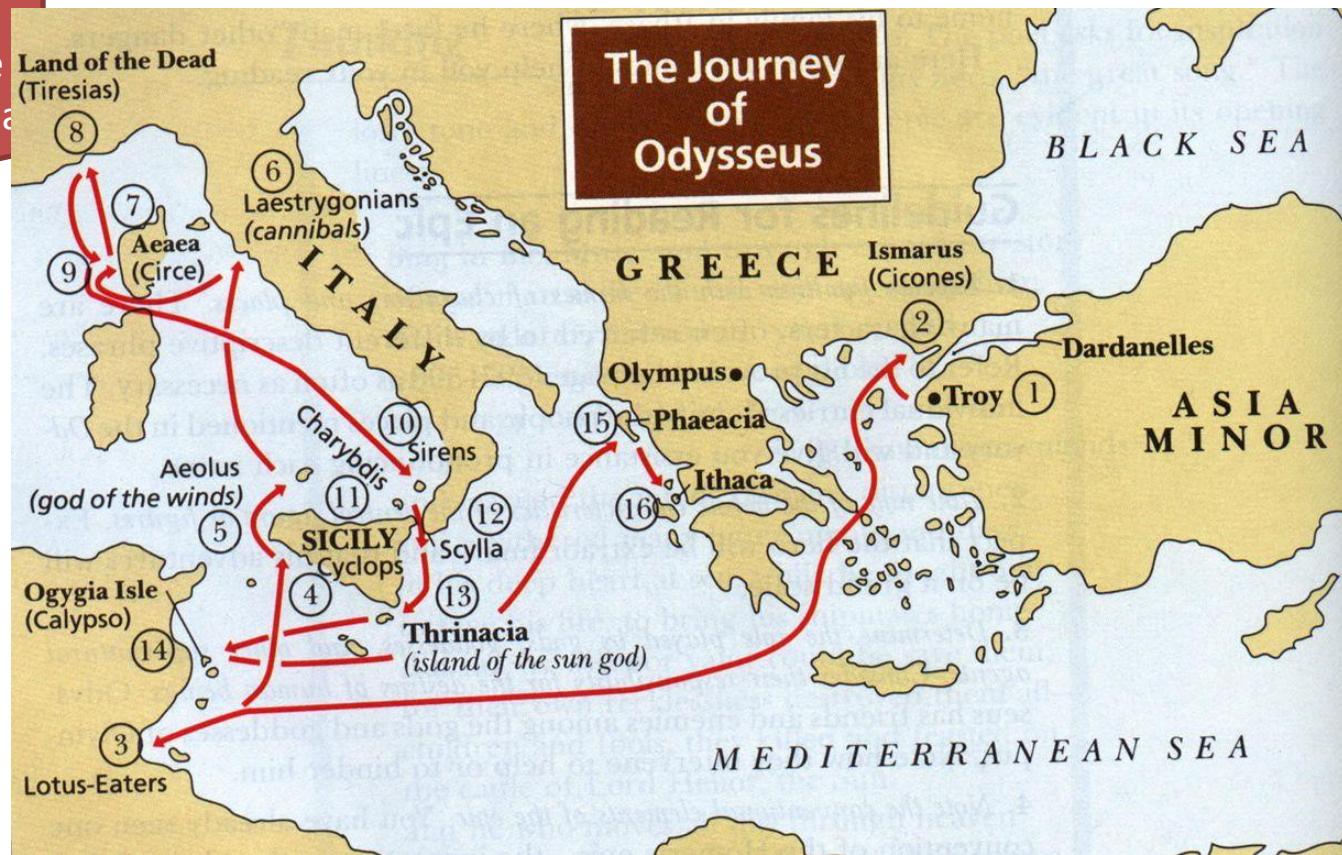
# OHDSI Evidence Generation

- Clinical characterization:
  - Descriptive statistics (e.g., natural history of a disease or patterns of medication use)
  - Quality improvement (e.g., performance measures)
- Population-level estimation
  - Safety surveillance (e.g., identifying new adverse event risks for drugs)
  - Comparative effectiveness (e.g. comparing interventional to non-interventional treatment of chronic back pain)
- Patient-level prediction
  - Incorporating patient medical history to provide personalized recommendations for therapy selection, adverse event risk, high value diagnostic studies



# The odyssey to evidence generation

Patient-level  
data in source  
system/ schema



evidence



# Open Science through Standardization

- The OHDSI community has standardized core components of the research process in order to
  - Promote transparent, reproducible science
  - Reveal data quality issues
  - ‘Calibrate’ datasets
  - Bring skillsets together from across the community (clinical, epi, stats, compSci)



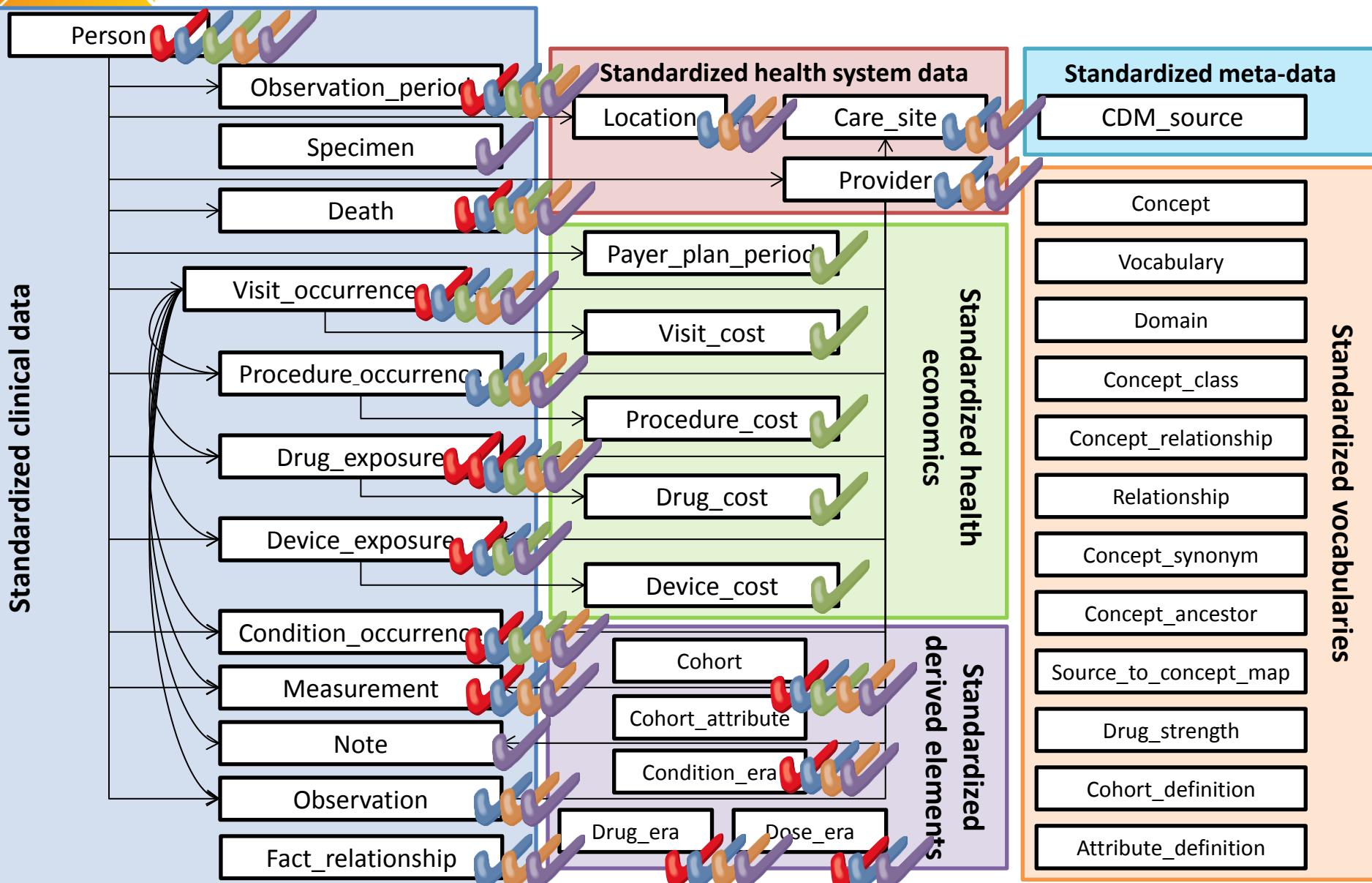
# Opportunities for standardization in the evidence generation process

Protocol

- **Data structure** : tables, fields, data types
- **Data content** : vocabulary to codify clinical domains
- **Data semantics** : conventions about meaning
- **Cohort definition** : algorithms for identifying the set of patients who meet a collection of criteria
- **Covariate construction** : logic to define variables available for use in statistical analysis
- **Analysis** : collection of decisions and procedures required to produce aggregate summary statistics from patient-level data
- **Results reporting** : series of aggregate summary statistics presented in tabular and graphical form

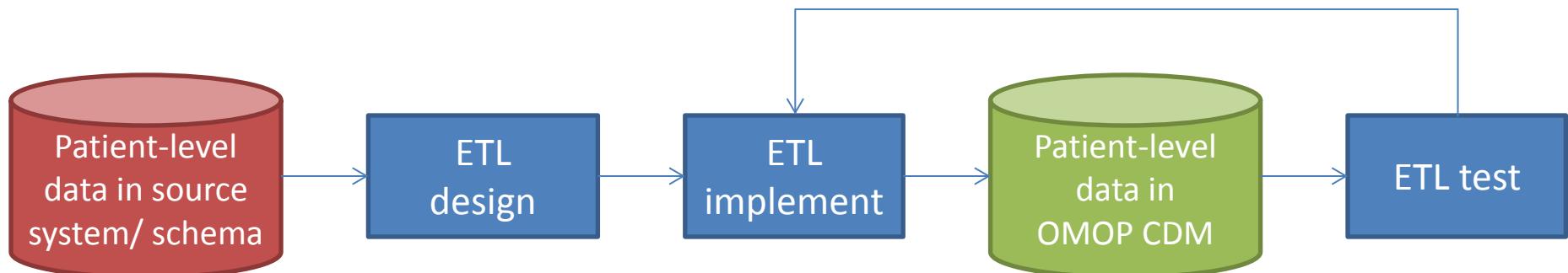


# One model, multiple use cases





# Preparing your data for analysis



OHDSI tools built to help

**WhiteRabbit:**  
profile your  
source data

**RabbitInAHat:**  
map your source  
structure to  
CDM tables and  
fields

**ATHENA:**  
standardized  
vocabularies  
for all CDM  
domains

**Usagi:**  
map your  
source codes  
to CDM  
vocabulary

**CDM:**  
DDL, index,  
constraints for  
Oracle, SQL

Server,  
PostgreSQL;  
Vocabulary tables  
with loading  
scripts

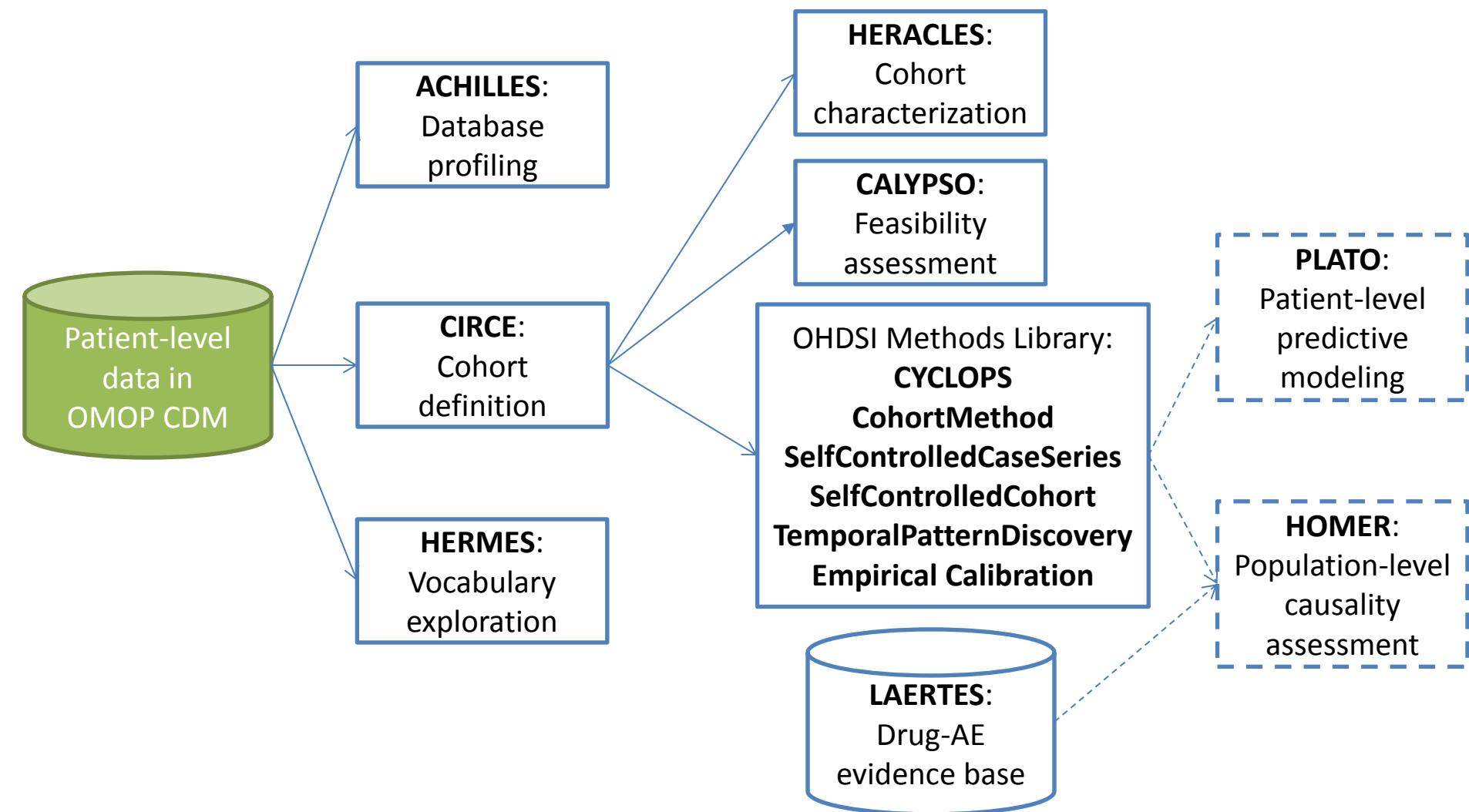
**ACHILLES:**  
profile your  
CDM data;  
review data  
quality  
assessment;  
explore  
population-  
level summaries

**OHDSI Forums:**

Public discussions for OMOP CDM Implementers/developers



# Standardized large-scale analytics tools under development within OHDSI





# OHDSI Software

- Community developed
- Apache 2.0 licensed
- Available on GitHub
- Common frameworks
  - Java
  - HTML5 / Javascript
  - R
  - Oracle / SQL Server / Postgres / Redshift / Netezza



# Motivating example to see the OHDSI tools in action

---



## MINI-SENTINEL MEDICAL PRODUCT ASSESSMENT A PROTOCOL FOR ASSESSMENT OF DABIGATRAN

Version 3

March 27, 2015

Prior versions:

Version 1: December 31, 2013

Version 2: March 18, 2014

**Prepared by:** Alan S. Go, MD<sup>1</sup>, Daniel Singer, MD<sup>2</sup>, T. Craig Cheetham, PharmD MS<sup>3</sup>, Darren Toh, ScD<sup>4</sup>, Marsha Reichman, PhD<sup>5</sup>, David Graham, MD MPH<sup>5</sup>, Mary Ross Southworth, PharmD<sup>6</sup>, Rongmei Zhang PhD<sup>7</sup>, Monika Houstoun, PharmD<sup>5</sup>, Yu-te Wu PhD<sup>7</sup>, Katrina Mott MS<sup>5</sup>, Joshua Gagne, PharmD ScD<sup>8</sup>

**Author Affiliations:** 1. Division of Research, Kaiser Permanente Northern California, Oakland, CA. 2. General Medicine Division, Massachusetts General Hospital, Boston, MA. 3. Kaiser Permanente Southern California, Downey, CA. 4. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA. 5. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD. 6. Division of Cardiovascular and Renal Products, Office of New Drugs, Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD. 7. Division of Biometric VII, Office of Biostatistics, Office of Translation Sciences, Food and Drug Administration (FDA), Silver Spring, MD. 8. Division of



### III. PROTOCOL DETAILS

#### A. ASSESSMENT DESIGN

This one-time assessment will employ a “new user” parallel cohort design.<sup>12</sup>

#### B. COHORT IDENTIFICATION

##### 1. Target Population

We will focus on the identification of **adult (age ≥21 years) patients with diagnosed nonvalvular atrial fibrillation and who are new users of dabigatran or warfarin.**

##### 2. Sample Inclusion and Exclusion Criteria

The target sample inclusion and exclusion criteria are summarized in **Table 1** below. Please see **Appendix A** and **Section D** for additional details, definitions and rationale.

**Table 1. Inclusion and exclusion criteria for comparison of adults with atrial fibrillation who are new users of dabigatran or warfarin in the MSDD.**

Inclusion Criteria	Exclusion Criteria
<ul style="list-style-type: none"><li>First dispensing of dabigatran or warfarin therapy from November 1, 2010 to the most recent data available in the MSDD from participating Data Partners *</li><li>Age 21 years or older at the first dispensing of dabigatran or warfarin therapy</li><li>One or more diagnoses of atrial fibrillation or atrial flutter based on ICD-9-CM codes (ICD-9-CM 427.31, 427.32) from any practice setting (inpatient or outpatient) any time before the first identified prescription for dabigatran or warfarin therapy during the study period *</li></ul>	<ul style="list-style-type: none"><li>Less than 180 days of continuous enrollment with prescription and medical coverage immediately preceding the date of the index dispensing (i.e., index date)</li><li>Any prior dispensing for warfarin, dabigatran, rivaroxaban or apixaban during the 180 days before index date **</li><li>Known mechanical heart valve or diagnosed mitral stenosis at index date based on corresponding administrative diagnosis and/or procedure codes</li><li>Chronic hemodialysis or peritoneal dialysis at index date based on corresponding administrative diagnosis and/or procedure codes</li><li>History of kidney transplant at index date based on corresponding administrative diagnosis and/or procedure codes</li><li>At a skilled nursing facility or nursing home at index date</li></ul>



# Let's ask the OHDSI network!

Observational Health Data Sciences and Informatics

Recent changes Media Manager Sitemap

Trace: [welcome](#) • [data\\_network](#)

resources: data\_network

The following table provides a list of databases which have been converted to the OMOP CDM

Database	Data Type	Country	# of Patients (000s)	CDM Status
Truven MarketScan Commercial Claims and Encounters (CCAE)	Claims	USA	113060	CDMv4 complete, ETL posted
Truven MarketScan Multi-state Medicaid (MDCD)	Claims	USA	16150	CDMv4 complete, ETL posted
Truven MarketScan Medicare Supplemental (MDCR)	Claims	USA	8710	CDMv4 complete, ETL posted
Optum ClinFormatics	Claims	USA	36230	CDMv4 complete, ETL posted
Premier	Hospital Billing	USA	100090	CDMv4 complete, ETL posted
HCUP NIS	Claims	USA	91980	CDMv4 complete, ETL posted
NHANES	Survey	USA	72	CDMv4 complete, ETL posted

Documentation Development Research Studies Projects & Workgroups Other Resources

- Community Forums
- Data Network
- Funding Opportunities
- Call for Papers
- Conferences
- Mailing Lists
- Realtime Chat (IRC)

+Add New Page



# OLYMPUS

## THE OHDSI APPLICATION LAUNCHER

There are remote WebAPIs configured. Applications that support toggling between WebAPIs will allow you to use these via the gear/settings.



ATHENA

OMOP Vocabulary  
Loader

HERMES

OMOP Vocabulary  
Explorer

ACHILLES

Dataset  
Characterization

CIRCE

Cohort Creation



HERACLES

Cohort  
Characterization

CALYPSO

Clinical Trial  
Feasibility



# Use ACHILLES to see if the databases have the required data elements

Achilles Data Sources Reports

OI Achilles Data Sources Reports

**OPTUM**

Drug Era Report

Warfarin

Drug Prevalence

Drug Exposure Prevalence by Month

BL A

Age At First Exposure

Length of Era Distribution

Days

Max: 3.252k  
P90: 477  
P75: 218  
Median: 90  
P25: 30  
P10: 30  
Min: 0

Length of Era

80

60

40

20

0

MALE

FEMALE

UNKNOWN

70

50

30

10

0

400

300

200

100

0



# Also use ACHILLES to check for any data quality issues

Data Quality Messages	
Message Type	Message
ERROR	101-Number of persons by age, with age at first observation period; should not have age < 0, (n=848)
ERROR	103 - Distribution of age at first observation period (count = 1); min value should not be negative
ERROR	114-Number of persons with observation period before year-of-birth; count (n=851) should not be > 0
ERROR	206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative
ERROR	301-Number of providers by specialty concept_id; 224 concepts in data are not in correct vocabulary (Specialty)
ERROR	400-Number of persons with at least one condition occurrence, by condition_concept_id; 115 concepts in data are not in correct vocabulary (SNOMED)
ERROR	406 - Distribution of age by condition_concept_id (count = 753); min value should not be negative



# Use HERMES to figure out how to find a particular condition, drug, procedure, or other concept

HERMES

warfarin

Drug RxNorm 11289 1310149 Ingredient V S

Warfarin

Concepts Related to Warfarin

Vocabulary

NDC (2328)	SPL (113)	RxNorm (93)	Multilex (71)	NDFRT (69)	VA Product (56)
Gemscript (28)	SNOMED (13)	Multum (10)	Genseqno (10)	ATC (5)	VA Class (2)
Cohort (1)	Mesh (1)				

Standard Concept

N (2636)	C (84)	S (80)
----------	--------	--------

Invalid Reason

V (2758)	D (31)	U (11)
----------	--------	--------

Class

11-digit NDC (2062)	9-digit NDC (266)	SPL (101)	Clinical Drug (80)	VA Product (56)	Ind / CI (37)
Gemscript (28)	Clinical Drug Comp (23)	Branded Drug Comp (21)	Branded Drug (21)	Physiologic Effect (12)	Prescription Drug (12)
Pharma/Biol Product (12)	Genseqno (10)	Multum (10)	Chemical Structure (10)	Brand Name (7)	Mechanism of Action (5)
Branded Drug Form (5)	Ingredient (5)	Pharma Preparation (4)	Clinical Drug Form (2)	VA Class (2)	Drug (1)
ATC 5th (1)	ATC 2nd (1)	ATC 4th (1)	ATC 1st (1)	Substance (1)	Cohort (1)
Pharmacologic Class (1)	ATC 3rd (1)				

Domain

Drug (2800)
-------------

Relationship

Standard to Non-standard map (OMOP) (2715)	Has ancestor of (72)	Has descendant of (71)	Has inferred drug class (OMOP) (68)	Ingredient of (RxNorm) (25)	Has trademark (RxNorm) (7)
RxNorm to SNOMED equivalent (RxNorm) (2)	RxNorm contained in DOI (OMOP) (1)	RxNorm to ATC equivalent by concept_name (OMOP) (1)	RxNorm to ATC (RxNorm) (1)	RxNorm to Multilex equivalent (OMOP) (2)	Has form (RxNorm) (2)
				NDF-RT to RxNorm equivalent by concept_name (OMOP) (1)	RxNorm to NDF-RT equivalent (RxNorm) (2)
					Non-standard to Standard map (OMOP) (1)

Distance

2 (2044)	0 (661)	1 (121)	3 (13)	4 (8)	5 (4)
6 (2)	7 (1)	8 (1)			

Show 100 entries

Search:

Concept Code Related Concept Class Domain Vocabulary

Concept Code	Related Concept	Class	Domain	Vocabulary
000560168	warfarin sodium 4mg/1 ORAL TABLET [coumadin]	9-digit NDC	Drug	NDC
00056016801	Warfarin Sodium 4 MG Oral Tablet [Coumadin]	11-digit NDC	Drug	NDC
00056016870	Warfarin Sodium 4 MG Oral Tablet [Coumadin]	11-digit NDC	Drug	NDC



# Use CIRCE to define the cohort of interest

CIRCE  
Cohort Inclusion and Restriction Criteria Expression

Cohort Definition List Help

Index Population: MiniSentinel replication - warfarin new users

Description:

Save

Expression Concept Sets Print Friendly Raw JSON Generate

People having any of the following: **Add Primary Event Filters...**

a drug era of warfarin **+ Add Filter...** Delete Filter

for the first time in the person's history

era start is: After 2010-11-01

with age at era start Greater or Equal To 21

with observation at least 180 days prior and 0 days after index

Limit primary events to: All Events per person.

**Add Additional Filters**

Limit cohort expression results to: All Events per person.

Show SQL Add Options



# Use CALYPSO to conduct feasibility assessment to evaluate the impact of study inclusion criteria

CAL  
Criteria /  
Population

Index Rule   Inclusion Rules   Concept Sets   Results

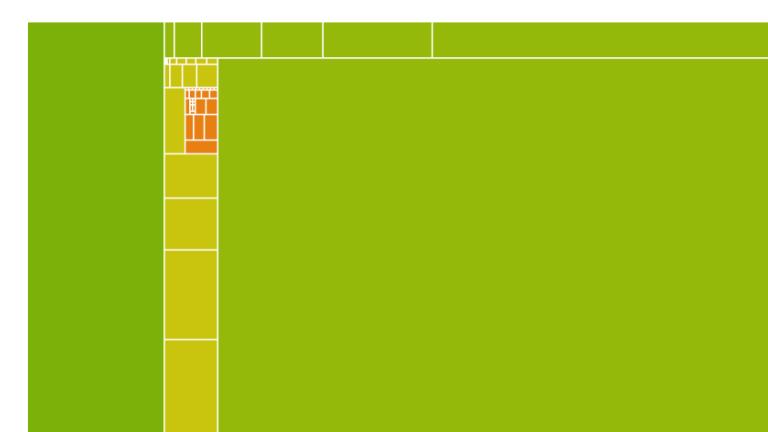
Source	Name	Dialect	Action
TRUVENCCAE	Truven CCAE (APS)	pdw	Generate
TRUVENMDCR	Truven MDCR (APS)	pdw	Generate
TRUVENMDCD	Truven MDCCD (APS)	pdw	Generate
OPTUM	Optum (APS)	pdw	Generate
CPRD	CPRD (APS)	pdw	Generate
PREMIER	Premier (APS)	pdw	Generate
JMDC	JMDC (APS)	pdw	Generate
NHANES	NHANES (APS)	pdw	Generate
VOCAB	Default Vocabulary	sql server	Generate
LAERTES	Laertes	postgresql	Generate

Overview   Reports

**Summary Statistics:**   Match Rate 18.15%   Matching Persons 12061   Total Persons 66443

Inclusion Rule	% Satisfied	% To-Gain
1. Prior atrial fibrillation	23.31%	71.19%
2. No prior warfarin ever	100.00%	0.00%
3. No prior dabigatran ever	98.80%	0.17%
4. No prior anticoagulants in past 183 days	98.05%	0.38%
5. No mechanical heart valve or mitral stenosis	94.99%	2.23%
6. No dialysis in last 30 days	98.97%	0.39%
7. No history of kidney transplant	99.61%	0.06%
8. Not at long-term care visit	97.29%	0.70%

**Population Visualization**





# Use HERACLES to characterize the cohorts you developed

OHDSI Heracles

«Back

Refresh

Truven MDCC (APS) ▾

Heracles Runner

Cohort Specific

Condition

Condition Eras

Conditions by Index

Dashboard

Data Density

Death

Drug Eras

Drug Exposures

Drugs by Index

Heracles Heel

Conditions by Index

Dashboard

Data Density

Death

Drug Eras

Drug Exposures

Drugs by Index

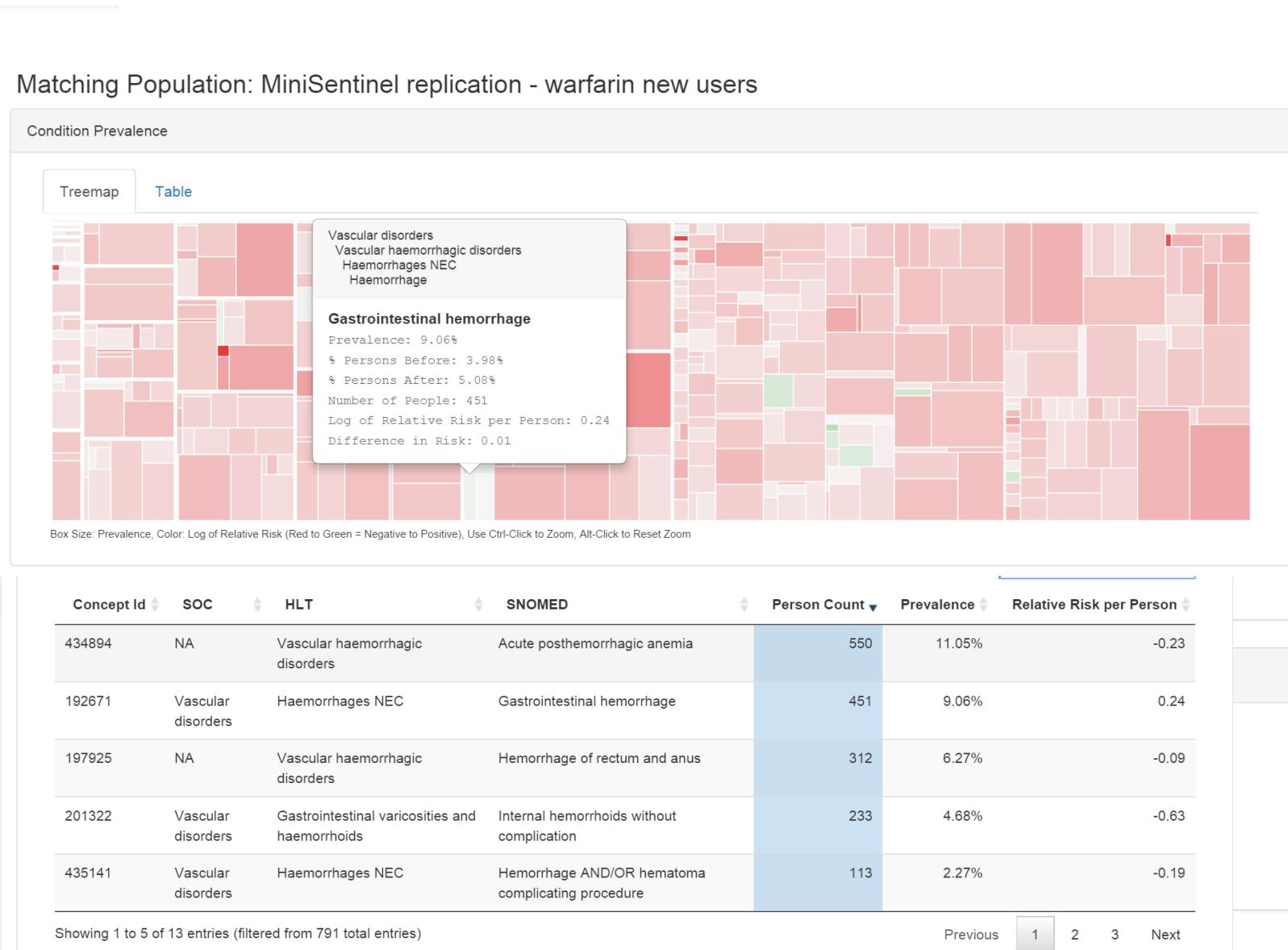
Heracles Heel

Measurements

Observation Periods

Observations

Person





# Use HERACLES to characterize the cohorts you developed

## OHDSI Heracles

«Back

Refresh

Truven MDCCD (APS) ▾

Heracles Runner

Cohort Specific

Condition

Condition Eras

Conditions by Index

Dashboard

Data Density

Death

Drug Eras

Drug Exposures

Drugs by Index

Heracles Heel

Measurements

Observation Periods

Observations

Person

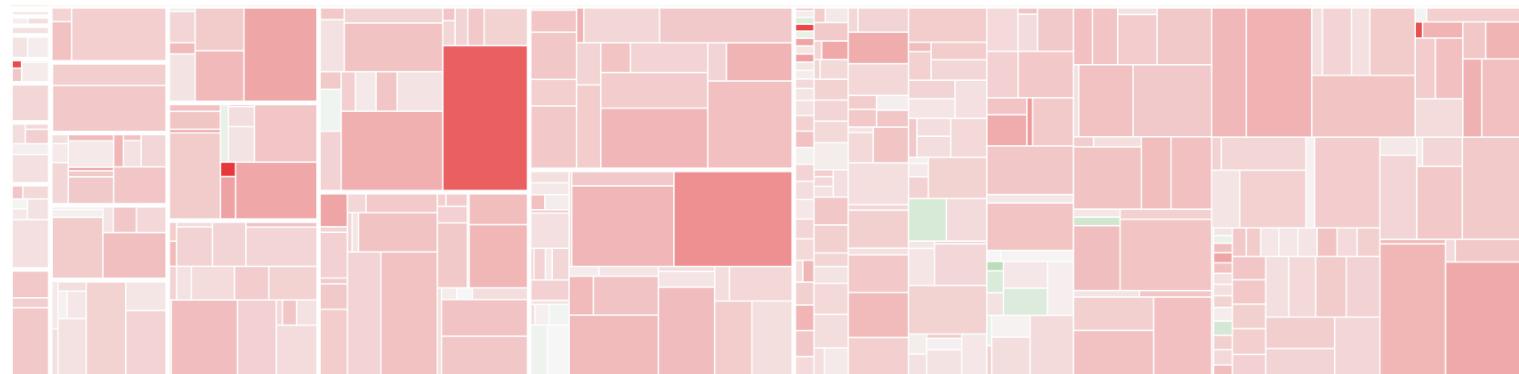
Procedures

Procedures by Index

Visits

Treemap

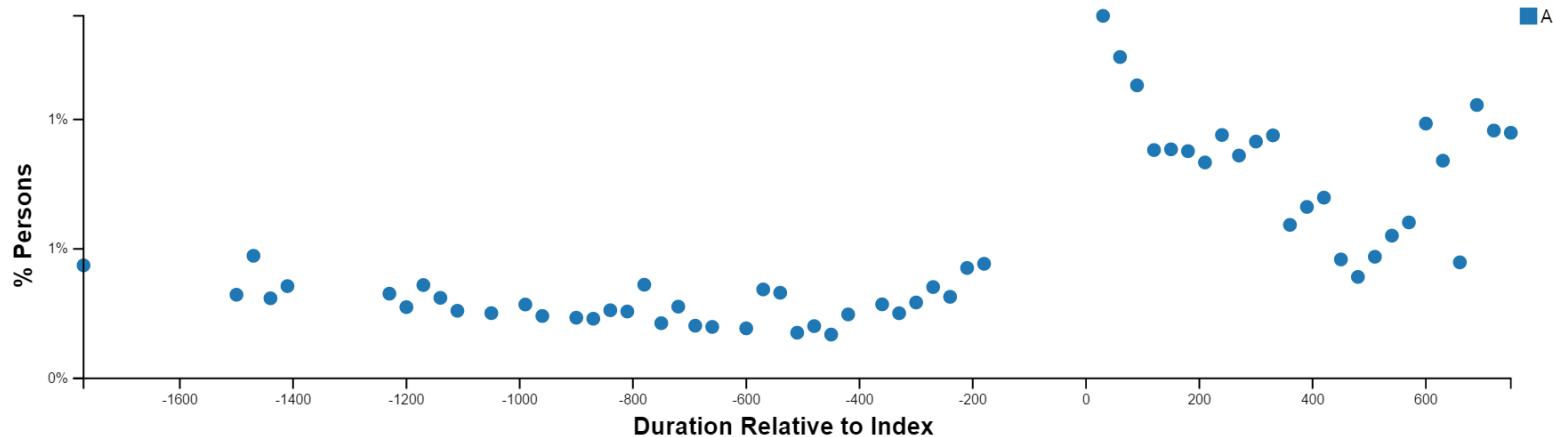
Table



Box Size: Prevalence, Color: Log of Relative Risk (Red to Green = Negative to Positive), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

## Gastrointestinal hemorrhage

First Condition Occurrence Relative to Index [Download](#)





# Use LAERTES to summarize evidence from existing data sources





# Step up to Advanced Analytic Methods

<https://github.com/OHDSI?utf8=%E2%9C%93&query=cohort>

Search GitHub

Explore Gist Blog Help

pbr6cornell + ·

## Observational Health Data Sciences and Informatics

http://ohdsi.org

Filters ▾  + New repository

### CohortMethod

An R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model.

Updated 10 days ago

### SelfControlledCohort

[Under development] Method to estimate risk by comparing time exposed with time unexposed among the exposed cohort

Updated on Dec 22, 2014

### People

35 >

Invite someone

### Teams

4 >

Jump to a team



# Open-source large-scale analytics through R

## Package ‘CohortMethod’

February 23, 2015

Type Package

Title New-user cohort method with large scale propensity and outcome models

Version 1.0.0

Date 2015-02-02

Author Martijn J. Schuemie [aut, cre], Marc A. Suchard [aut], Patrick B. Ryan [aut]

Maintainer Martijn J. Schuemie <schuemie@ohdsi.org>

Description CohortMethod is an R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model. It extracts the necessary data from a database in OMOP Common Data Model format, and uses a large set of covariates for both the propensity and outcome model, including for example all drugs, diagnoses, procedures, as well as age, comorbidity indexes, etc. Large scale regularized regression is used to fit the propensity and outcome models. Functions are included for trimming, stratifying and matching on propensity scores, as well as diagnostic functions, such as propensity score distribution plots and plots showing covariate balance before and after matching and/or trimming. Supported outcome models are (conditional) logistic regression, (conditional) Poisson regression, and (conditional) Cox regression.

License Apache License 2.0

VignetteBuilder knitr

Depends R (>= 3.1.0), bit, DatabaseConnector, Cyclops (>= 1.0.0)

Imports ggplot2, ff, ffbase, plyr, Rcpp (>= 0.11.2), RJDBC, SqlRender (>= 1.0.0), survival

Suggests testthat, pROC, gnm, knitr, rmarkdown

LinkingTo Rcpp

NeedsCompilation yes

## Why is this a novel approach?

- Large-scale analytics, scalable to ‘big data’ problems in healthcare:
  - millions of patients
  - millions of covariates
  - millions of questions
- End-to-end analysis, from CDM through evidence
  - No longer de-coupling ‘informatics’ from ‘statistics’ from ‘epidemiology’



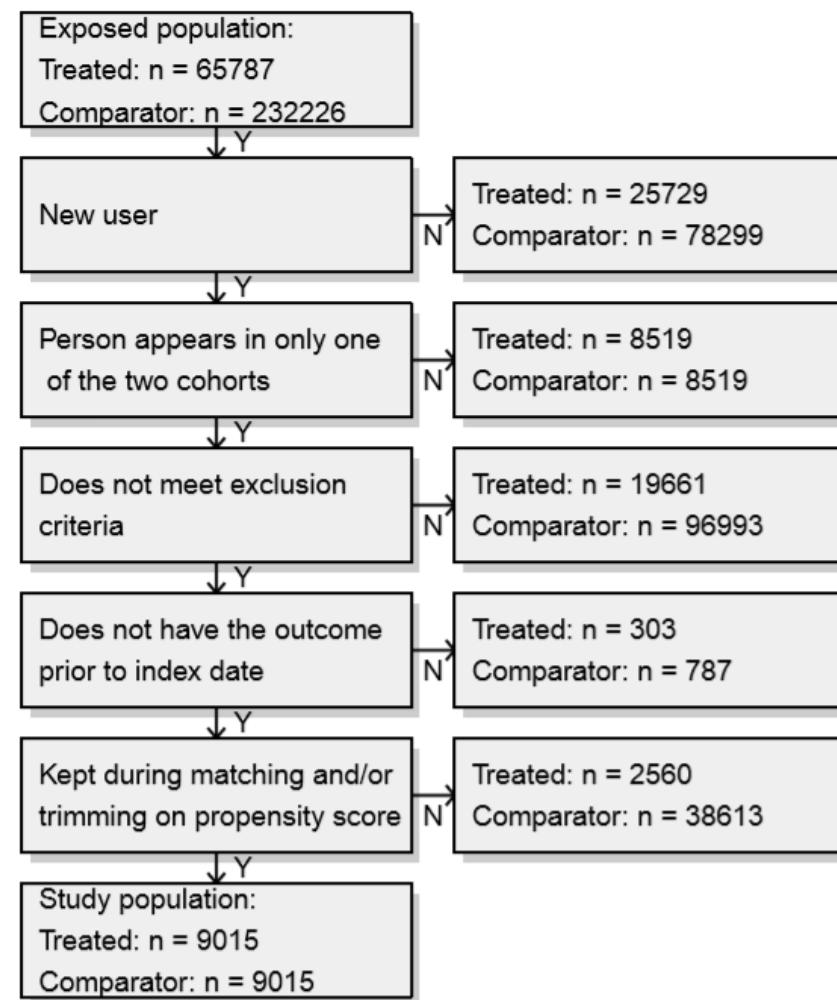
# Standardize Analysis and Results Reporting

```
summary(outcomeModel)
```

```
#> Model type: cox  
#> Status: c plotKaplanMeier(outcomeModel)  
#>  
#> Counts  
#>  
#> Nr. of persons  
#> Nr. of events  
#> Person time  
#>  
#> Model  
#> N  
#>  
#> Coefficients  
#>  
#> treatment  
#>  
#> Prior variables
```

The plot shows survival probability on the y-axis (ranging from 0.95 to 1.00) versus time in days on the x-axis (ranging from 0 to 400). Two curves are shown: a blue curve for the treated group and a red curve for the comparator group. Both curves start at 1.00 at day 0. The blue curve remains higher than the red curve throughout the observed period, indicating better survival probability for the treated group.

```
drawAttritionDiagram(outcomeModel)
```





# Demo!

**OLYMPUS**  
THE OHDSI APPLICATION LAUNCHER

There are remote WebAPIs configured. Applications that support toggling between WebAPIs will allow you to use these via the gear/settings.

ATHENA OMOP Vocabulary Loader	HERMES OMOP Vocabulary Explorer	ACHILLES Dataset Characterization
CIRCE Cohort Creation	HERACLES Cohort Characterization	CALYPSO Clinical Trial Feasibility



# Concluding Thoughts

- Open science requires optimized technical infrastructure, community infrastructure, and dedication
- But open science is not charity!
  - The payoff can be both for individual participants and the community
- A diversity of skillsets brings value to all and greatly accelerates generation of high quality evidence



# Join the journey

Interested in OHDSI?  
Questions or comments?

Contact:

[jonduke@regenstrief.org](mailto:jonduke@regenstrief.org)