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

Slide Credits: Patrick Ryan

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:

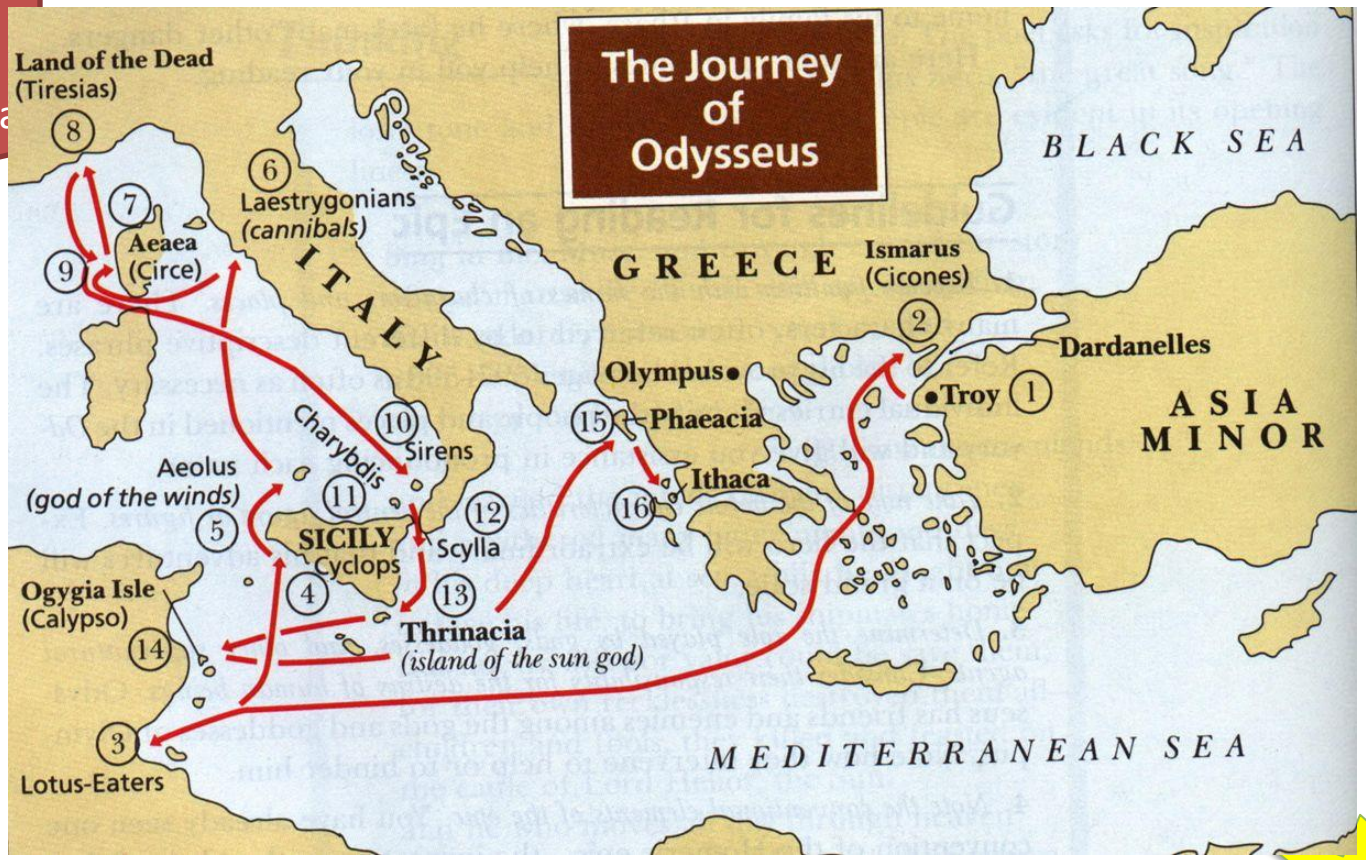
- >50 databases standardized to OMOP common data model
- >680 million patients

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

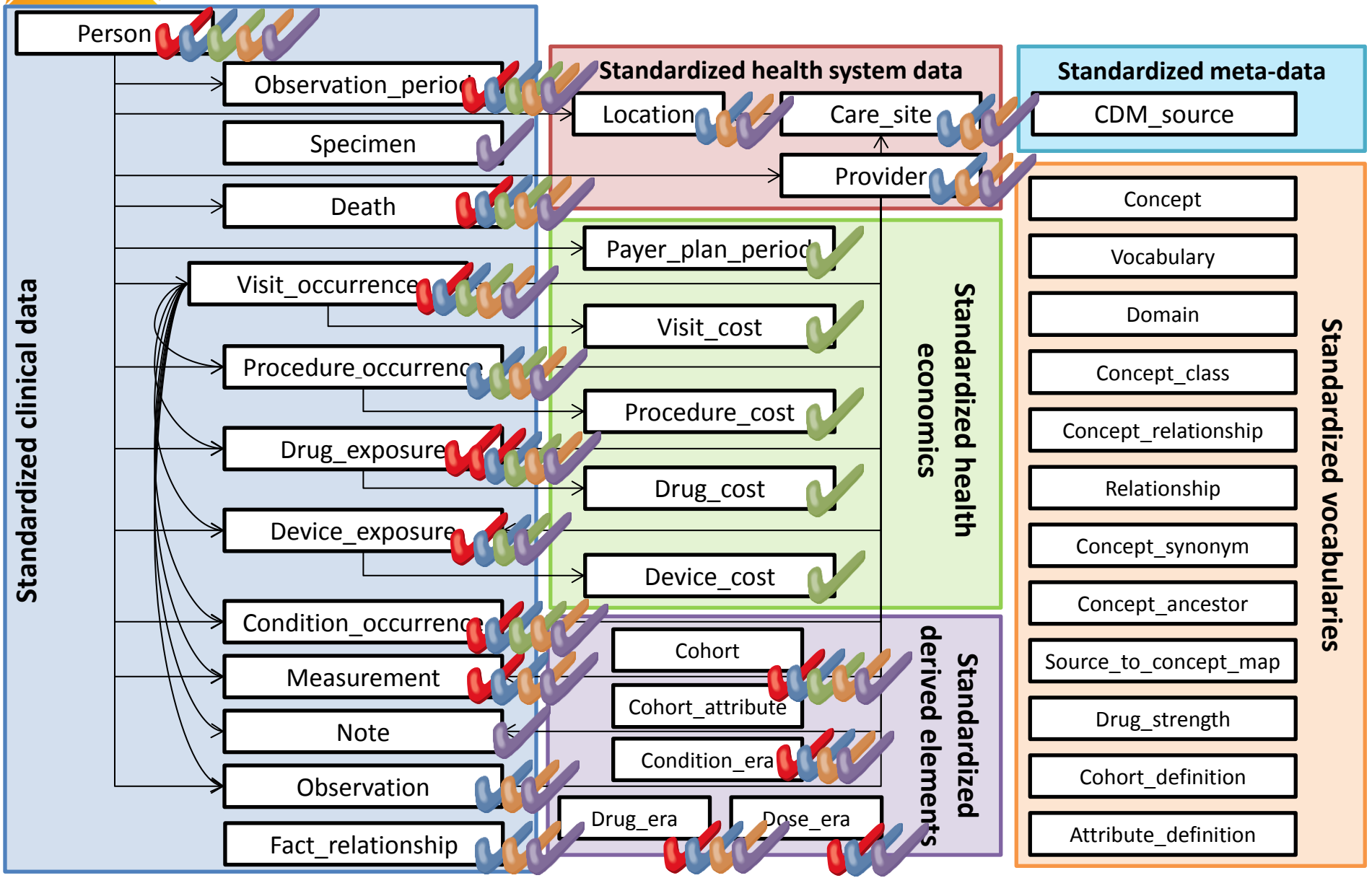
- The OHDSI community has standardized core components of the research process in order to
  - Promote transparent, reproducible science
  - Reveal data quality issues
  - ‘Calibrate’ datasets
  - Bring skillsets together from across the community (clinical, epi, stats, compSci)

Opportunities for standardization in the evidence generation process

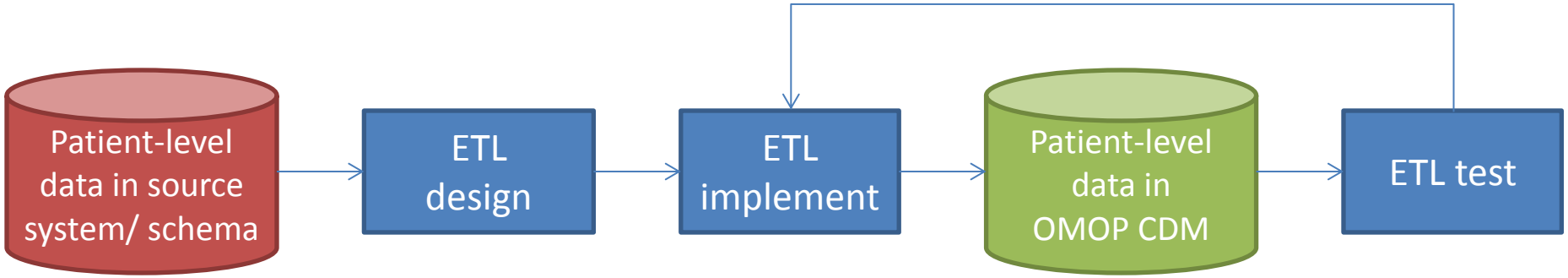
Protocol

- **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

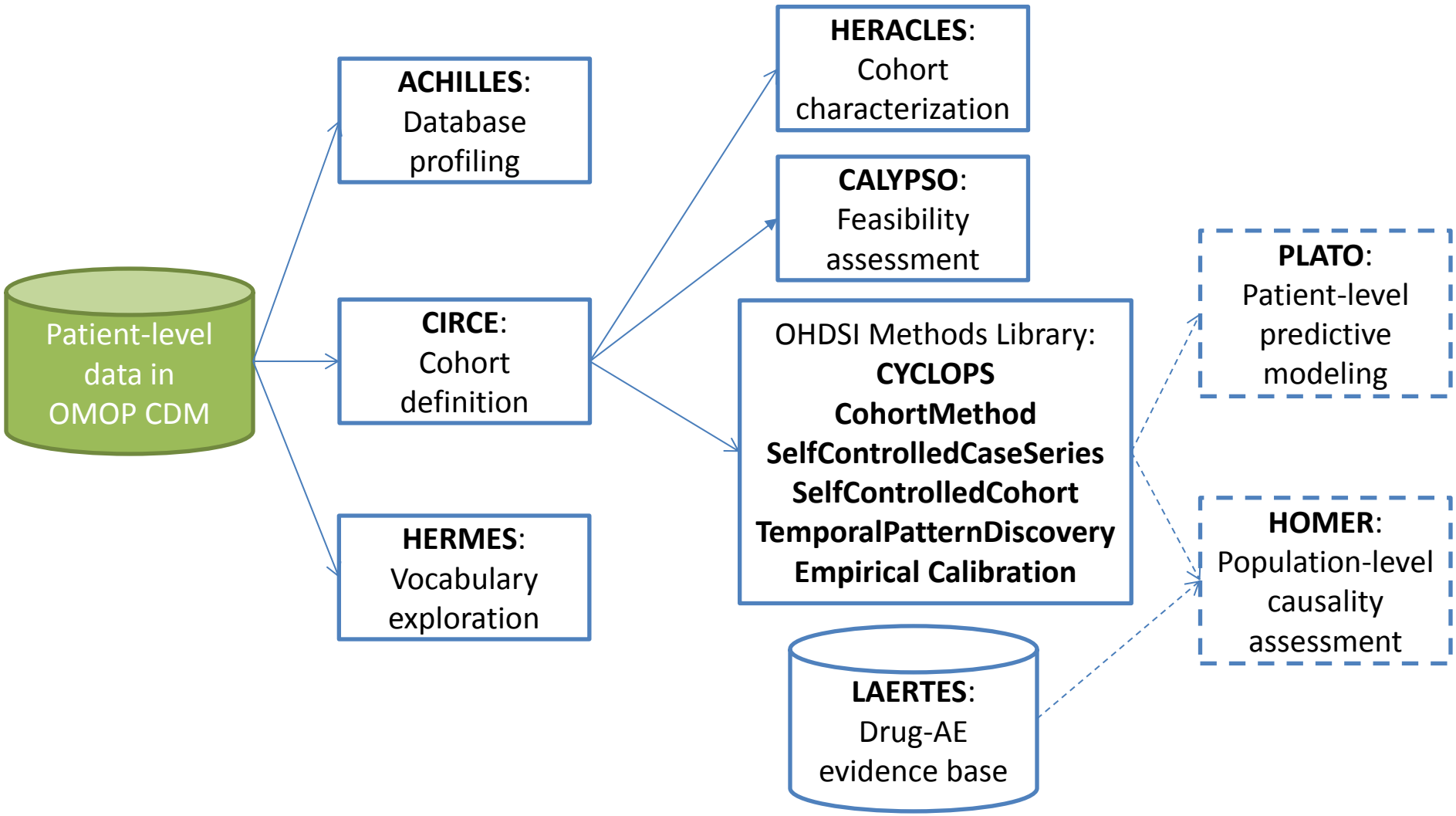
Usagi: map your source codes to CDM vocabulary

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

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[1], Daniel Singer, MD[2], T. Craig Cheetham, PharmD MS[3], Darren Toh, ScD[4], Marsha Reichman, PhD[5], David Graham, MD MPH[5], Mary Ross Southworth, PharmD[6], Rongmei Zhang PhD[7], Monika Houstoun, PharmD[5], Yu-te Wu PhD[7], Katrina Mott MS[5], Joshua Gagne, PharmD ScD[8]

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

III. PROTOCOL DETAILS

A. ASSESSMENT DESIGN

This one-time assessment will employ a "new user" parallel cohort design.[12]

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**.

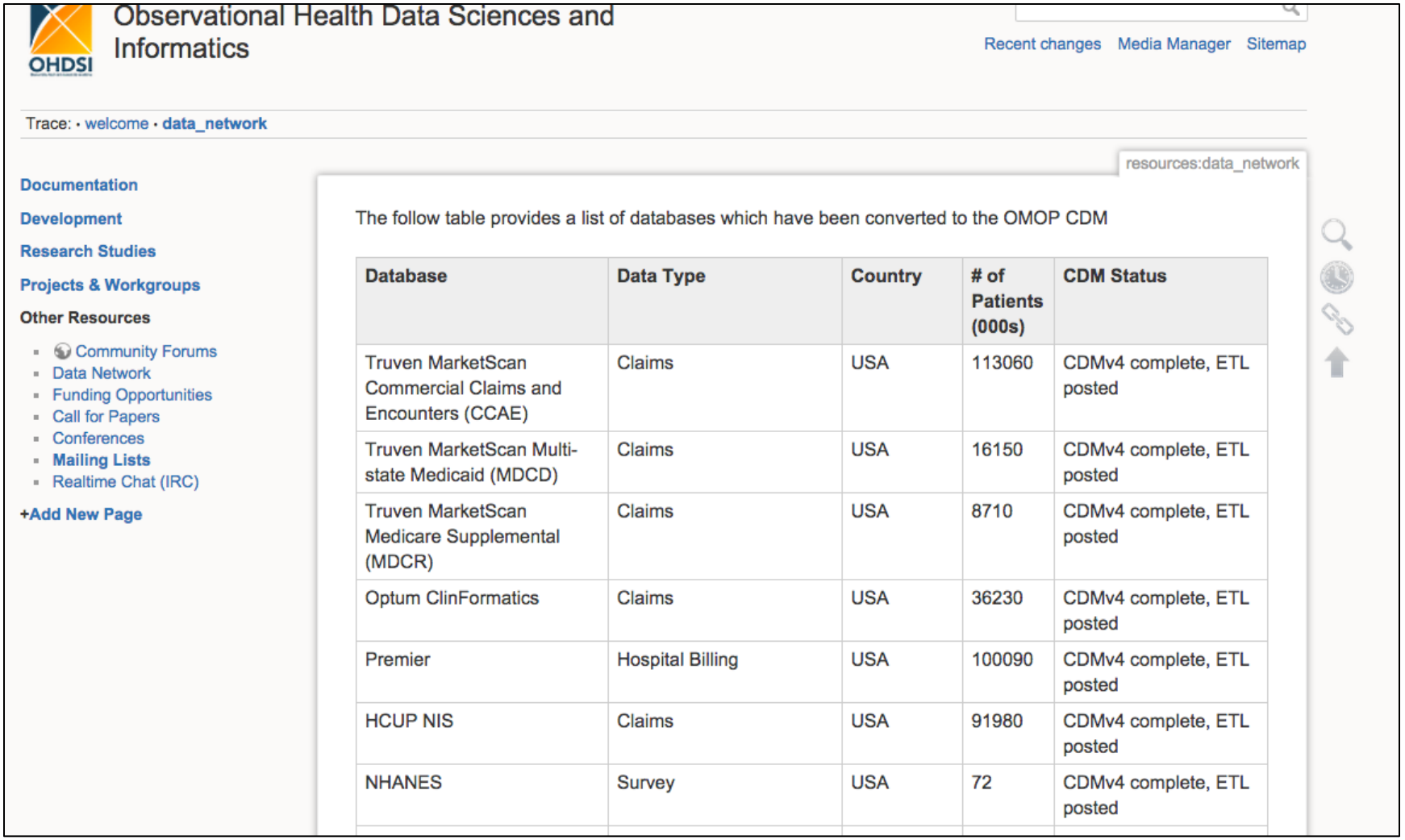
2. Sample Inclusion and Exclusion Criteria

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
• First dispensing of dabigatran or warfarin therapy from November 1, 2010 to the most recent data available in the MSDD from participating Data Partners *	• Less than 180 days of continuous enrollment with prescription and medical coverage immediately preceding the date of the index dispensing (i.e., index date)
• Age 21 years or older at the first dispensing of dabigatran or warfarin therapy	• Any prior dispensing for warfarin, dabigatran, rivaroxaban or apixaban during the 180 days before index date **
• 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 *	• Known mechanical heart valve or diagnosed mitral stenosis at index date based on corresponding administrative diagnosis and/or procedure codes • Chronic hemodialysis or peritoneal dialysis at index date based on corresponding administrative diagnosis and/or procedure codes • History of kidney transplant at index date based on corresponding administrative diagnosis and/or procedure codes • At a skilled nursing facility or nursing home at index date

Let's ask the OHDSI network!



Observational Health Data Sciences and Informatics

OHDSI

Recent changes Media Manager Sitemap

Trace: • welcome • data_network

resources:data_network

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

The follow 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

OHDSI

Home | Applications | Job Status | Settings | Logout

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

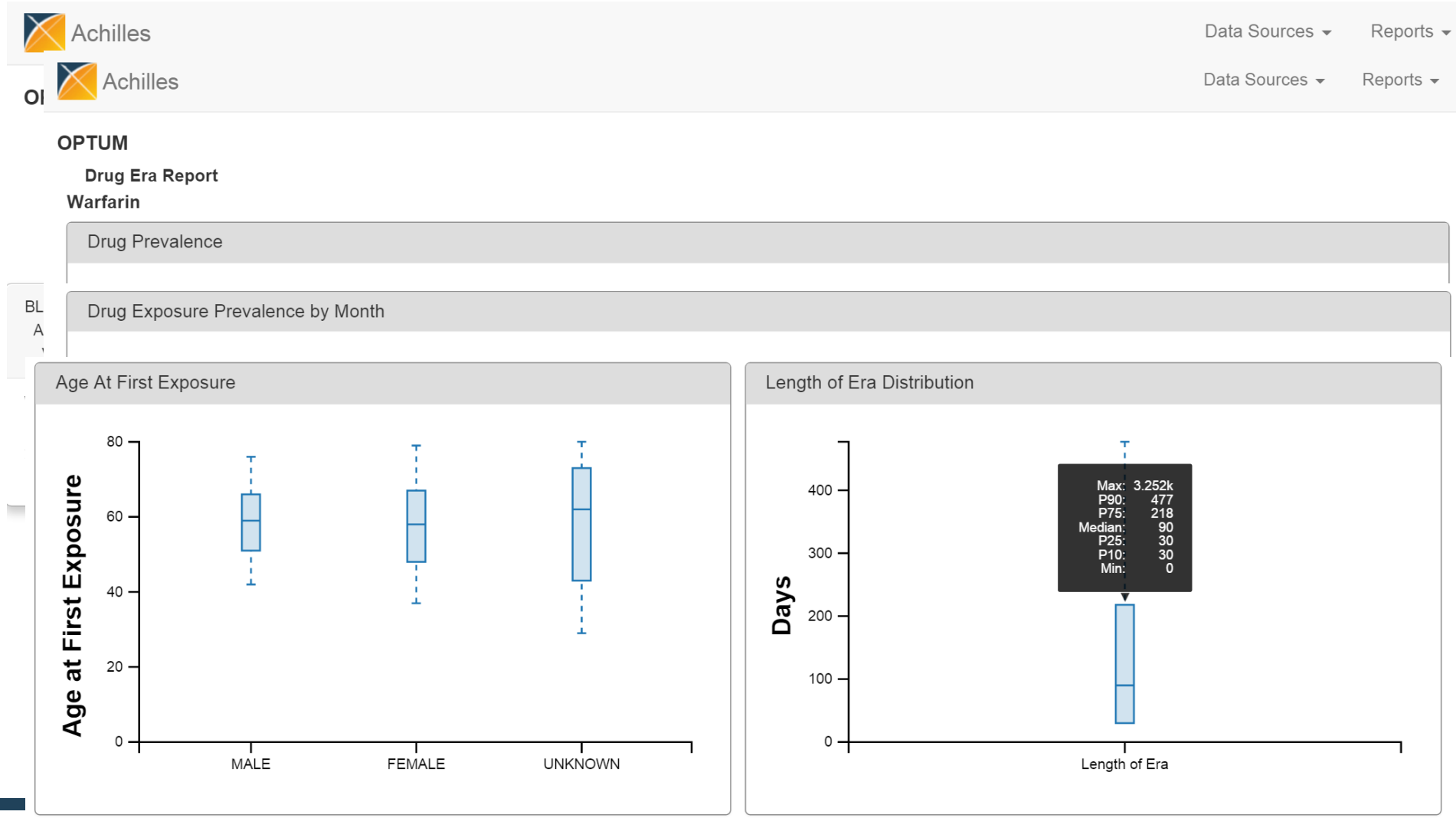
HERACLES

Cohort Characterization

CALYPSO

Clinical Trial Feasibility

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



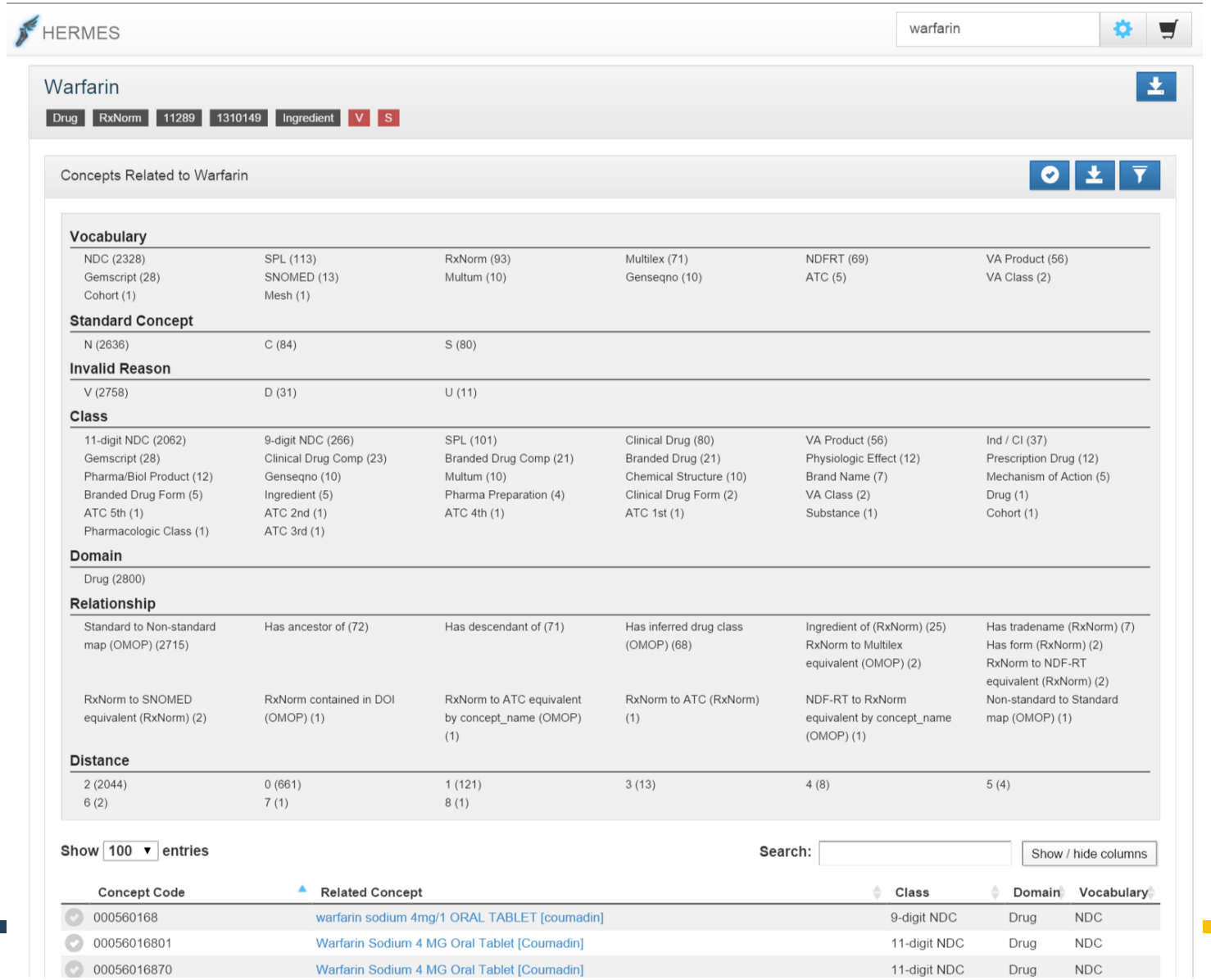
Also use ACHILLES to check for any data quality issues

Data Quality Messages

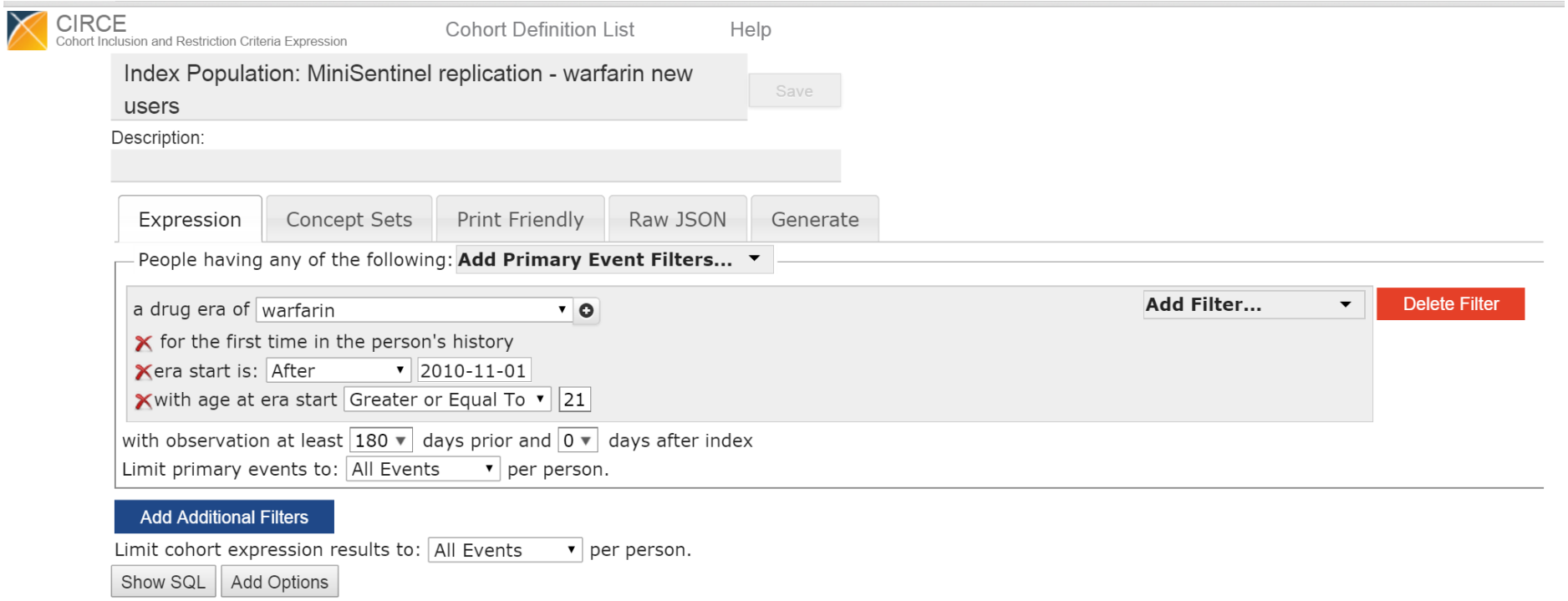
Search: Show / hide columns

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



Use CIRCE to define the cohort of interest



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

CAL
Criteria A
Populatio

Index Rule | Inclusion Rules | Concept Sets | Results

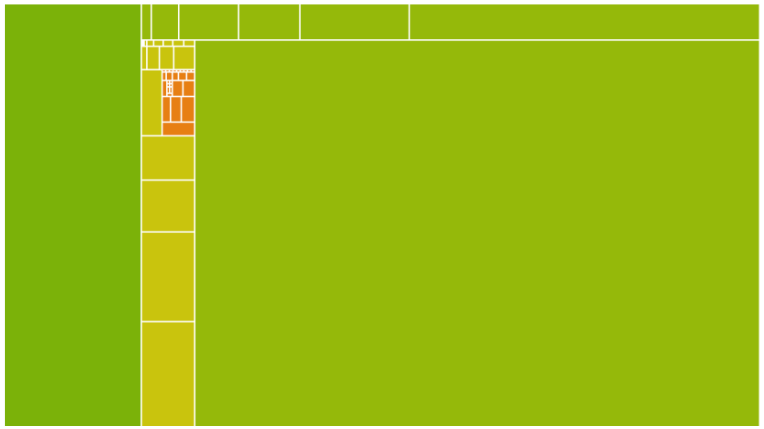
Source	Name	Dialect	
TRUVENCCAE	Truven CCAE (APS)	pdw	Generate
TRUVENMDCR	Truven MDCR (APS)	pdw	Generate
TRUVENMDCD	Truven MDCD (APS)	pdw	Generate
OPTUM	Optum (APS)	pdw	Generate
CPRD	CPRD (APS)	pdw	Generate
PREMIER	Premier (APS)	pdw	Generate
JMDC	JMDC (APS)	pdw	Generate
NHANES	NHANES (APS)	pdw	Generate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

Overview | Reports

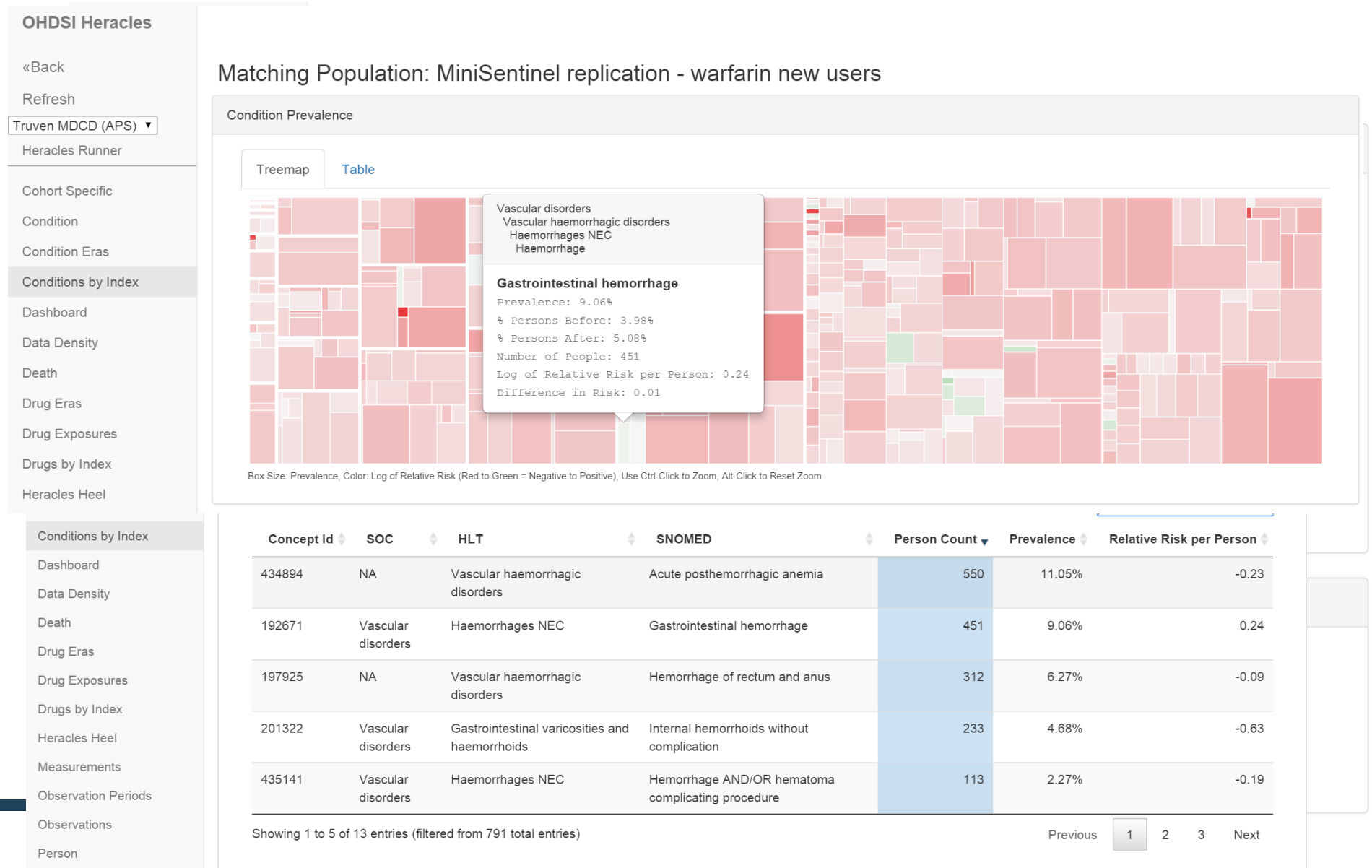
	Match Rate	Matching Persons	Total Persons
Summary Statistics:	18.15%	12061	66443

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 value 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

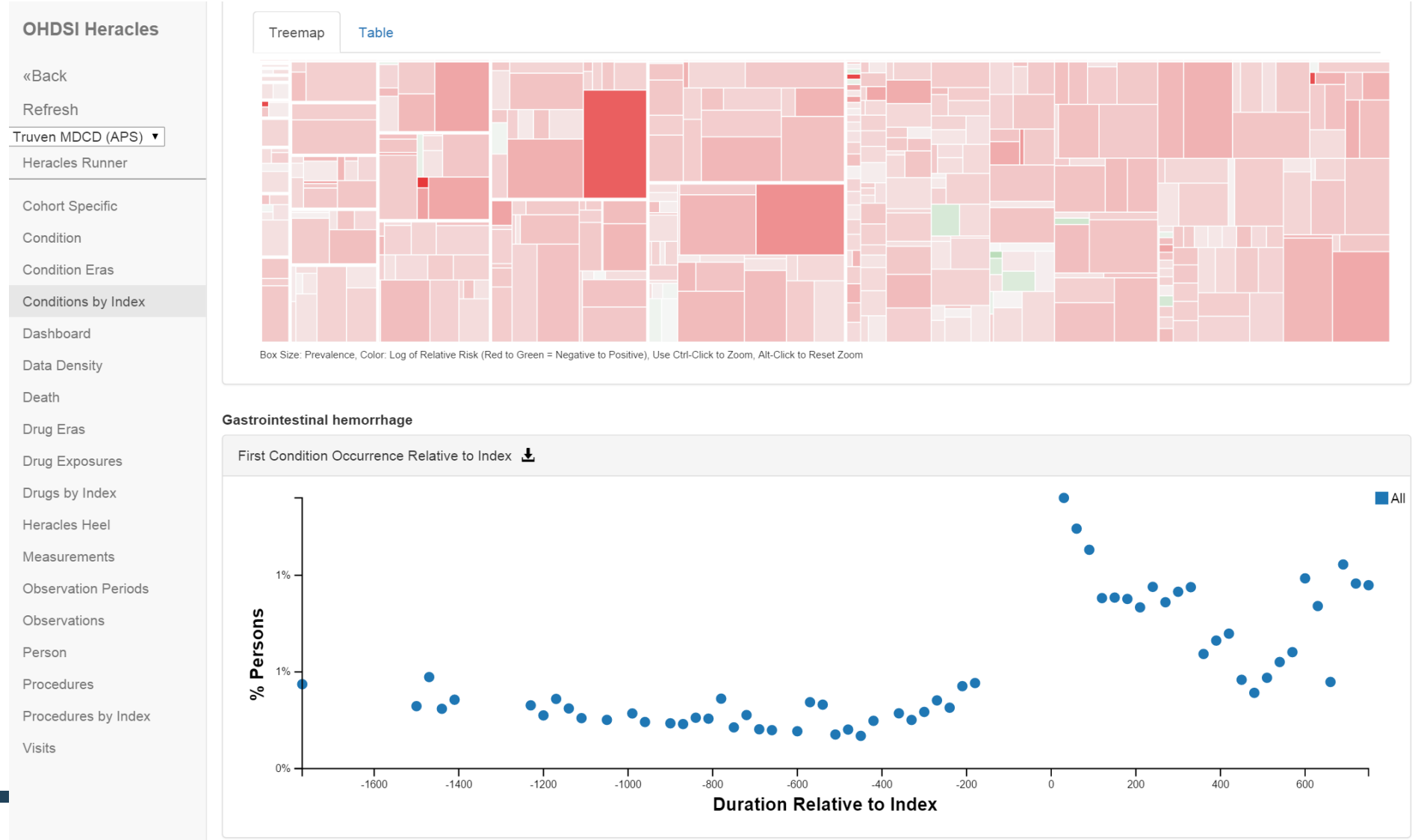


Concept Id	SOC	HLT	SNOMED	Person Count	Prevalence	Relative Risk per Person
434894	NA	Vascular haemorrhagic disorders	Acute posthemorrhagic anemia	550	11.05%	-0.23
192671	Vascular disorders	Haemorrhages NEC	Gastrointestinal hemorrhage	451	9.06%	0.24
197925	NA	Vascular haemorrhagic disorders	Hemorrhage of rectum and anus	312	6.27%	-0.09
201322	Vascular disorders	Gastrointestinal varicosities and haemorrhoids	Internal hemorrhoids without complication	233	4.68%	-0.63
435141	Vascular disorders	Haemorrhages NEC	Hemorrhage AND/OR hematoma complicating procedure	113	2.27%	-0.19

Showing 1 to 5 of 13 entries (filtered from 791 total entries)

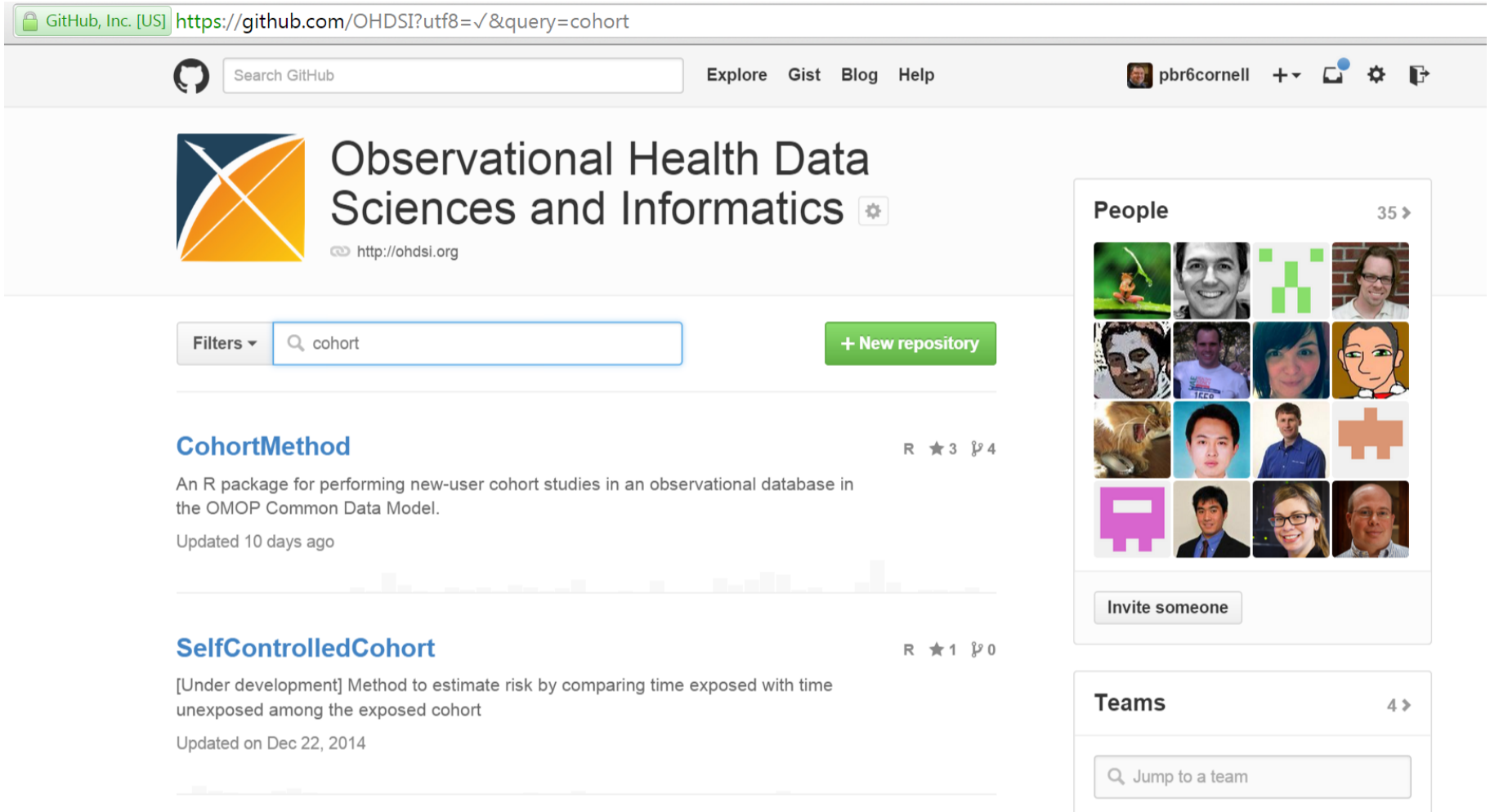
Previous 1 2 3 Next

Use HERACLES to characterize the cohorts you developed



Use LAERTES to summarize evidence from existing data sources

Step up to Advanced Analytic Methods



http://github.com/OHDSI

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?

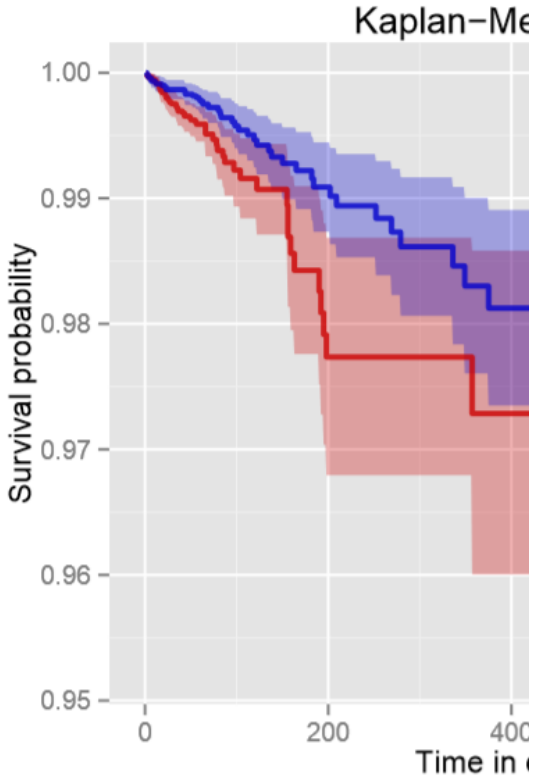
- Large-scale analytics, scalable to 'big data' problems in healthcare:
  - millions of patients
  - millions of covariates
  - millions of questions
- End-to-end analysis, from CDM through evidence
  - No longer de-coupling 'informatics' from 'statistics' from 'epidemiology'

Standardize Analysis and Results Reporting

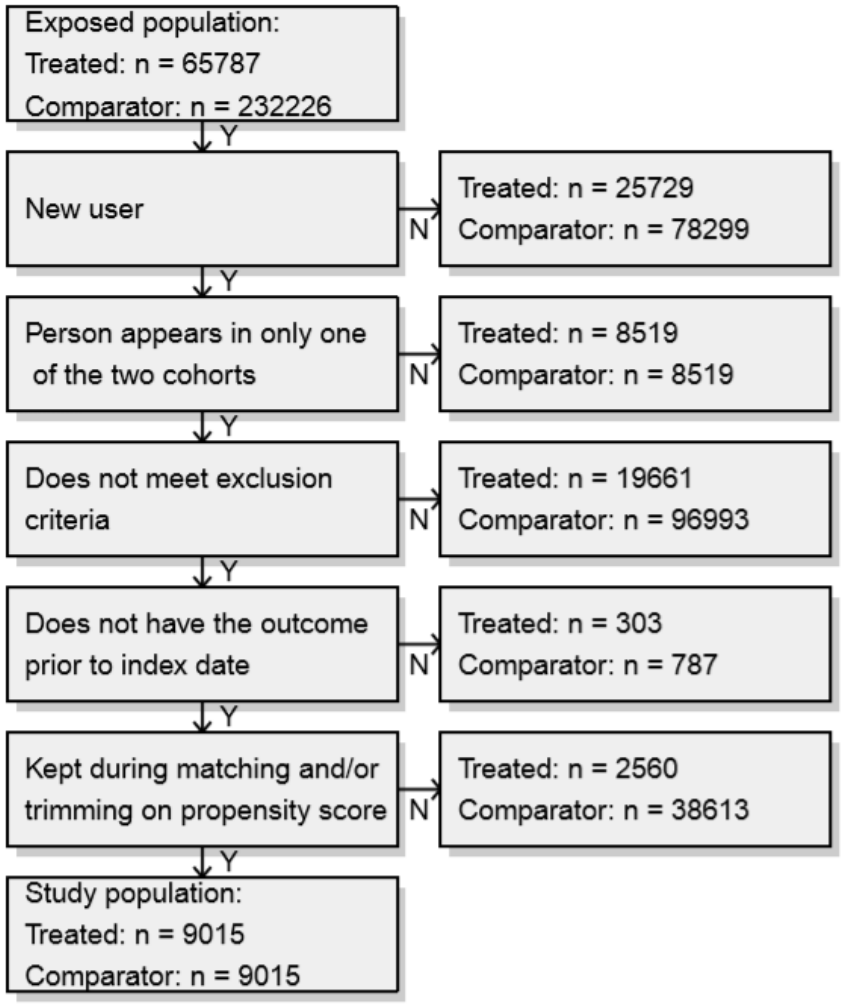
```
summary(outcomeModel)
```

```
#> Model type: cox
#> Status: C
#>
#> Counts
#>
#> Nr. of per
#> Nr. of eve
#> Person tim
#>
#> Model
#>
#>
#>
#> Coefficier
#>
#> treatment
#>
#> Prior vari
```

```
plotKaplanMeier(outcome
```




```
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