OHDSI

Observational Health Data Sciences and Informatics

Open-Source Big Data Analytics in Healthcare

Jon Duke, George Hripcsak, Patrick Ryan

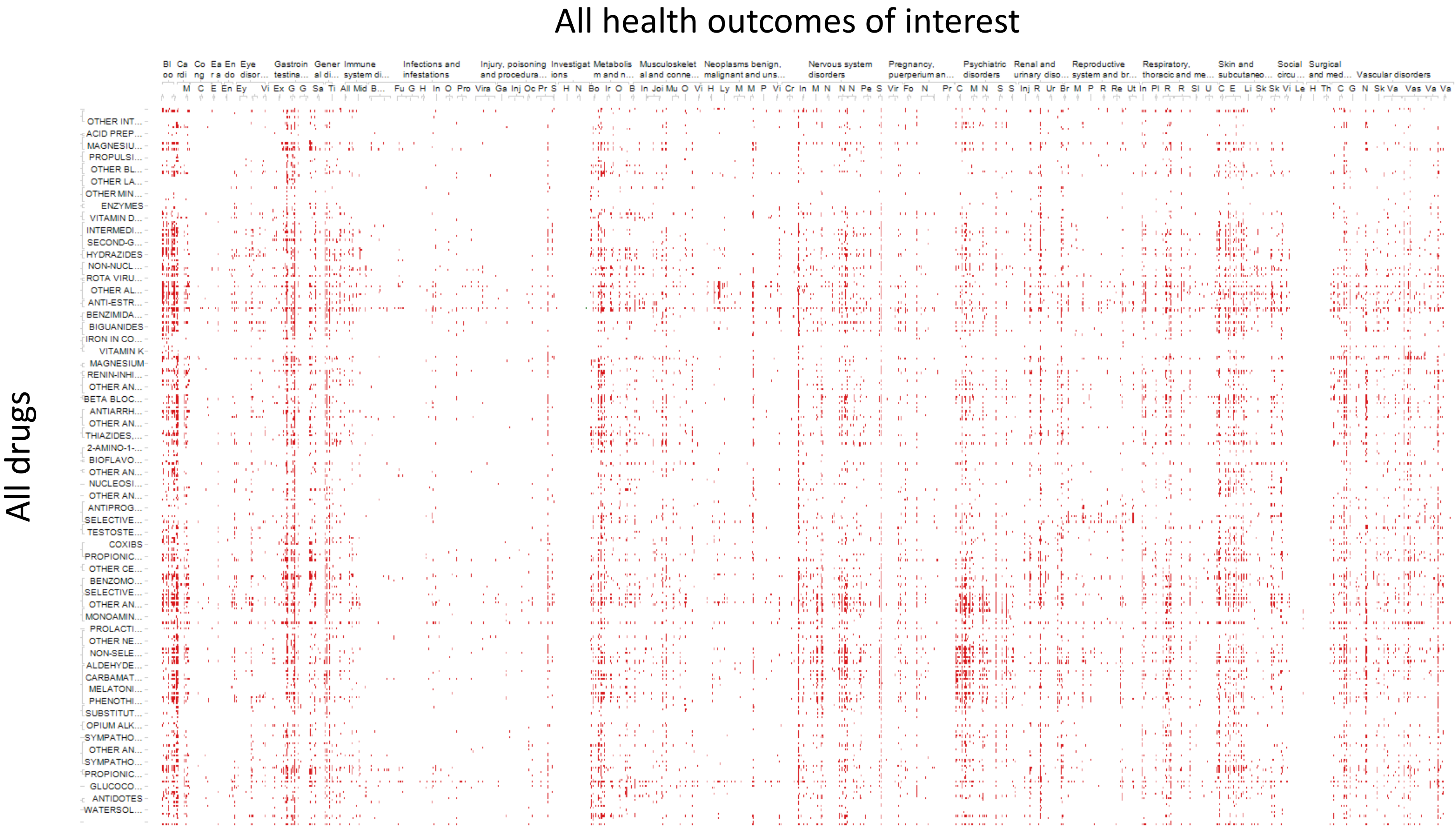
www.ohdsi.org/medinfo-2015-tutorial

Introduction

Introducing OHDSI

- The Observational Health Data Sciences and Informatics (OHDSI) program is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics
- OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University

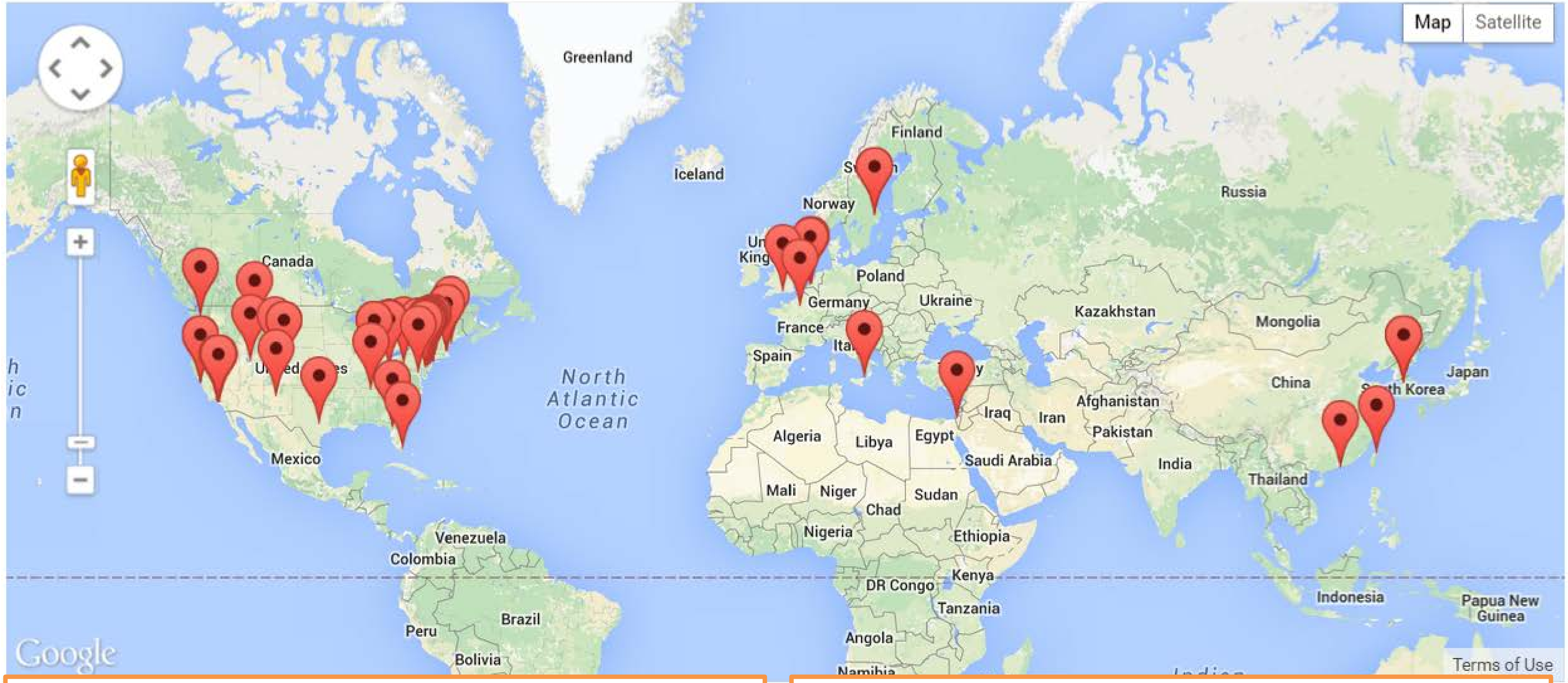
Why large-scale analysis is needed in healthcare



OHDSI’s vision

OHDSI collaborators access a network of **1,000,000,000 patients to generate evidence** about all aspects of healthcare. Patients and clinicians and other decision-makers around the world use OHDSI tools and evidence every day.

OHDSI: a global community



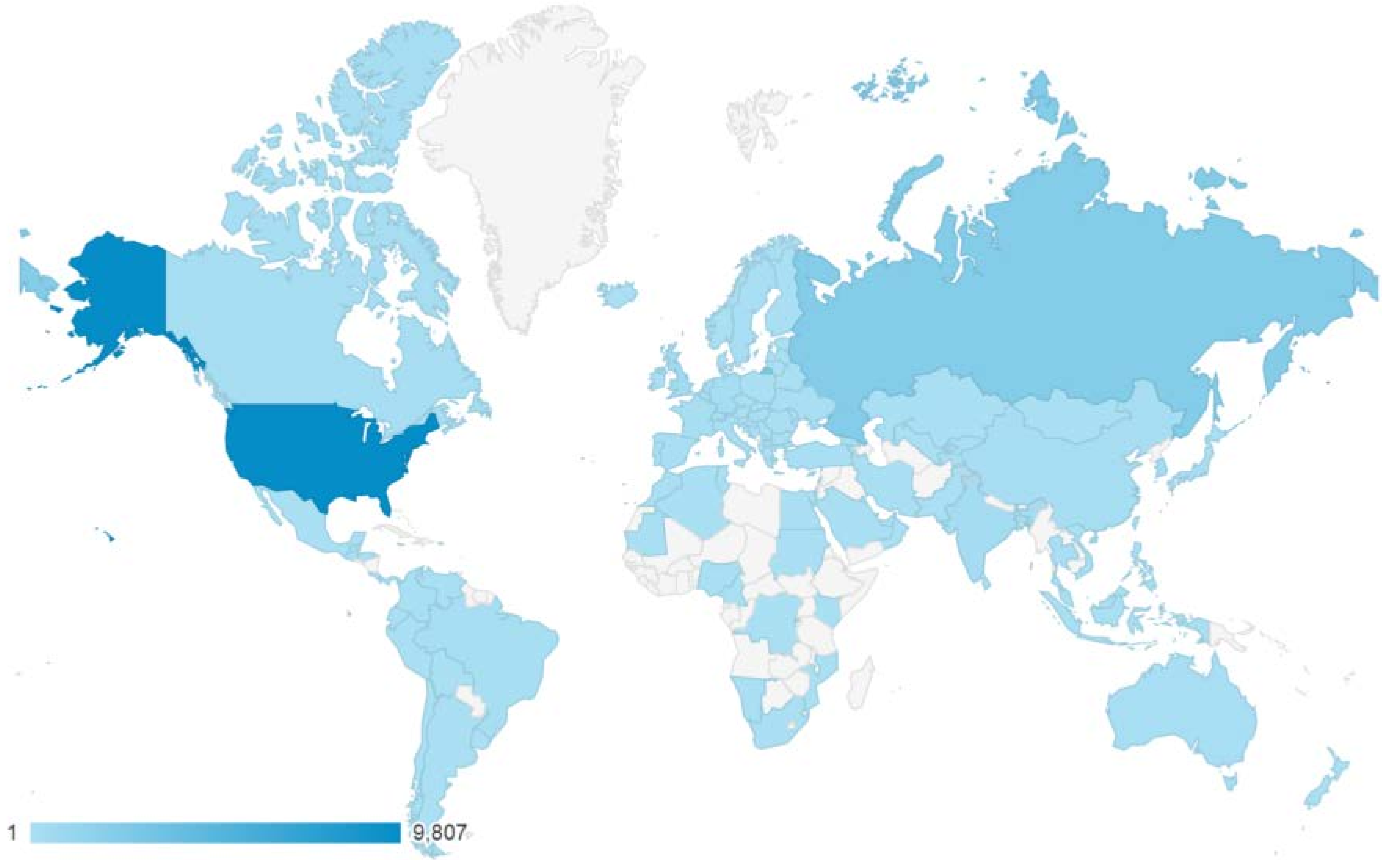
OHDSI Collaborators:

- >100 researchers in academia, industry and government
- >10 countries

OHDSI Data Network:

- >40 databases standardized to OMOP common data model
- >500 million patients

Global reach of ohdsi.org



- >10,000 distinct viewers from 110 countries in 2015

OHDSI’s guiding principles

- **Evidence-based**: OHDSI’s scientific research and development will be driven by objective, empirical evidence to ensure accuracy and reliability in everything we do
- **Practical**: OHDSI will go beyond methodological research, developing applied solutions and generating clinical evidence
- **Comprehensive**: OHDSI aims to generate reliable scientific evidence for all interventions and all outcomes
- **Transparent**: All work products within OHDSI will be open source and publicly available, including source code, analysis results, and other evidence generated in all our activities. Best practices for large-scale open source collaboration will guide development activities
- **Inclusive**: OHDSI encourages active participation from all stakeholders – patients, providers, payers, government, industry, academia – in all phases of research and development
- **Secure**: OHDSI will protect patient privacy and respect data holder interests at all times in our work

http://OHDSI.org

- To achieve the principle of inclusivity, OHDSI is an open collaborative. Anyone who can give time, data, or funding is welcome, and participation in the operation of OHDSI is expected.

Evidence OHDSI seeks to generate from observational data

- Clinical characterization:
  - Natural history: Who are the patients who have diabetes? Among those patients, who takes metformin?
  - Quality improvement: what proportion of patients with diabetes experience disease-related complications?
- Population-level estimation
  - Safety surveillance: Does metformin cause lactic acidosis?
  - Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?
- Patient-level prediction
  - Precision medicine: Given everything you know about me and my medical history, if I start taking metformin, what is the chance that I am going to have lactic acidosis in the next year?
  - Disease interception: Given everything you know about me, what is the chance I will develop diabetes?

OHDSI ongoing collaborative activities

	Methodological research	Open-source analytics development	Clinical applications
Observational data management	• Data quality assessment • Common Data Model evaluation • ATHENA for standardized vocabularies	• WhiteRabbit for CDM ETL • Usagi for vocabulary mapping • HERMES for vocabulary exploration • ACHILLES for database profiling	
Clinical characterization	• Phenotype evaluation	• CIRCE for cohort definition • CALYPSO for feasibility assessment • HERACLES for cohort characterization	• Chronic disease therapy pathways
Population-level estimation	• Empirical calibration • LAERTES for evidence synthesis	• CohortMethod • SelfControlledCaseSeries • SelfControlledCohort • TemporalPatternDiscovery	• HOMER for causality assessment
Patient-level prediction	• Evaluation framework and benchmarking	• PatientLevelPrediction • APHRODITE for predictive phenotyping	• PENELOPE for patient-centered product labeling

Open Science through Standardization

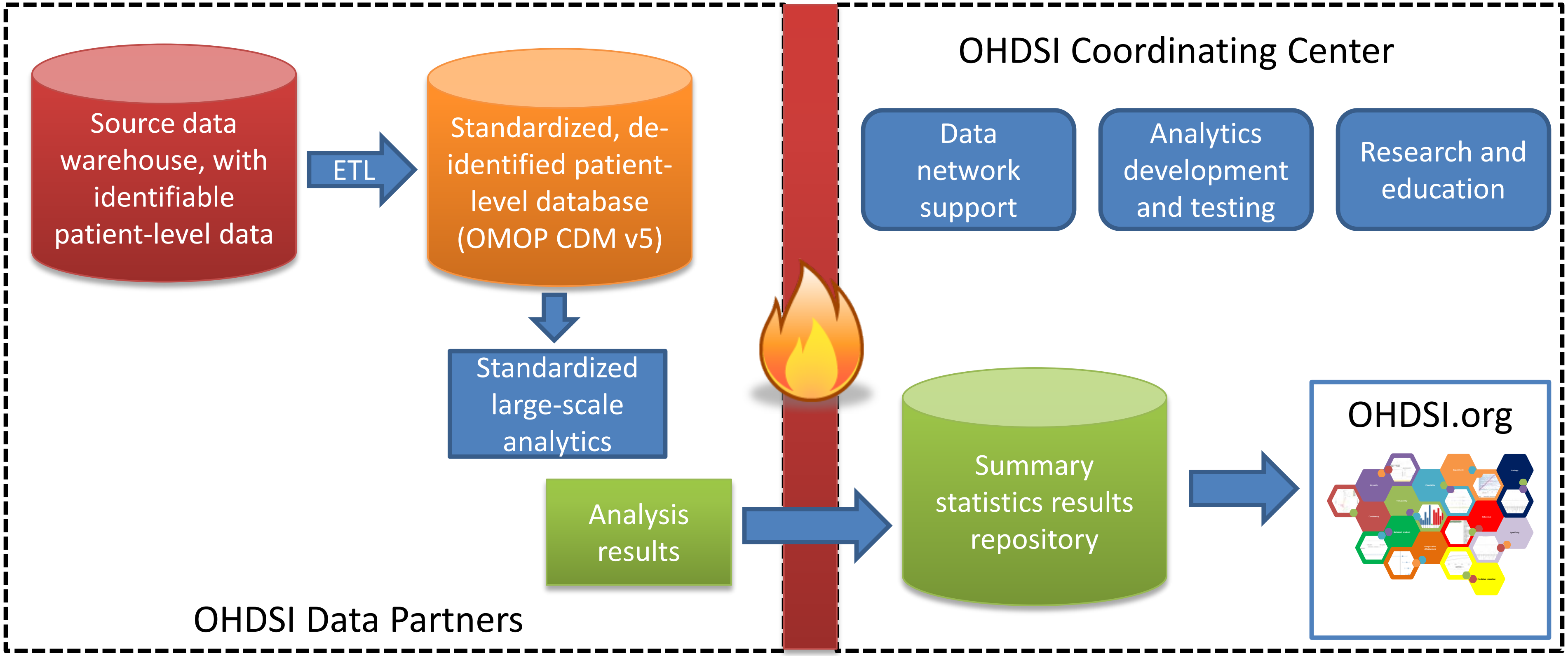
- The OHDSI community has standardized core components of the research process in order to
  - Promote transparent, reproducible science
  - Reveal data quality issues
  - ‘Calibrate’ datasets
  - Bring skillsets together from across the community (clinical, epi, stats, compSci)

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 for a given interval of time
- **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

How OHDSI Works



OHDSI Data Partners

Source data warehouse, with identifiable patient-level data

ETL

Standardized, de-identified patient-level database (OMOP CDM v5)

Standardized large-scale analytics

Analysis results

OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

Summary statistics results repository

OHDSI.org

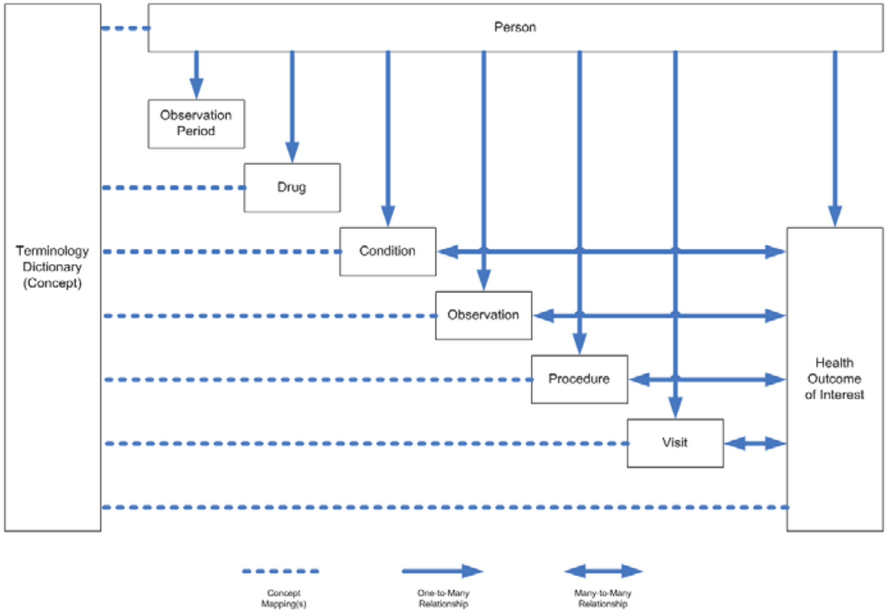
Objectives in OMOP Common Data Model development

- One model to accommodate both administrative claims and electronic health records
  - Claims from private and public payers, and captured at point-of-care
  - EHRs from both inpatient and outpatient settings
  - Also used to support registries and longitudinal surveys
- One model to support collaborative research across data sources both within and outside of US
- One model that can be manageable for data owners and useful for data users (efficient to put data IN and get data OUT)
- Enable standardization of structure, content, and analytics focused on specific use cases

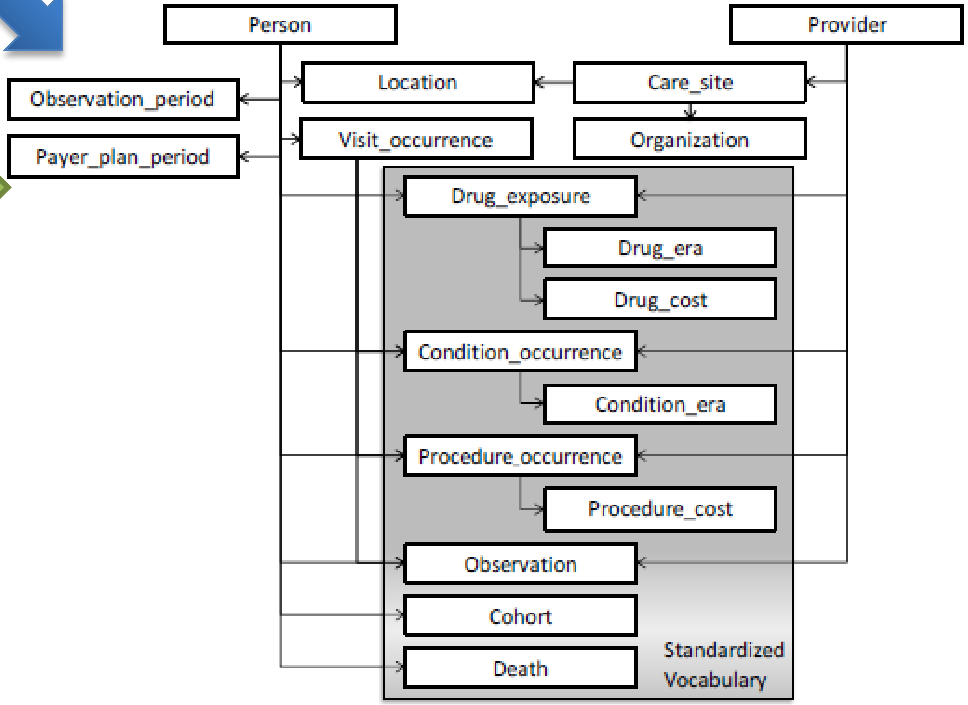
Evolution of the OMOP Common data model

OMOP CDM now Version 5, following multiple iterations of implementation, testing, modifications, and expansion based on the experiences of the OMOP community who bring on a growing landscape of research use cases.

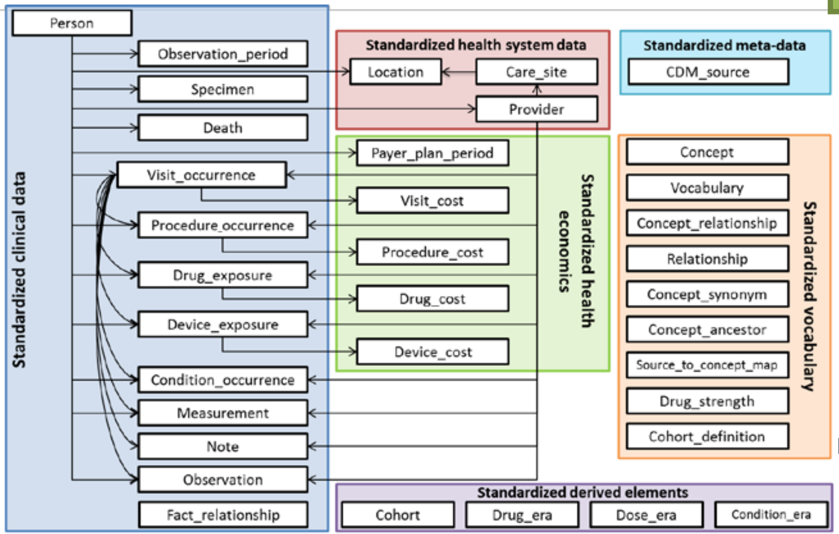
OMOP CDMv2



OMOP CDMv4

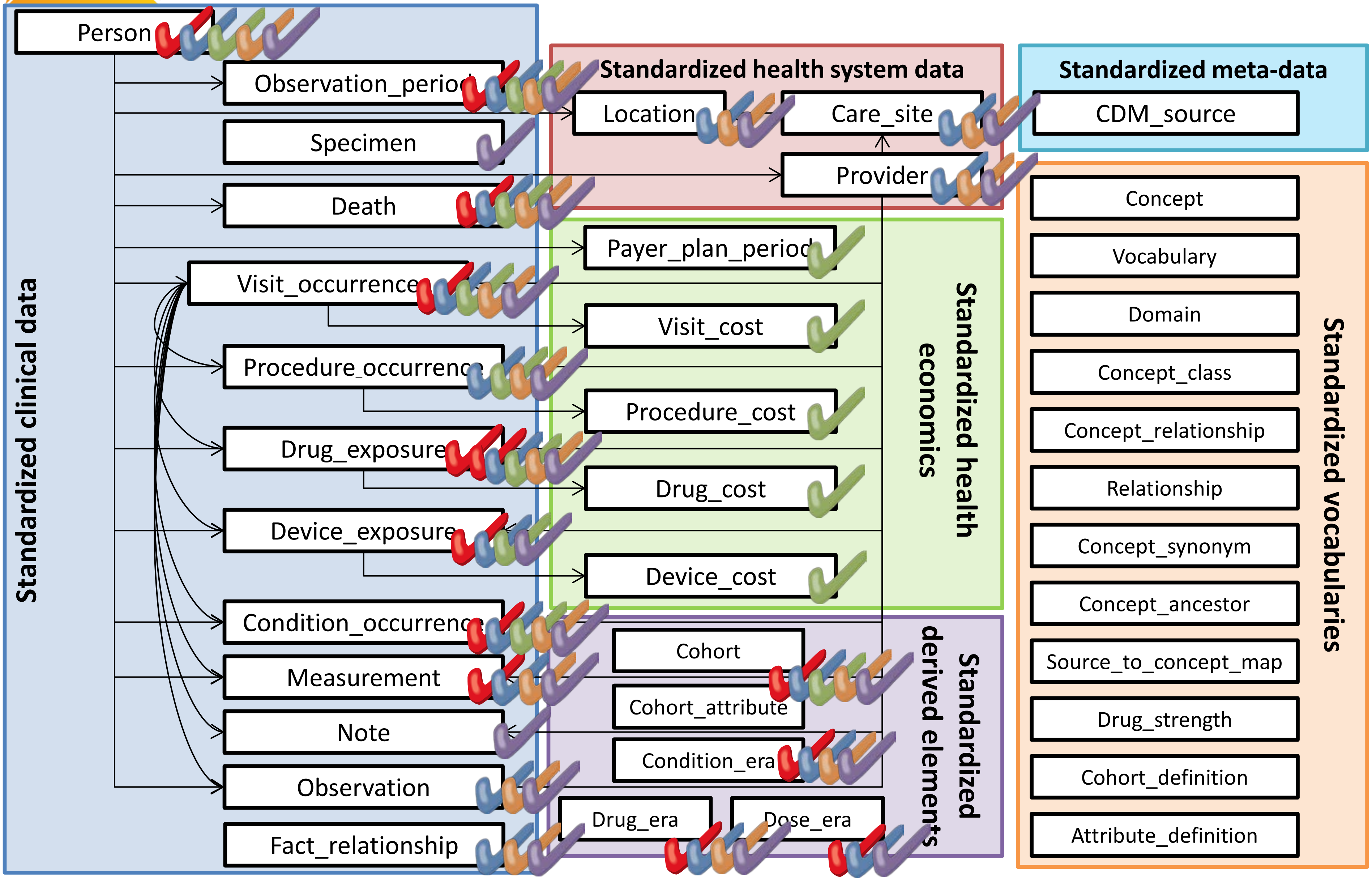


OMOP CDMv5



http://omop.org/CDM

One model, multiple use cases



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Visit_cost

Procedure_cost

Drug_cost

Device_cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

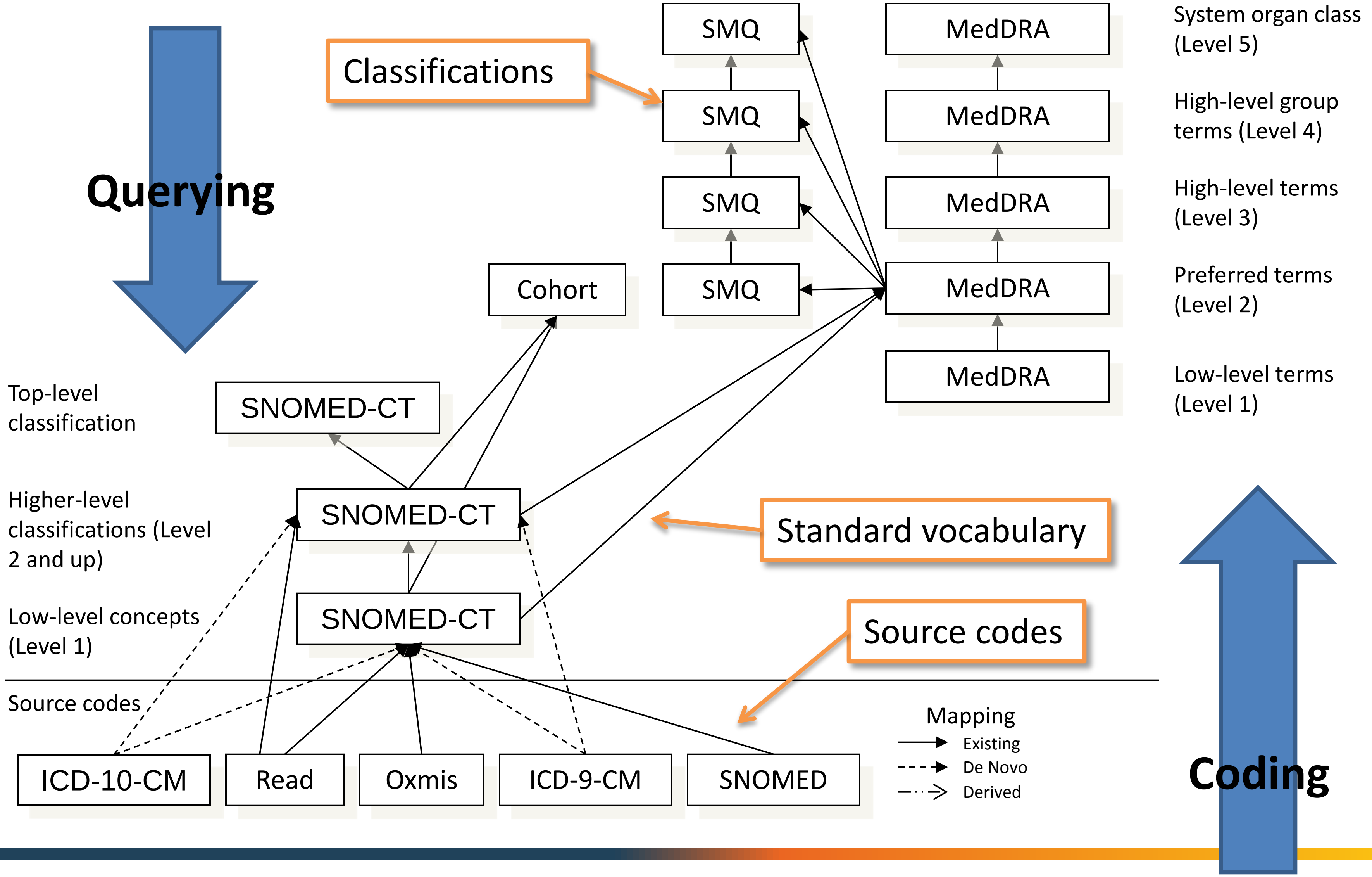
Source_to_concept_map

Drug_strength

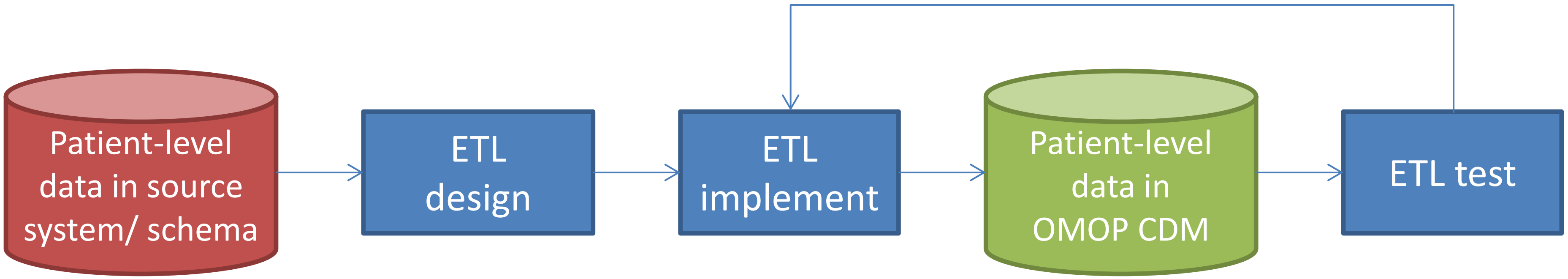
Cohort_definition

Attribute_definition

Standardized Vocabularies: Conditions



Preparing your data for analysis


Patient-level data in source system/ schema
ETL design
ETL implement
Patient-level data in OMOP CDM
ETL test

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

http://github.com/OHDSI

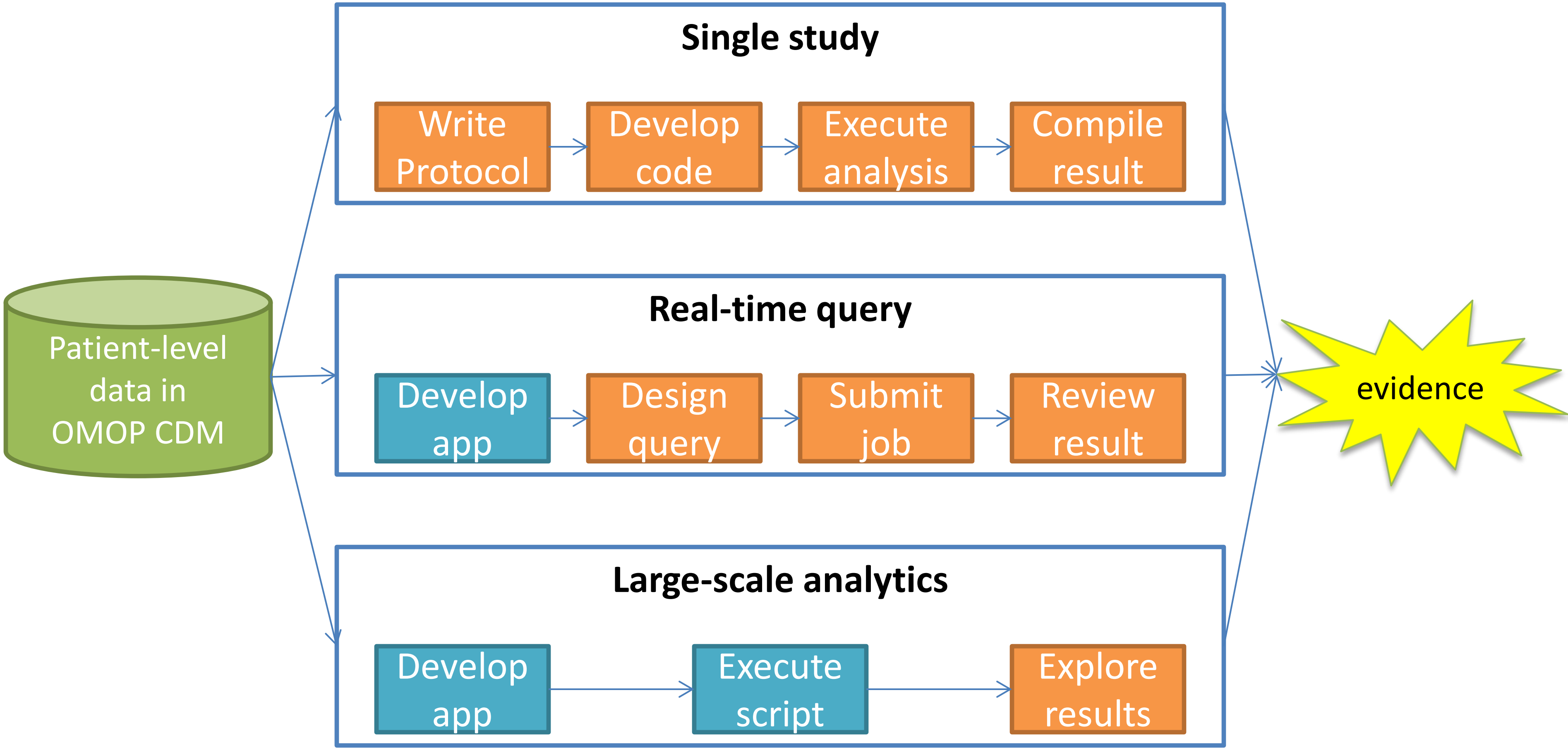
The odyssey to evidence generation

Patient-level data in source system/ schema

The Journey of Odysseus
Land of the Dead
(Tiresias)
8
7
Aeaea
(Circe)
9
6
Laestrygonians
(cannibals)
ITALY
GREECE
Ismarus
(Cicones)
2
Dardanelles
Troy
1
ASIA MINOR
BLACK SEA
Olympus
15
Phaeacia
Ithaca
16
10
Sirens
Charybdis
11
Aeolus
(god of the winds)
5
SICILY
Cyclops
12
Scylla
4
13
Ogygia Isle
(Calypso)
14
Thrinacia
(island of the sun god)
3
Lotus-Eaters
MEDITERRANEAN SEA

evidence

~~Data~~ Evidence sharing paradigms



Patient-level data in OMOP CDM

Single study

Write Protocol → Develop code → Execute analysis → Compile result

Real-time query

Develop app → Design query → Submit job → Review result

Large-scale analytics

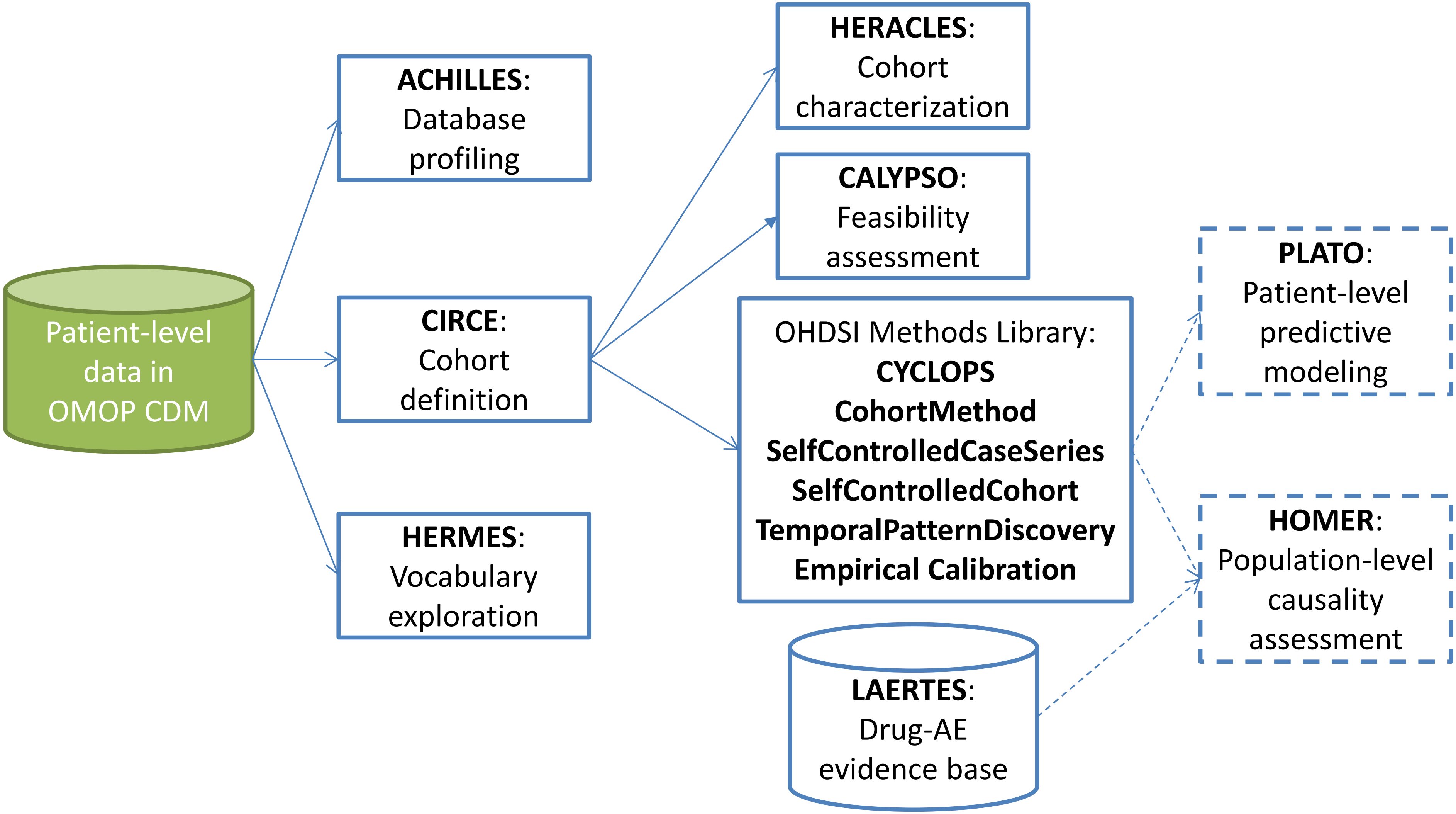
Develop app → Execute script → Explore results

evidence

One-time

Repeated

Standardized large-scale analytics tools under development within OHDSI



Patient-level data in OMOP CDM

ACHILLES: Database profiling

CIRCE: Cohort definition

HERMES: Vocabulary exploration

HERACLES: Cohort characterization

CALYPSO: Feasibility assessment

OHDSI Methods Library:
CYCLOPS
CohortMethod
SelfControlledCaseSeries
SelfControlledCohort
TemporalPatternDiscovery
Empirical Calibration

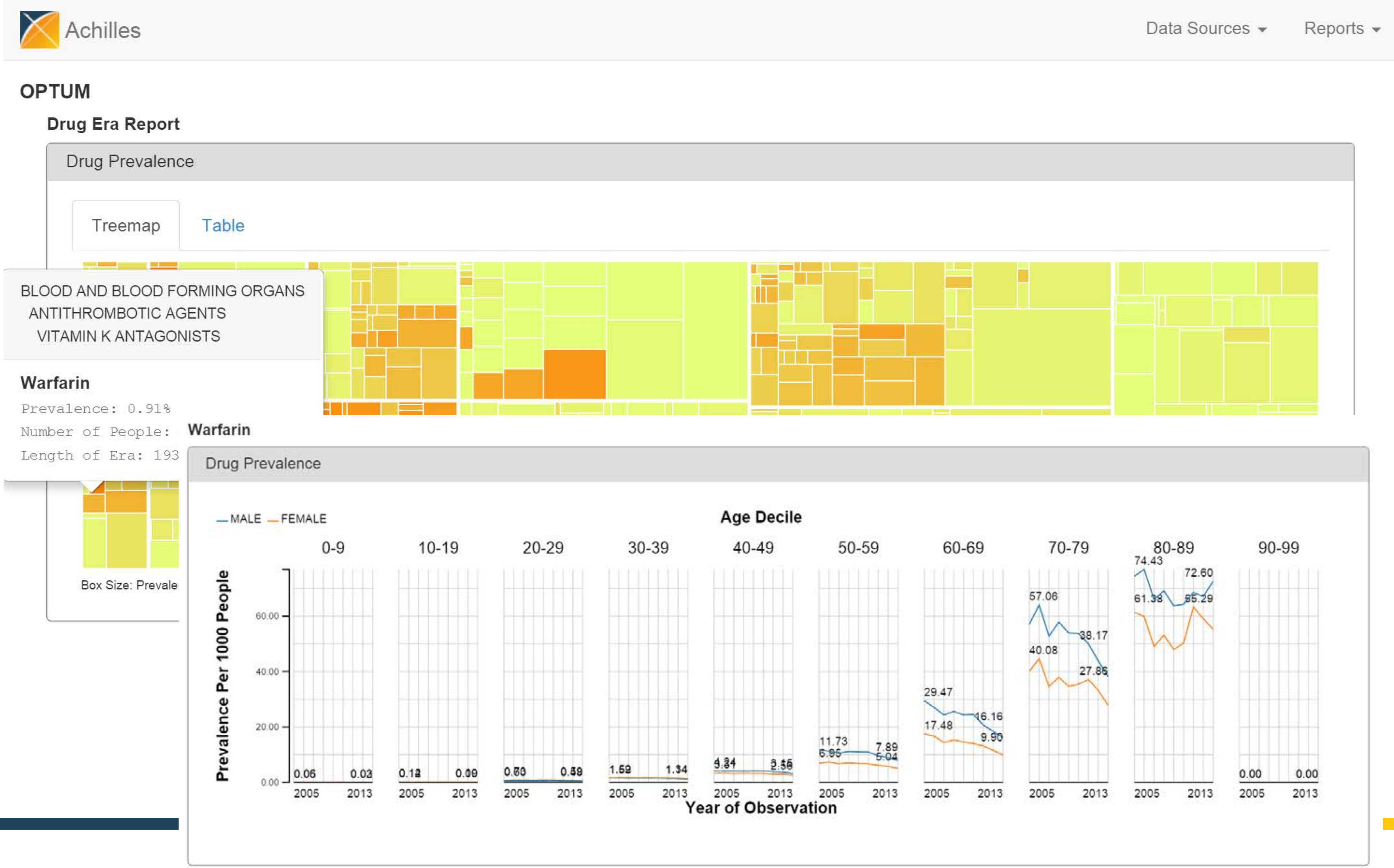
LAERTES: Drug-AE evidence base

PLATO: Patient-level predictive modeling

HOMER: Population-level causality assessment

http://github.com/OHDSI

ACHILLES: Database characterization to examine if the data have elements required for the analysis



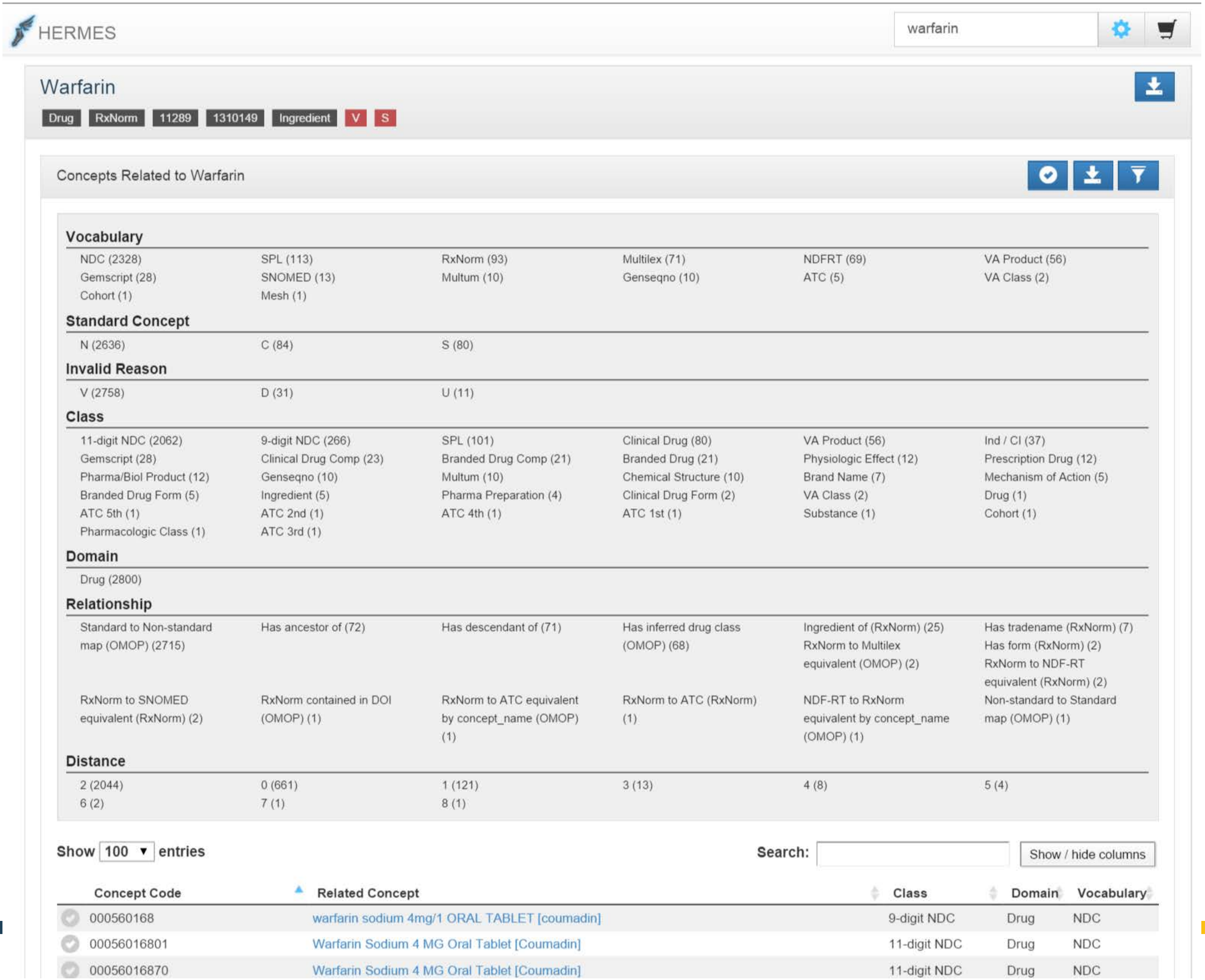
Warfarin

Prevalence: 0.91%

Number of People:

Length of Era: 193

HERMES: Explore the standardized vocabularies to define exposures, outcomes, and covariates



Warfarin

Drug | RxNorm | 11289 | 1310149 | Ingredient | V | S

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 tradename (RxNorm) (7)
RxNorm to Multilex equivalent (OMOP) (2) · Has form (RxNorm) (2) · RxNorm to NDF-RT equivalent (RxNorm) (2)
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) · NDF-RT to RxNorm equivalent by concept_name (OMOP) (1) · 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)

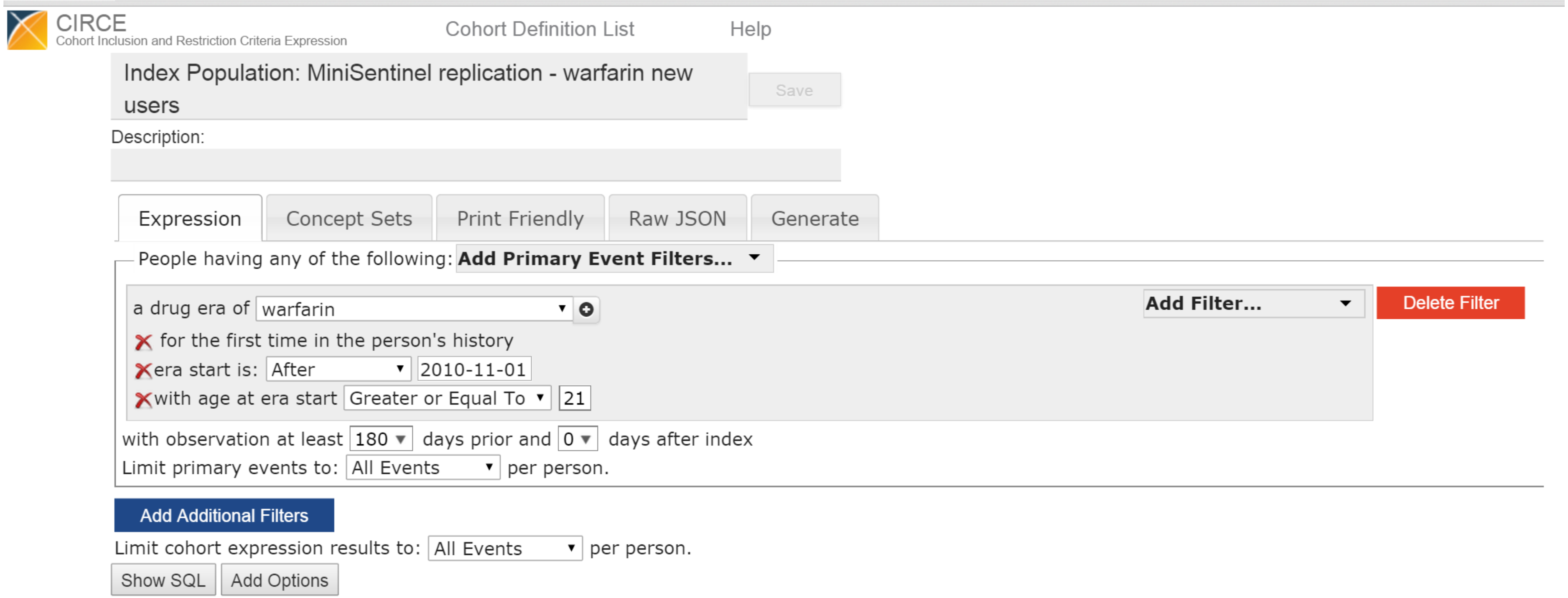
Show 100 entries

Search:

Show / hide columns

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

CIRCE: Define cohorts of interest



CIRCE
Cohort Inclusion and Restriction Criteria Expression

Index Population: MiniSentinel replication - warfarin new users

Description:

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

a drug era of warfarin

✗ for the first time in the person's history

✗ era start is: After 2010-11-01

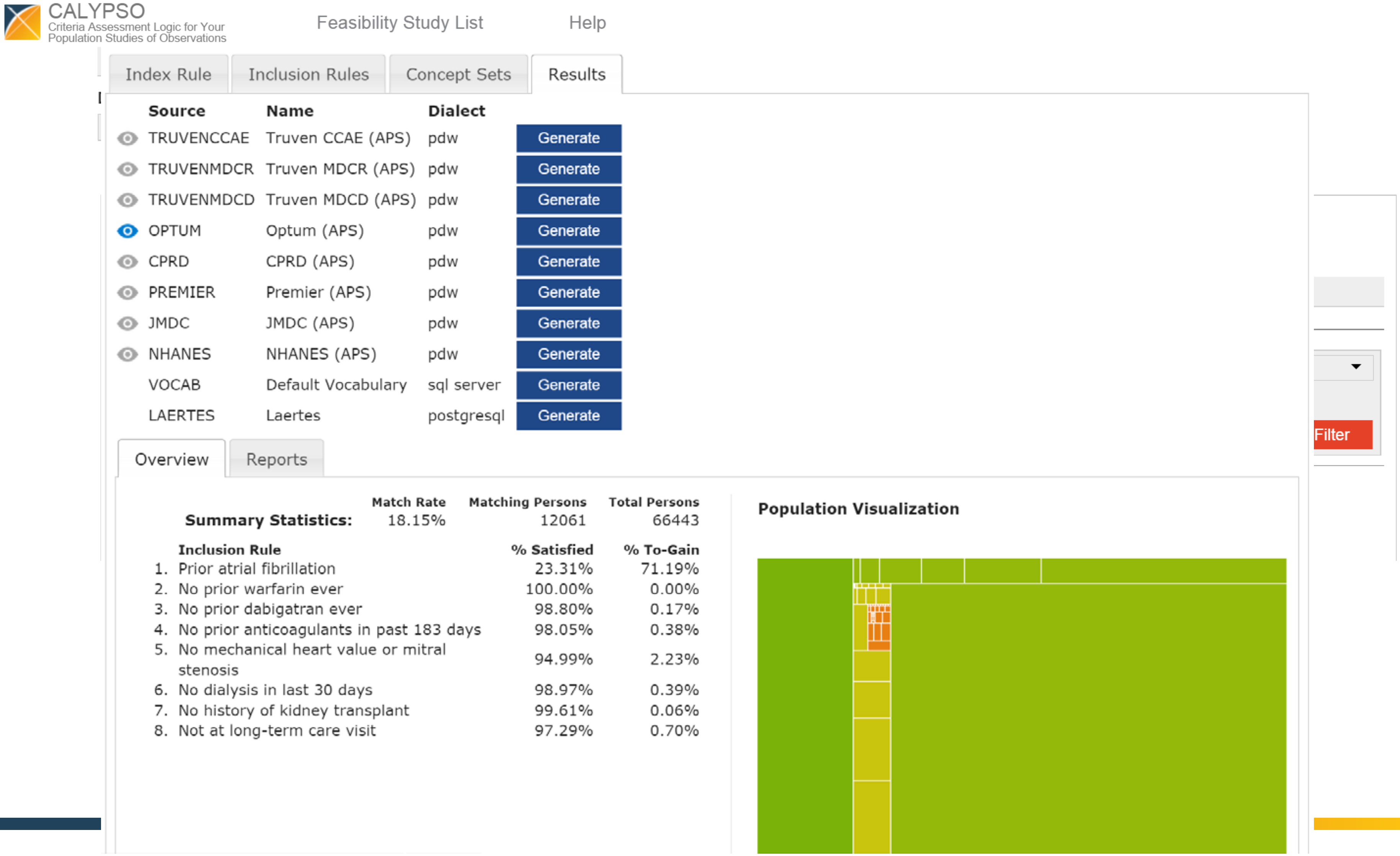
✗ with age at era start Greater or Equal To 21

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

Limit primary events to: All Events per person.

Limit cohort expression results to: All Events per person.

CALYPSO: Conduct feasibility assessment to evaluate the impact of study inclusion criteria



CALYPSO
Criteria Assessment Logic for Your Population Studies of Observations

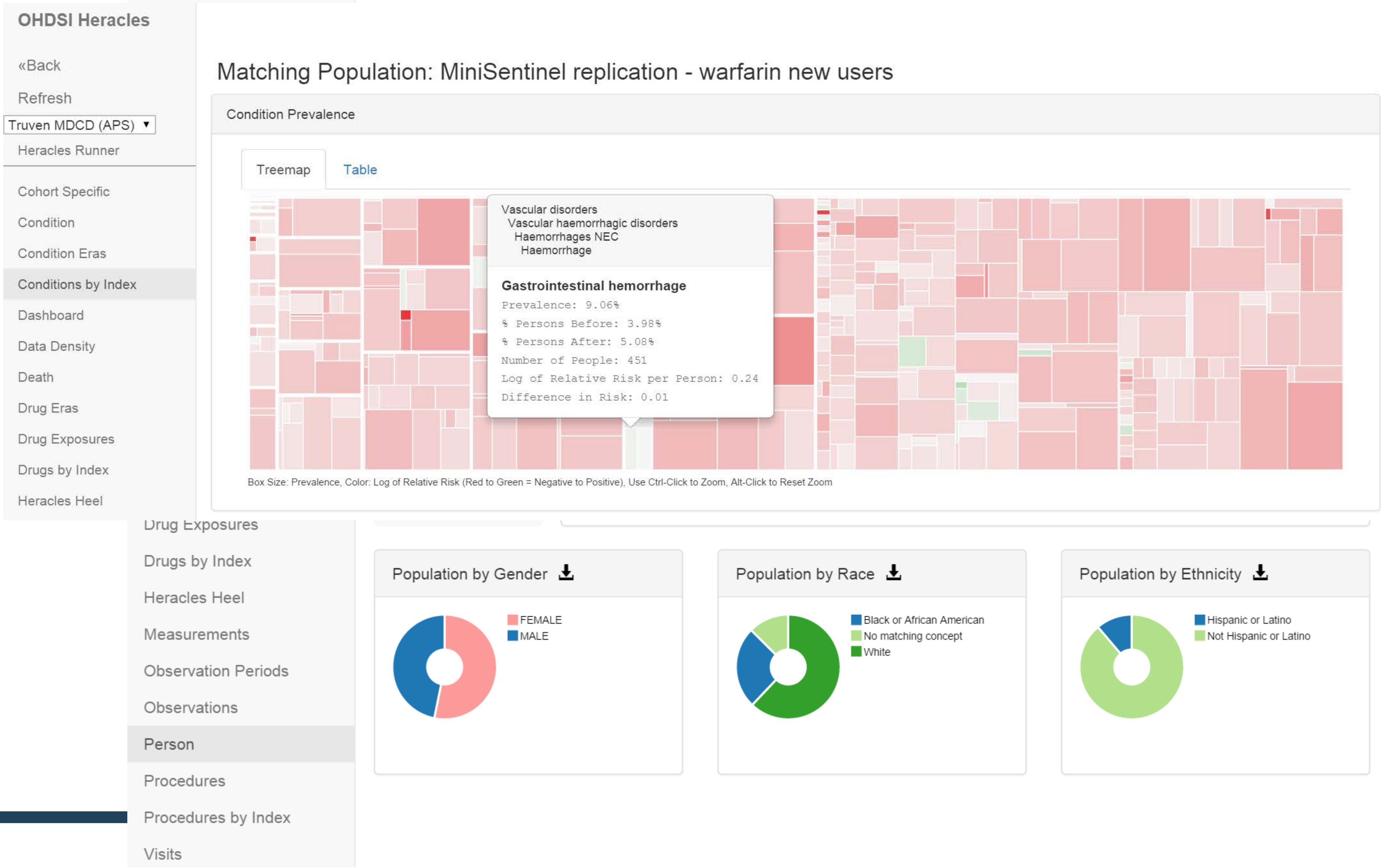
Source	Name	Dialect	
TRUVENCCAE	Truven CCAE (APS)	pdw	Generate
TRUVENMDCR	Truven MDCR (APS)	pdw	Generate
TRUVENMDCD	Truven MDCD (APS)	pdw	Generate
OPTUM	Optum (APS)	pdw	Generate
CPRD	CPRD (APS)	pdw	Generate
PREMIER	Premier (APS)	pdw	Generate
JMDC	JMDC (APS)	pdw	Generate
NHANES	NHANES (APS)	pdw	Generate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

	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

HERACLES: Characterize the cohorts of interest



Matching Population: MiniSentinel replication - warfarin new users

Condition Prevalence

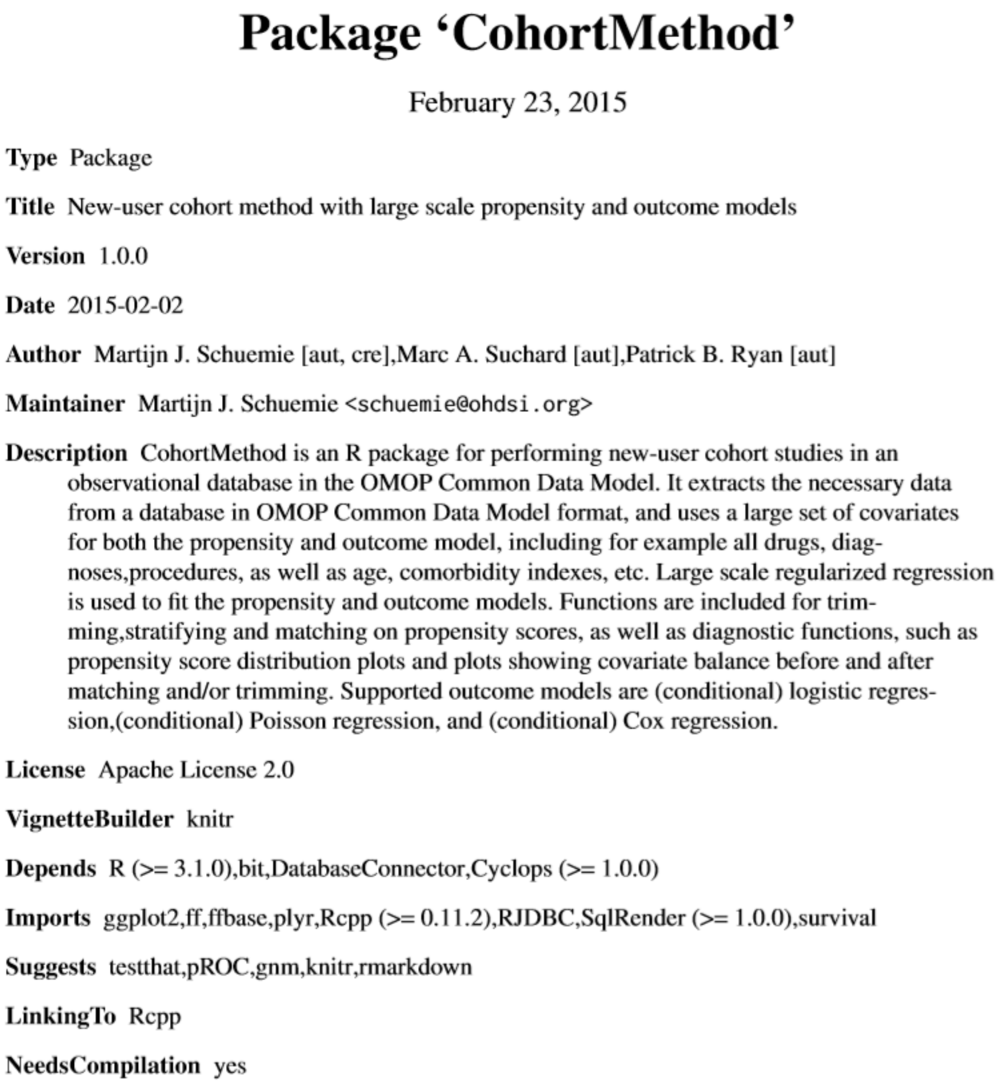
```
Vascular disorders
  Vascular haemorrhagic disorders
    Haemorrhages NEC
      Haemorrhage
```

Gastrointestinal hemorrhage

```
Prevalence: 9.06%
% Persons Before: 3.98%
% Persons After: 5.08%
Number of People: 451
Log of Relative Risk per Person: 0.24
Difference in Risk: 0.01
```

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

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

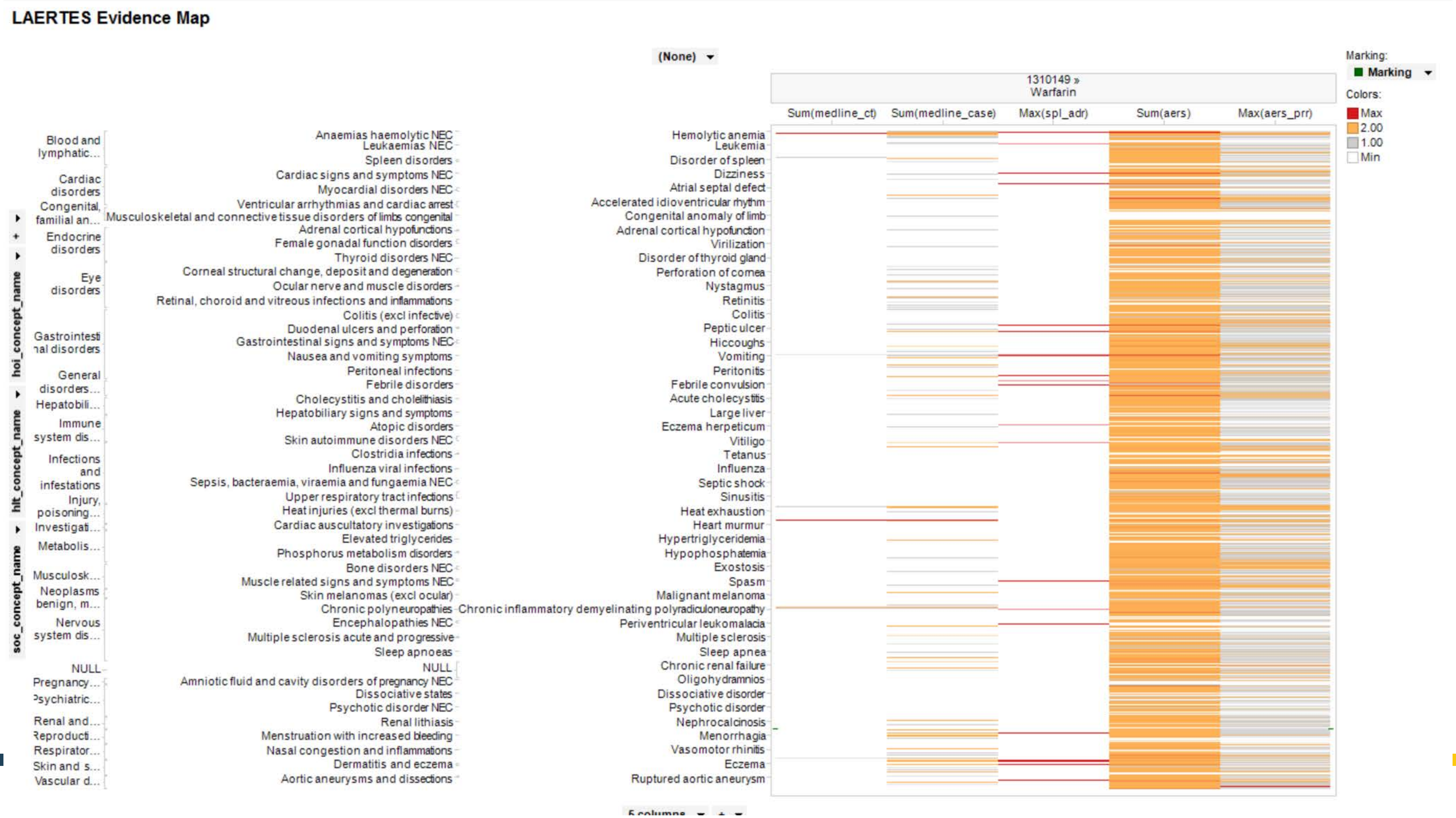
LinkingTo Rcpp

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’

LAERTES: Summarizing evidence from existing data sources: literature, labeling, spontaneous reporting



Steps to Standardized Data

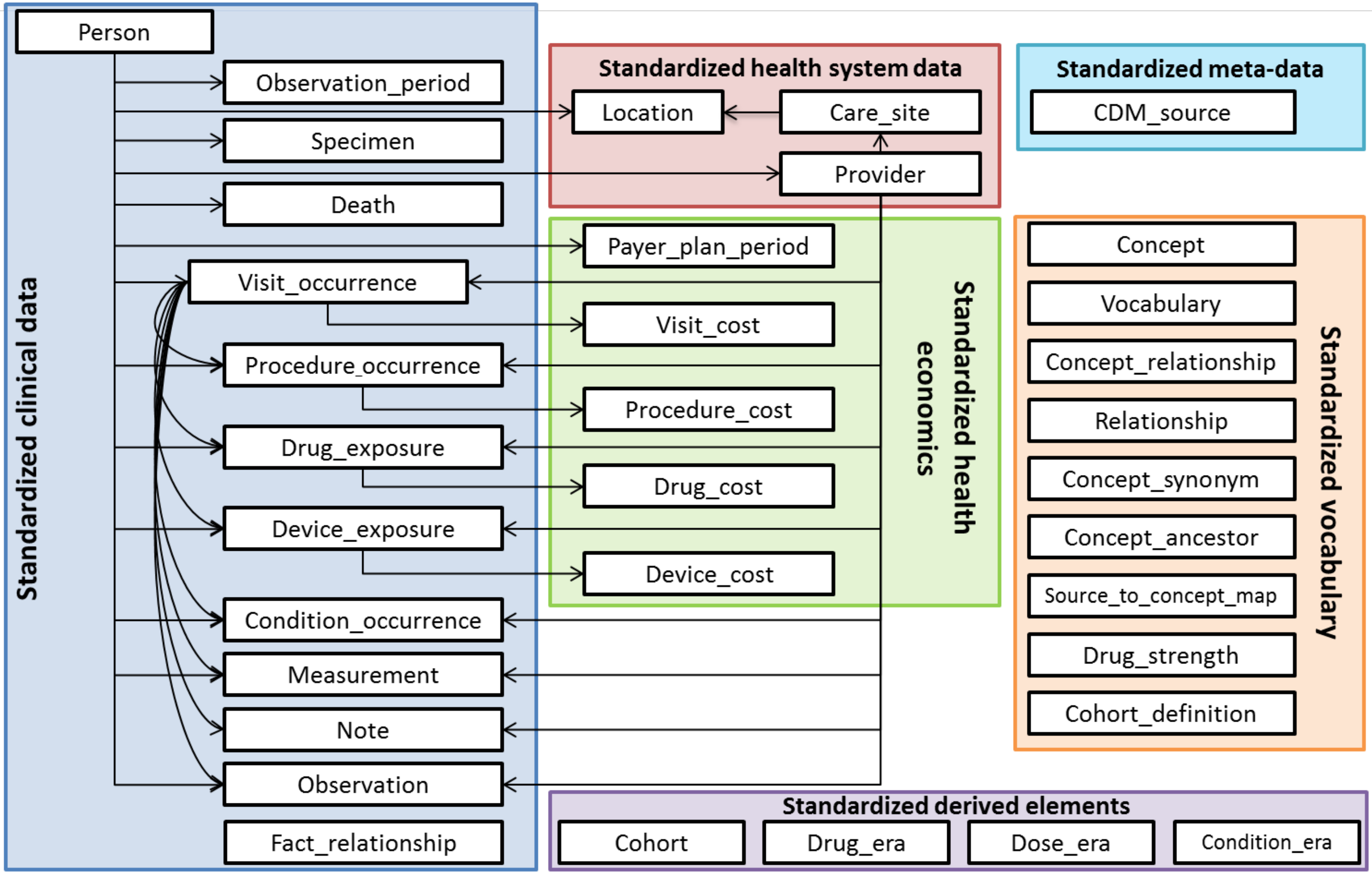
Getting Your Data into the OMOP CDM

- Everyone’s data starts messy!
- To get into a standardized model, you need
  - Someone familiar with the source dataset
  - Someone familiar with healthcare
  - Someone who can write SQL
- Fortunately, OHDSI has great tools (and people!) to help you out

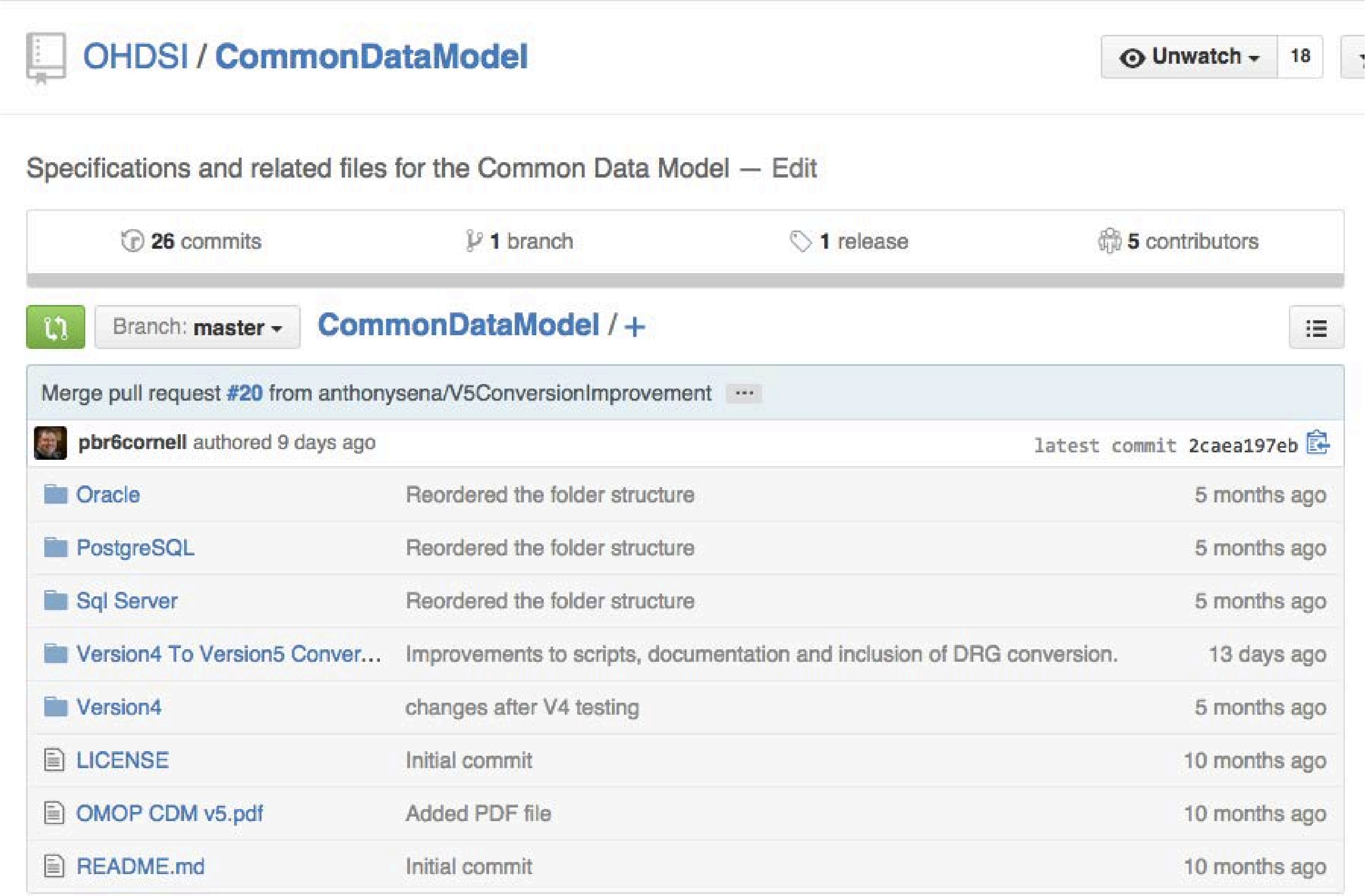
Interactive Example

- The U.S. Centers for Medicare and Medicaid Services (CMS) releases a variety of public data sets
- For this example, we will use ‘SynPUF’, a synthetic claims dataset based on real patient data
- We will cover the steps of mapping this over to OMOP CDM V5

OMOP CDM V5


Standardized clinical data
Person
Observation_period
Specimen
Death
Visit_occurrence
Procedure_occurrence
Drug_exposure
Device_exposure
Condition_occurrence
Measurement
Note
Observation
Fact_relationship
Standardized health system data
Location
Care_site
Provider
Standardized health economics
Payer_plan_period
Visit_cost
Procedure_cost
Drug_cost
Device_cost
Standardized meta-data
CDM_source
Standardized vocabulary
Concept
Vocabulary
Concept_relationship
Relationship
Concept_synonym
Concept_ancestor
Source_to_concept_map
Drug_strength
Cohort_definition
Standardized derived elements
Cohort
Drug_era
Dose_era
Condition_era

Where to find the CDM?



OHDSI / **CommonDataModel**

Specifications and related files for the Common Data Model — Edit

CommonDataModel / +

Merge pull request #20 from anthonysena/V5ConversionImprovement ···

pbr6cornell authored 9 days ago latest commit 2caea197eb

Oracle	Reordered the folder structure	5 months ago
PostgreSQL	Reordered the folder structure	5 months ago
Sql Server	Reordered the folder structure	5 months ago
Version4 To Version5 Conver…	Improvements to scripts, documentation and inclusion of DRG conversion.	13 days ago
Version4	changes after V4 testing	5 months ago
LICENSE	Initial commit	10 months ago
OMOP CDM v5.pdf	Added PDF file	10 months ago
README.md	Initial commit	10 months ago

Our Source Data

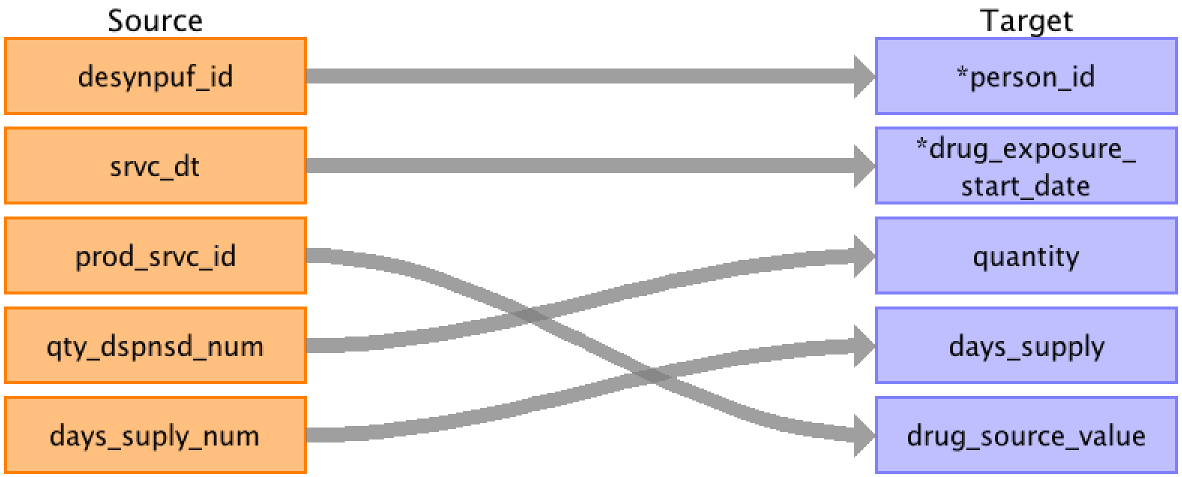
- Synthetic Public Use Files
  - Beneficiary Summary
  - Carrier claims
  - Inpatient claims
  - Outpatient claims
  - Prescription drug events
- CSV format

Step 1: What is in your dataset? **WhiteRabbit**

- WhiteRabbit, a tool that lets you
  - Scans your dataset
  - Extracts summary information on the contents
  - Produces a file that can be consumed for ETL planning

Step 2: Map Your Dataset to CDM
Rabbit In a Hat

- Rabbit-In-a-Hat is a tool that uses the WhiteRabbit output and lets you match up your dataset with the CDM model



Source

desynpuf_id

srvc_dt

prod_srvc_id

qty_dspnsd_num

days_suply_num

Target

*person_id

*drug_exposure_start_date

quantity

days_supply

drug_source_value

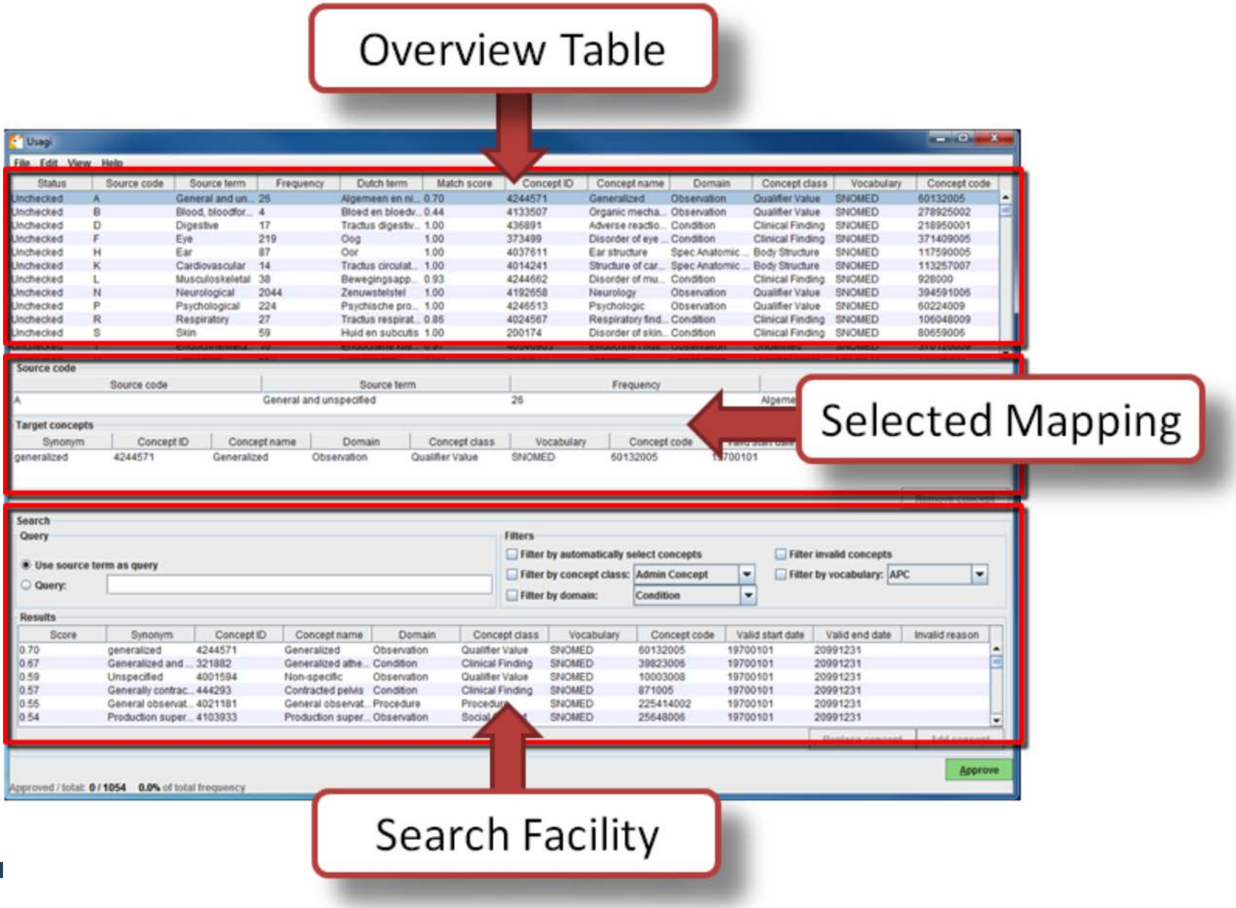
OHDSI Has Extensive Vocabulary Maps

Athena

1	SNOMED	Systematic Nomenclature of Medicine - Clinical Terms (IHDSTO)
2	ICD9CM	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 1 and 2 (NCHS)
3	ICD9Proc	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 3 (NCHS)
4	CPT4	Current Procedural Terminology version 4 (AMA)
5	HCPCS	Healthcare Common Procedure Coding System (CMS)
6	LOINC	Logical Observation Identifiers Names and Codes (Regenstrief Institute)
7	NDFRT	National Drug File - Reference Terminology (VA)
8	RxNorm	RxNorm (NLM)
9	NDC	National Drug Code (FDA and manufacturers)
10	GPI	Medi-Span Generic Product Identifier (Wolters Kluwer Health)
11	UCUM	Unified Code for Units of Measure (Regenstrief Institute)
12	Gender	OMOP Gender
13	Race	Race and Ethnicity Code Set (USBC)
14	Place of Service	Place of Service Codes for Professional Claims (CMS)
15	MedDRA	Medical Dictionary for Regulatory Activities (MSSO)
16	Multum	Cerner Multum (Cerner)
17	Read	NHS UK Read Codes Version 2 (HSCIC)
18	OXMIS	Oxford Medical Information System (OCHP)
19	Indication	Indications and Contraindications (FDB)
20	ETC	Enhanced Therapeutic Classification (FDB)
21	ATC	WHO Anatomic Therapeutic Chemical Classification
22	Multilex	Multilex (FDB)
28	VA Product	VA National Drug File Product (VA)
31	SMQ	Standardised MedDRA Queries (MSSO)
32	VA Class	VA National Drug File Class (VA)
33	Cohort	Legacy OMOP HOI or DOI cohort
34	ICD10	International Classification of Diseases, 10th Revision, (WHO)
35	ICD10PCS	ICD-10 Procedure Coding System (CMS)
40	DRG	Diagnosis-related group (CMS)
41	MDC	Major Diagnostic Categories (CMS)
42	APC	Ambulatory Payment Classification (CMS)
43	Revenue Code	UB04/CMS1450 Revenue Codes (CMS)
44	Ethnicity	OMOP Ethnicity
46	MeSH	Medical Subject Headings (NLM)
47	NUCC	National Uniform Claim Committee Health Care Provider Taxonomy Code Set (NUCC)
48	Specialty	Medicare provider/supplier specialty codes (CMS)
50	SPL	Structured Product Labeling (FDA)
53	Genseqno	Generic sequence number (FDB)
54	CCS	Clinical Classifications Software for ICD-9-CM (HCUP)
55	OPCS4	OPCS Classification of Interventions and Procedures version 4 (NHS)
56	Gemscript	Gemscript NHS dictionary of medicine and devices (NHS)
57	HES Specialty	Hospital Episode Statistics Specialty (NHS)
60	PCORNet	National Patient-Centered Clinical Research Network (PCORI)
65	Currency	International Currency Symbol (ISO 4217)
70	ICD10CM	International Classification of Diseases, 10th Revision, Clinical Modification (NCHS)
72	CIEL	Columbia International eHealth Laboratory (Columbia University)

Additional Vocabulary Support

- If you use non-standard vocabularies, you can also utilize our vocabulary mapper tool **Usagi**



Overview Table

Selected Mapping

Search Facility

Status	Source code	Source term	Frequency	Dutch term	Match score	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
Unchecked	A	General and un...	26	Algemeen en ni...	0.70	4244571	Generalized	Observation	Qualifier Value	SNOMED	60132005
Unchecked	B	Blood, bloodfor...	4	Bloed en bloedv...	0.44	4133507	Organic mecha...	Observation	Qualifier Value	SNOMED	278925002
Unchecked	D	Digestive	17	Tractus digestiv...	1.00	436891	Adverse reactio...	Condition	Clinical Finding	SNOMED	218950001
Unchecked	F	Eye	219	Oog	1.00	373499	Disorder of eye ...	Condition	Clinical Finding	SNOMED	371409005
Unchecked	H	Ear	87	Oor	1.00	4037611	Ear structure	Spec Anatomic ...	Body Structure	SNOMED	117590005
Unchecked	K	Cardiovascular	14	Tractus circulat...	1.00	4014241	Structure of car...	Spec Anatomic ...	Body Structure	SNOMED	113257007
Unchecked	L	Musculoskeletal	38	Bewegingsapp...	0.93	4244662	Disorder of mu...	Condition	Clinical Finding	SNOMED	928000
Unchecked	N	Neurological	2044	Zenuwstelstel	1.00	4192658	Neurology	Observation	Qualifier Value	SNOMED	394591006
Unchecked	P	Psychological	224	Psychische pro...	1.00	4246513	Psychologic	Observation	Qualifier Value	SNOMED	60224009
Unchecked	R	Respiratory	27	Tractus respirat...	0.86	4024567	Respiratory find...	Condition	Clinical Finding	SNOMED	106048009
Unchecked	S	Skin	59	Huid en subcutis	1.00	200174	Disorder of skin...	Condition	Clinical Finding	SNOMED	80659006

Source code

Source code	Source term	Frequency	
A	General and unspecified	26	Algeme

Target concepts

Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date
generalized	4244571	Generalized	Observation	Qualifier Value	SNOMED	60132005	19700101

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by automatically select concepts

□ Filter by concept class: Admin Concept

□ Filter by domain: Condition

□ Filter invalid concepts

□ Filter by vocabulary: APC

Results

Score	Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
0.70	generalized	4244571	Generalized	Observation	Qualifier Value	SNOMED	60132005	19700101	20991231	
0.67	Generalized and ...	321882	Generalized athe...	Condition	Clinical Finding	SNOMED	39823006	19700101	20991231	
0.59	Unspecified	4001594	Non-specific	Observation	Qualifier Value	SNOMED	10003008	19700101	20991231	
0.57	Generally contrac...	444293	Contracted pelvis	Condition	Clinical Finding	SNOMED	871005	19700101	20991231	
0.55	General observat...	4021181	General observat...	Procedure	Procedure	SNOMED	225414002	19700101	20991231	
0.54	Production super...	4103933	Production super...	Observation	Social	SNOMED	25648006	19700101	20991231	

Approved / total: **0 / 1054** **0.0%** of total frequency

Step 3: Turn the Crank

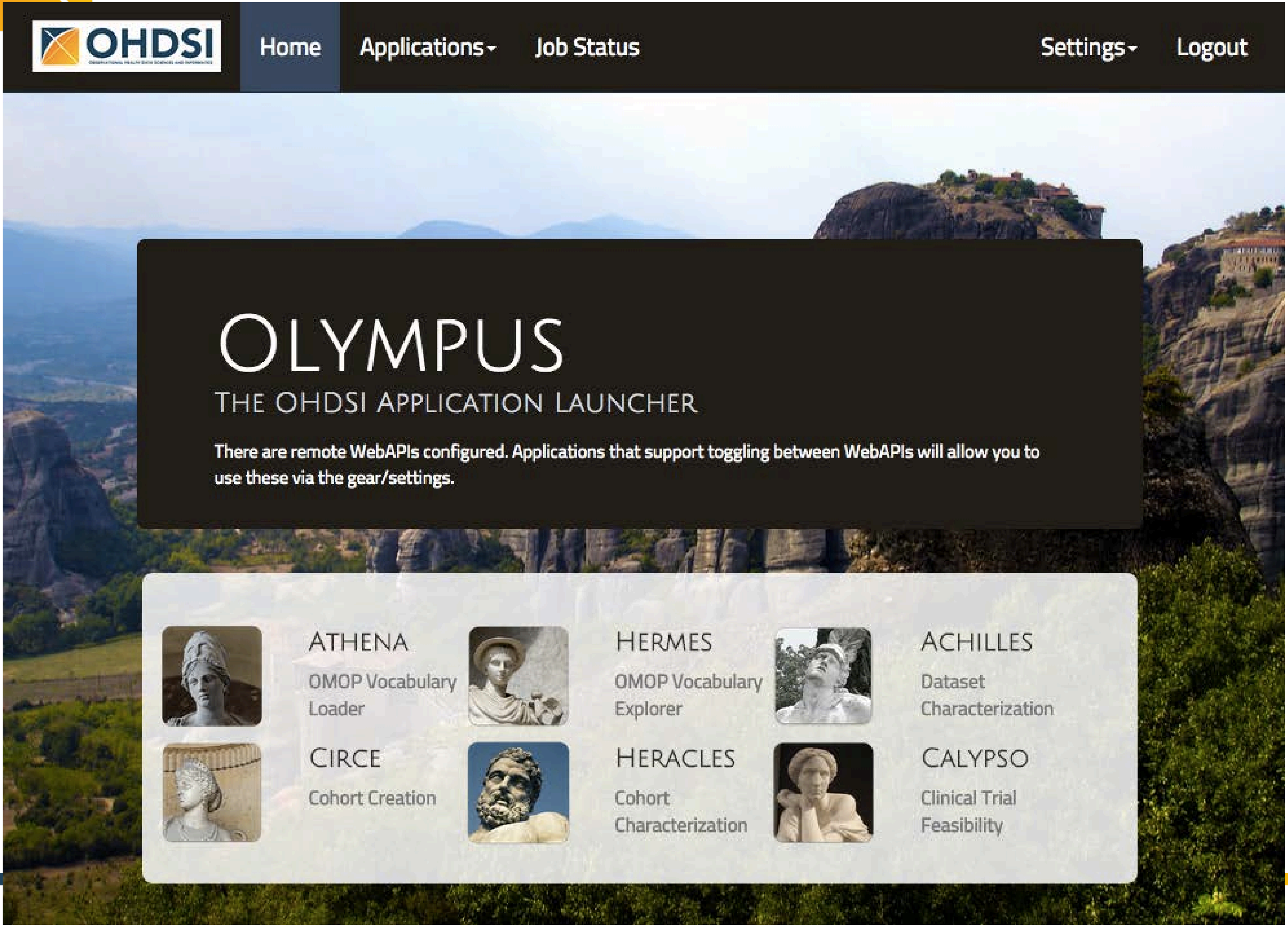
- Write the SQL using the generated ETL doc as you guide
- Get help on the forums from the many folks who have done it before
- We provide tools to explore and analyze your data and data quality as you go along so you can iterate as needed

Exploring Populations and Cohorts

Getting Value from Your Data

- Once your data has been transformed, the OHDSI platform opens up a variety of ways to explore it

The OHDSI Web Application Suite



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

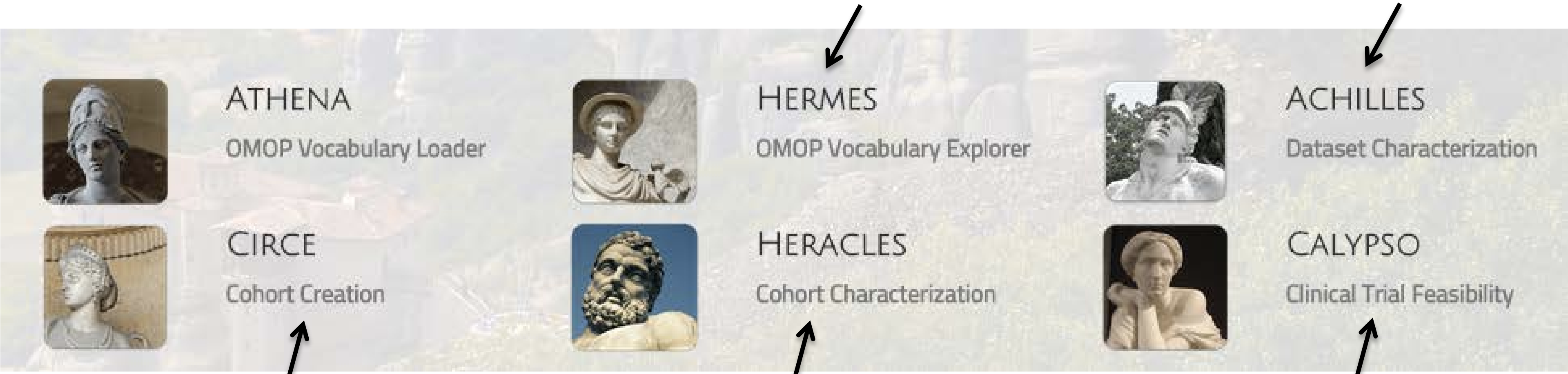
CALYPSO

Clinical Trial Feasibility

OHDSI Web Tools

HERMES:
Explore the OMOP
Vocabulary

ACHILLES:
Explore Population
Level Data



ATHENA
OMOP Vocabulary Loader

HERMES
OMOP Vocabulary Explorer

ACHILLES
Dataset Characterization

CIRCE
Cohort Creation

HERACLES
Cohort Characterization

CALYPSO
Clinical Trial Feasibility

CIRCE:
Define Patient
Cohorts

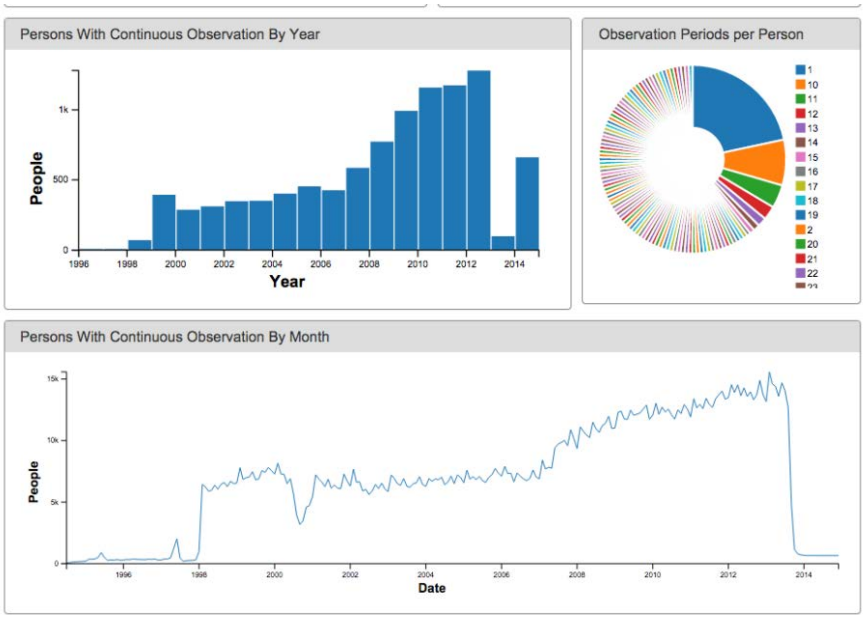
HERACLES:
Explore Cohort
Level Data

CALYPSO:
Explore Trial
Feasibility

Characterization in OHDSI

- In OHDSI, characterization = generating a comprehensive overview of a patient dataset
  - Clinical (e.g., conditions, medications, procedures)
  - Metadata (e.g., observation periods, data density)
- Supports
  - Feasibility studies
  - Hypothesis generation
  - Data quality assessment
  - Data sharing (aggregate-level)

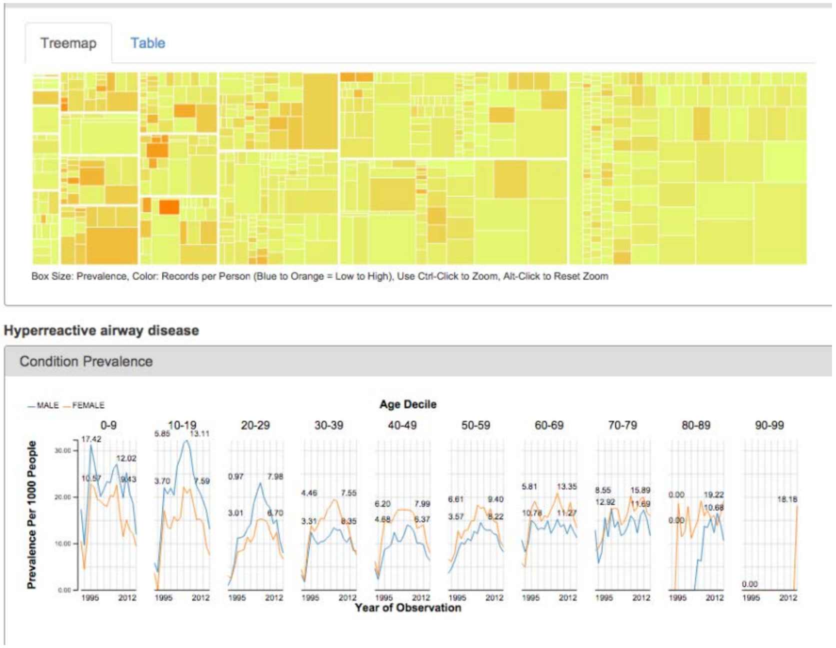
ACHILLES



Persons With Continuous Observation By Year

Observation Periods per Person

Persons With Continuous Observation By Month



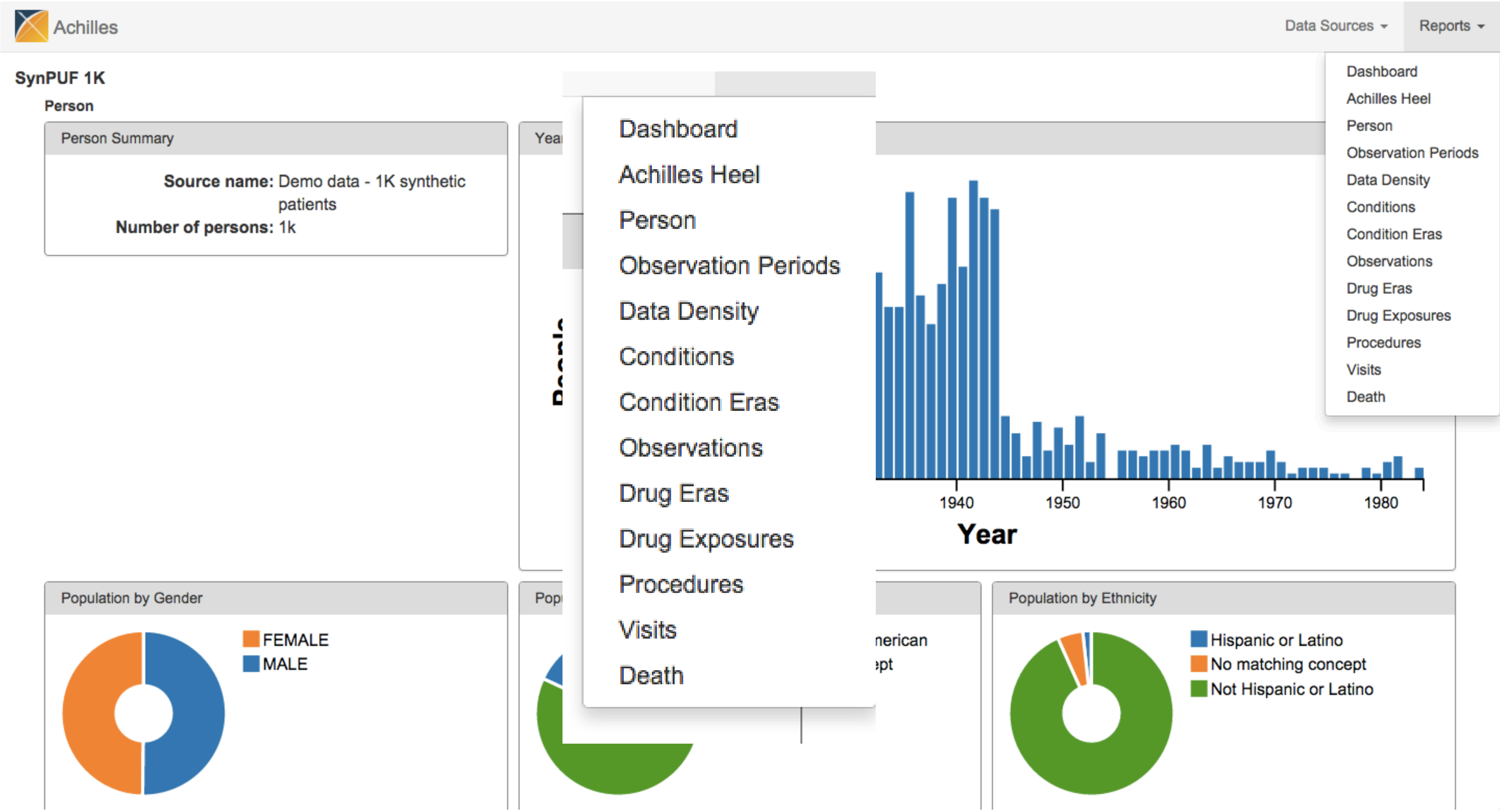
Treemap | Table

Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

Hyperreactive airway disease

Condition Prevalence

ACHILLES Report Types



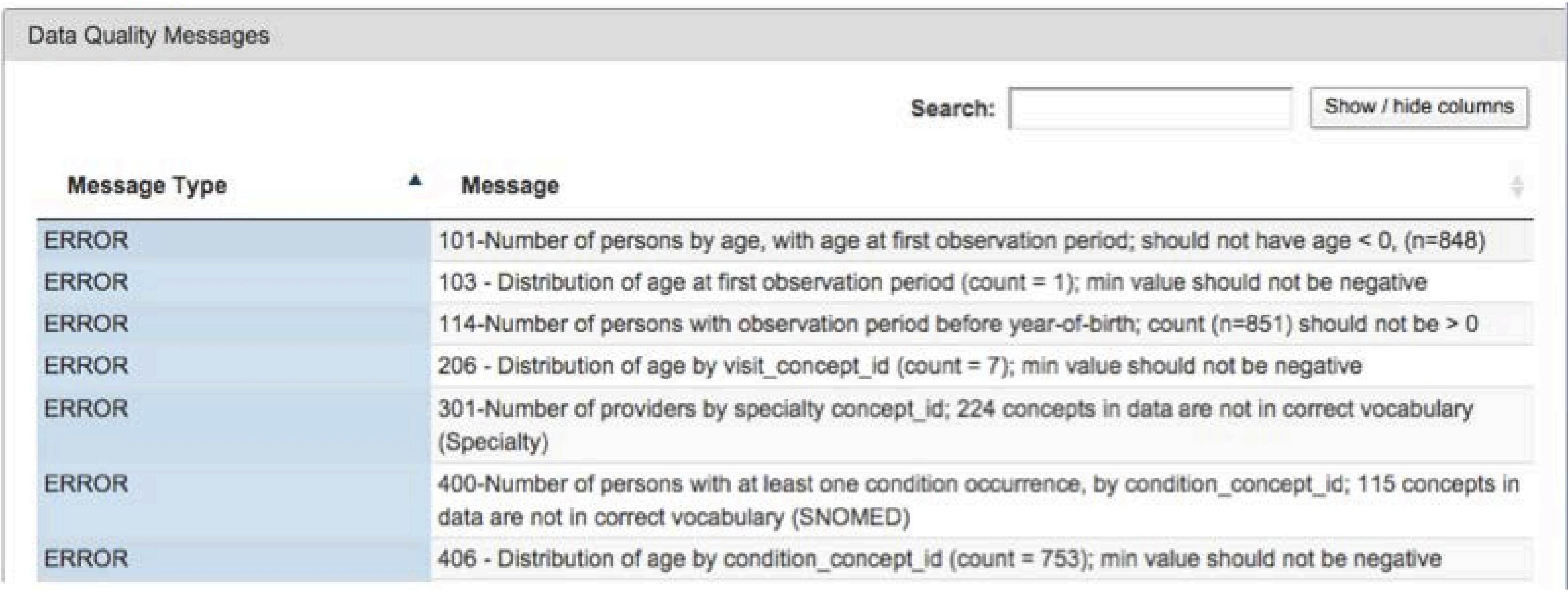
SynPUF 1K

Person

Person Summary

Source name: Demo data - 1K synthetic patients
Number of persons: 1k

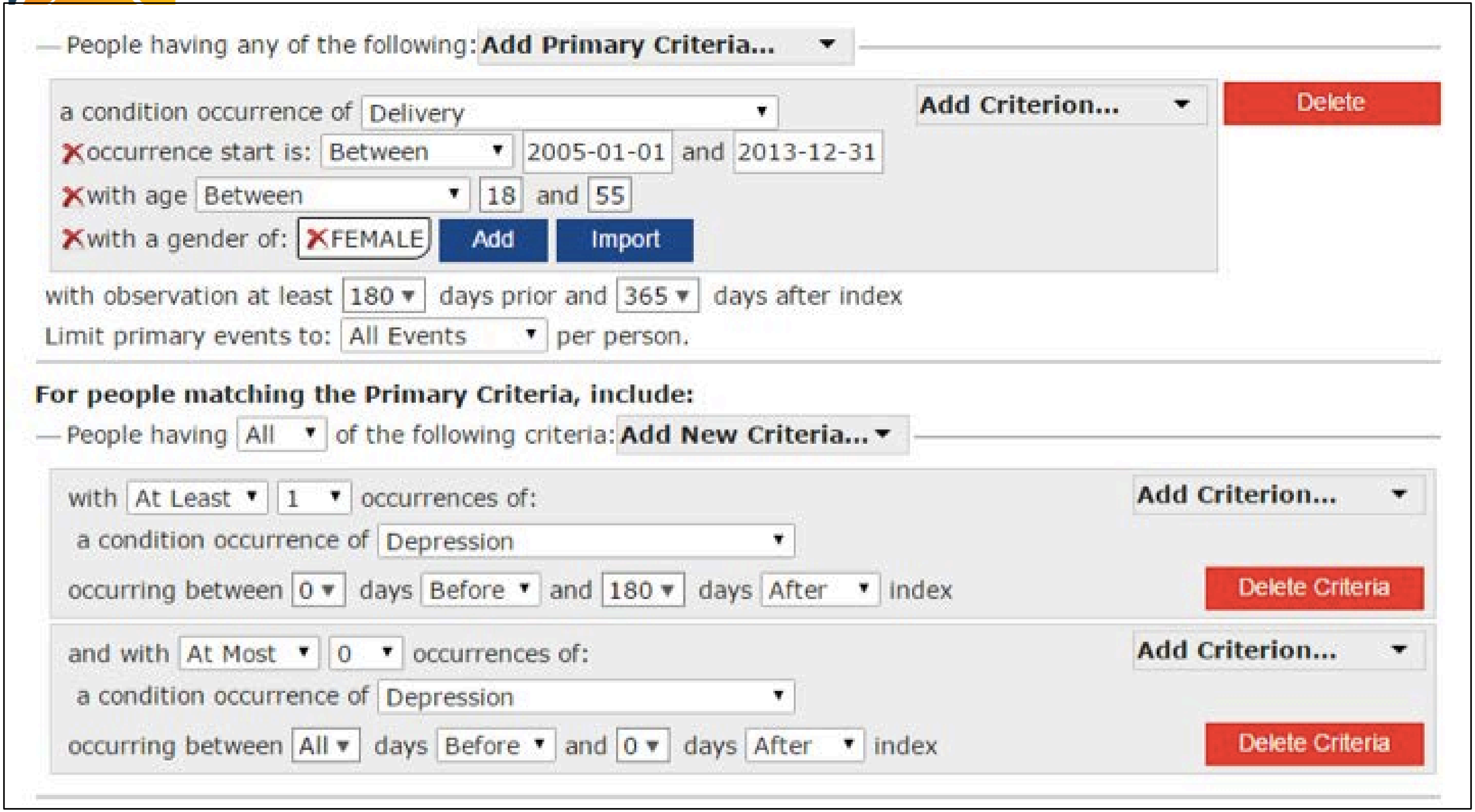
ACHILLES Heel Helps You Validate Your Data Quality



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

From Populations to Cohorts

- Once you’ve explored your overall dataset, designing cohorts allows you to analyze individual populations, conduct studies, explore trial feasibility, and so forth
- CIRCE provides a graphical interface for defining patient cohorts



People having any of the following: Add Primary Criteria...

a condition occurrence of Delivery

occurrence start is: Between 2005-01-01 and 2013-12-31

with age Between 18 and 55

with a gender of: FEMALE

with observation at least 180 days prior and 365 days after index

Limit primary events to: All Events per person.

For people matching the Primary Criteria, include:

People having All of the following criteria: Add New Criteria...

with At Least 1 occurrences of:

a condition occurrence of Depression

occurring between 0 days Before and 180 days After index

and with At Most 0 occurrences of:

a condition occurrence of Depression

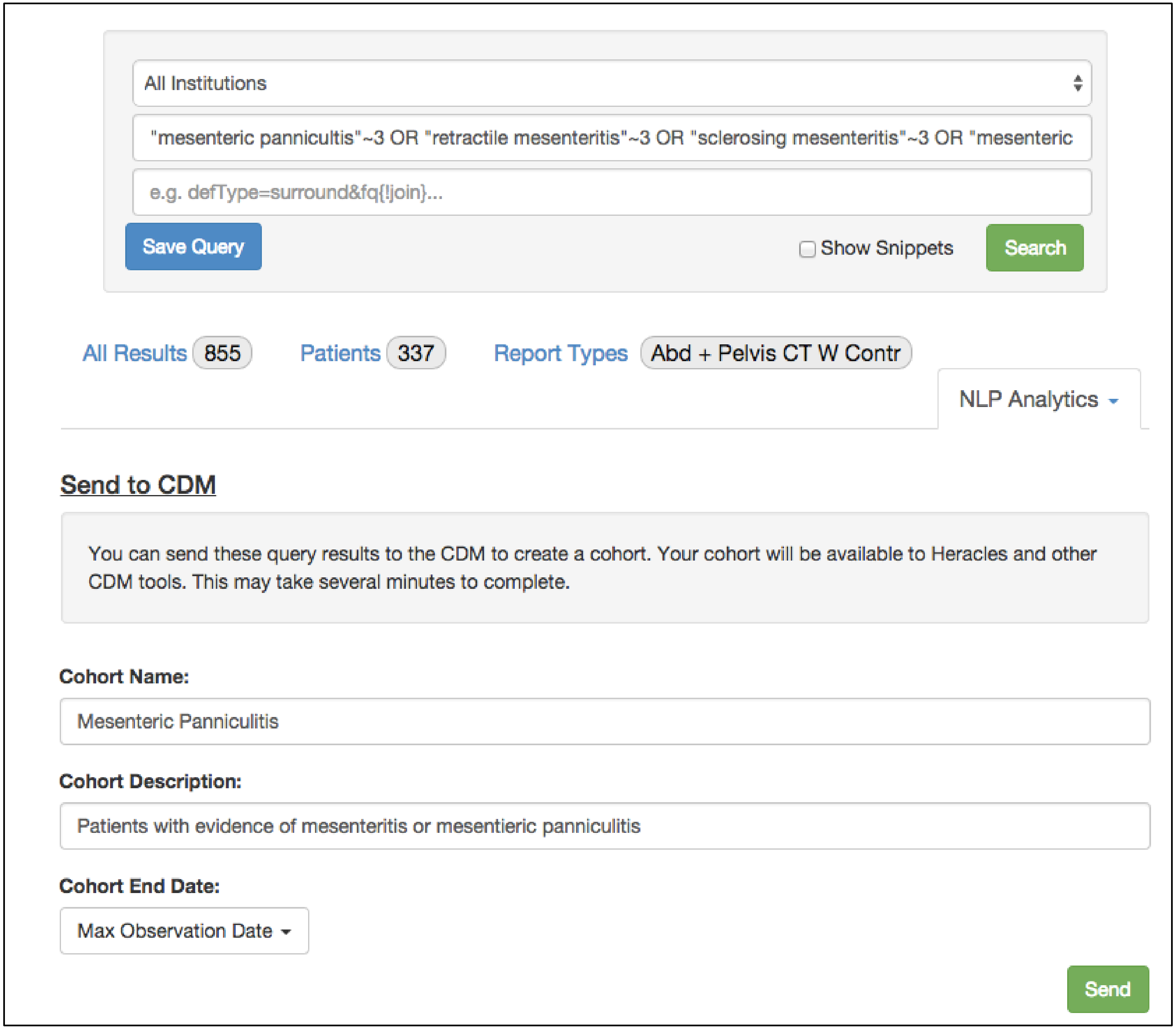
occurring between All days Before and 0 days After index

Building Cohorts

- When building cohorts, it is very helpful to reference ACHILLES data to see frequently used concepts
- This data-driven approach can similarly be achieved through the Hermes vocabulary explorer

Building Cohorts

- In addition to the graphical tools, cohorts can also be generated by manual SQL queries or imported from external sources



All Institutions

"mesenteric panniculitis"~3 OR "retractile mesenteritis"~3 OR "sclerosing mesenteritis"~3 OR "mesenteric

e.g. defType=surround&fq{!join}...

☐ Show Snippets

All Results 855

Patients 337

Report Types Abd + Pelvis CT W Contr

Send to CDM

You can send these query results to the CDM to create a cohort. Your cohort will be available to Heracles and other CDM tools. This may take several minutes to complete.

Cohort Name: Mesenteric Panniculitis

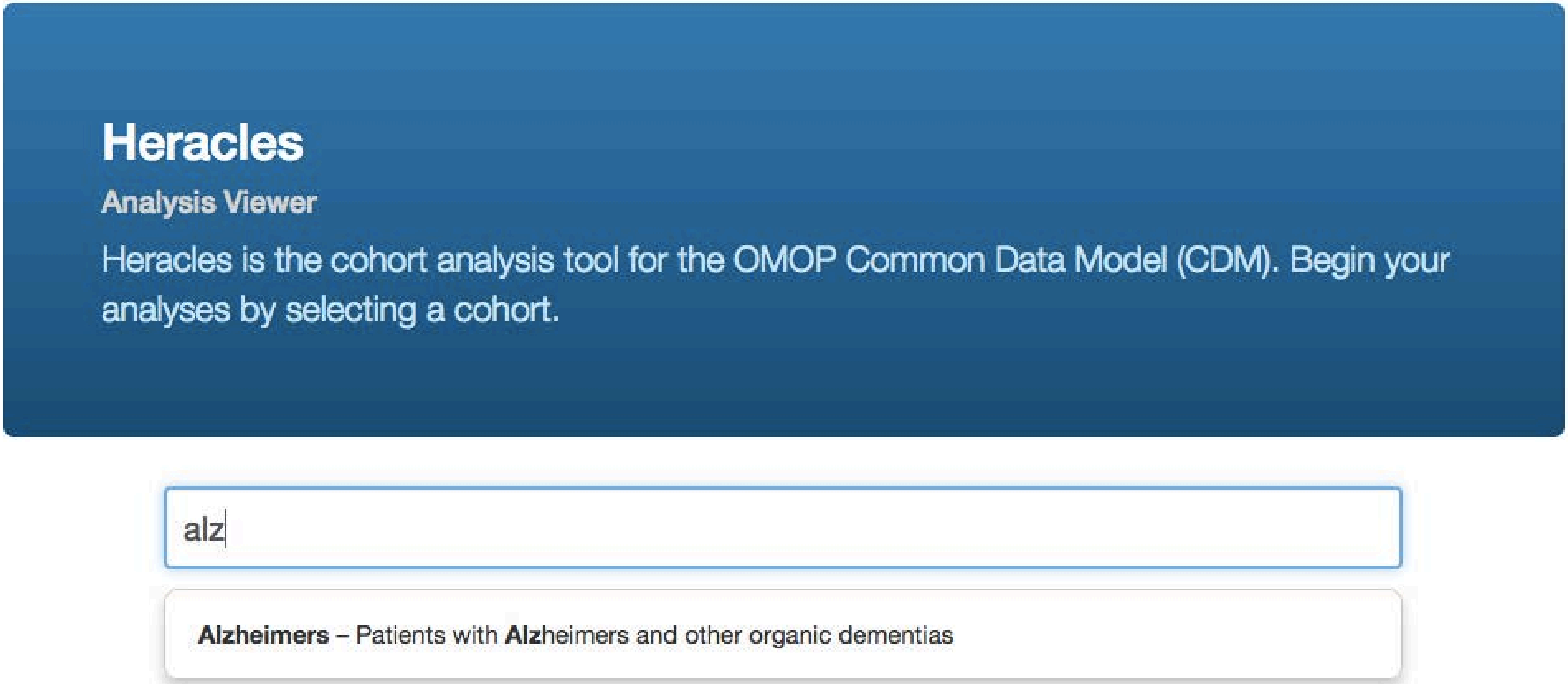
Cohort Description: Patients with evidence of mesenteritis or mesentieric panniculitis

Cohort End Date: Max Observation Date

Cohort Creation vs Analysis

- Currently, cohort definition and analysis are separate in the OHDSI stack
- This was designed to facilitate sharing of cohorts, but may ultimately be merged
- Cohort definition is performed by Circe
- Cohort analyses are performed using Heracles

HERACLES



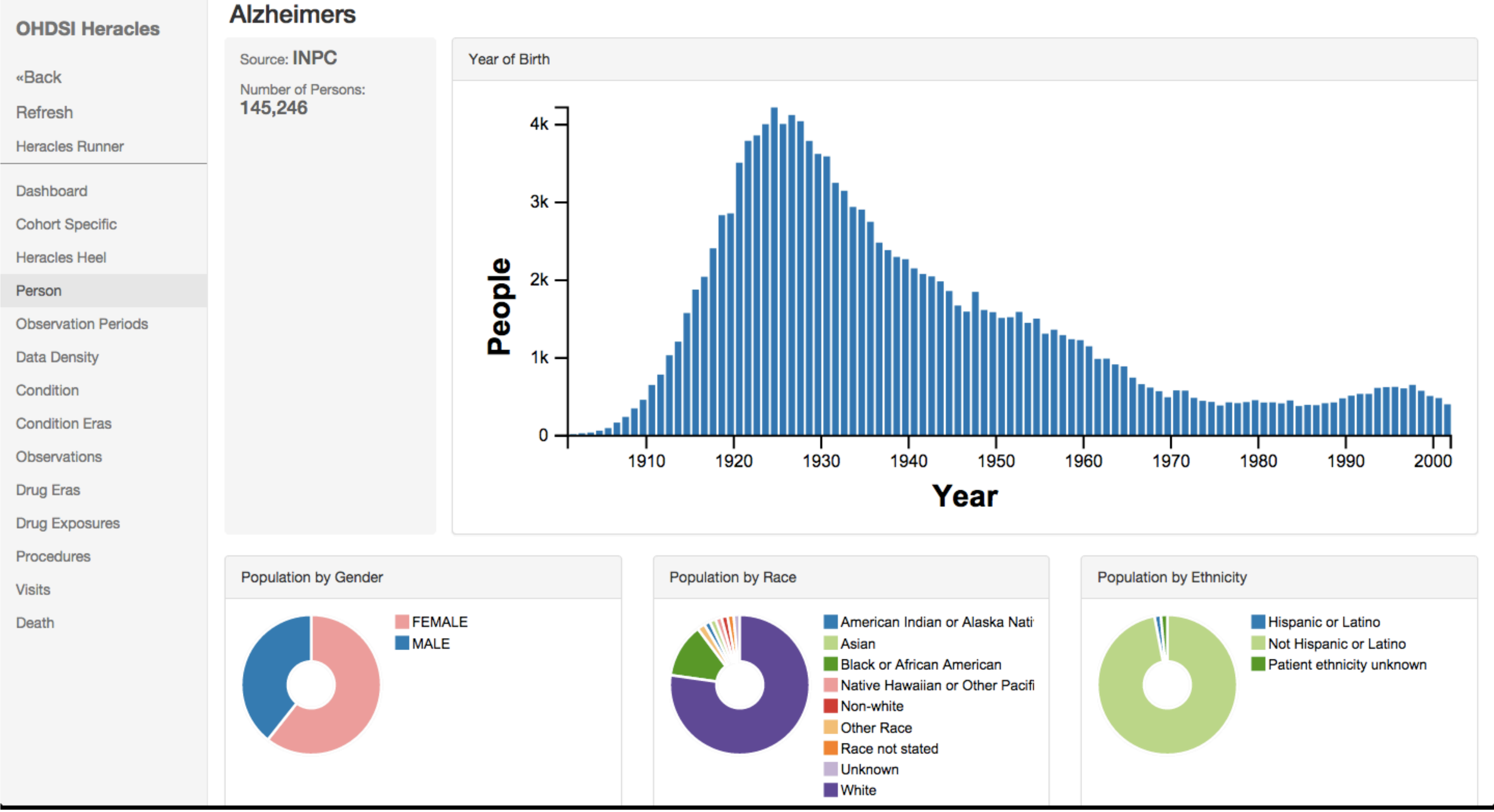
Heracles

Analysis Viewer

Heracles is the cohort analysis tool for the OMOP Common Data Model (CDM). Begin your analyses by selecting a cohort.

alz

Alzheimers – Patients with **Alz**heimers and other organic dementias



OHDSI Heracles

«Back

Refresh

Heracles Runner

Dashboard

Cohort Specific

Heracles Heel

Person

Observation Periods

Data Density

Condition

Condition Eras

Observations

Drug Eras

Drug Exposures

Procedures

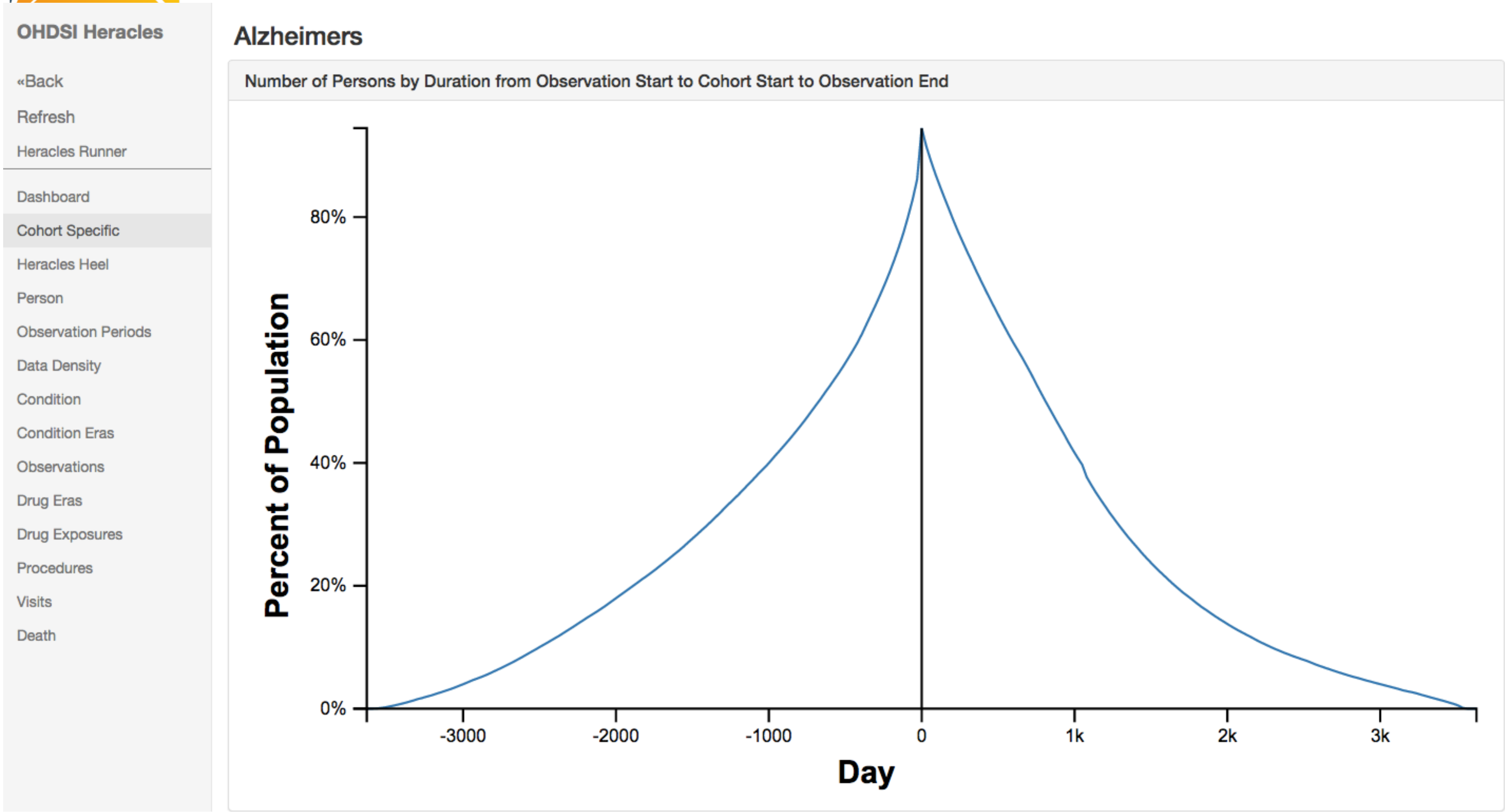
Visits

Death

Alzheimers

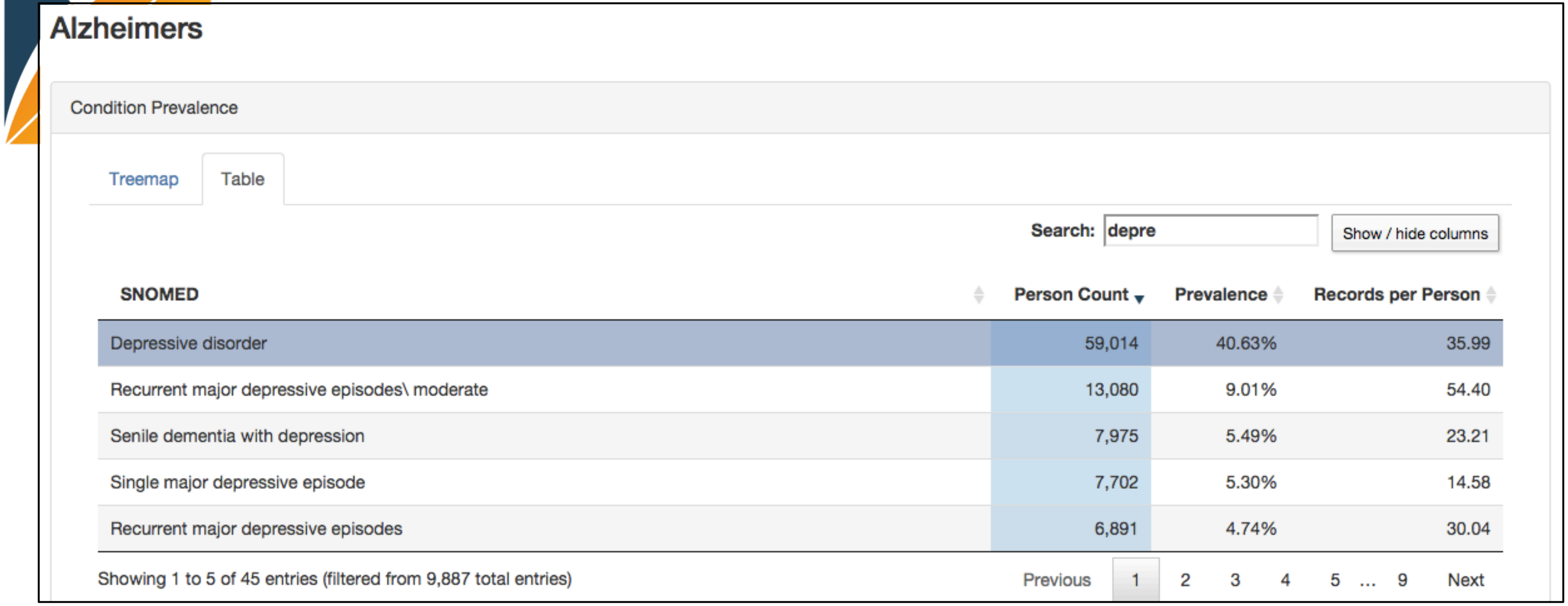
Source: INPC

Number of Persons: 145,246



Alzheimers

Number of Persons by Duration from Observation Start to Cohort Start to Observation End



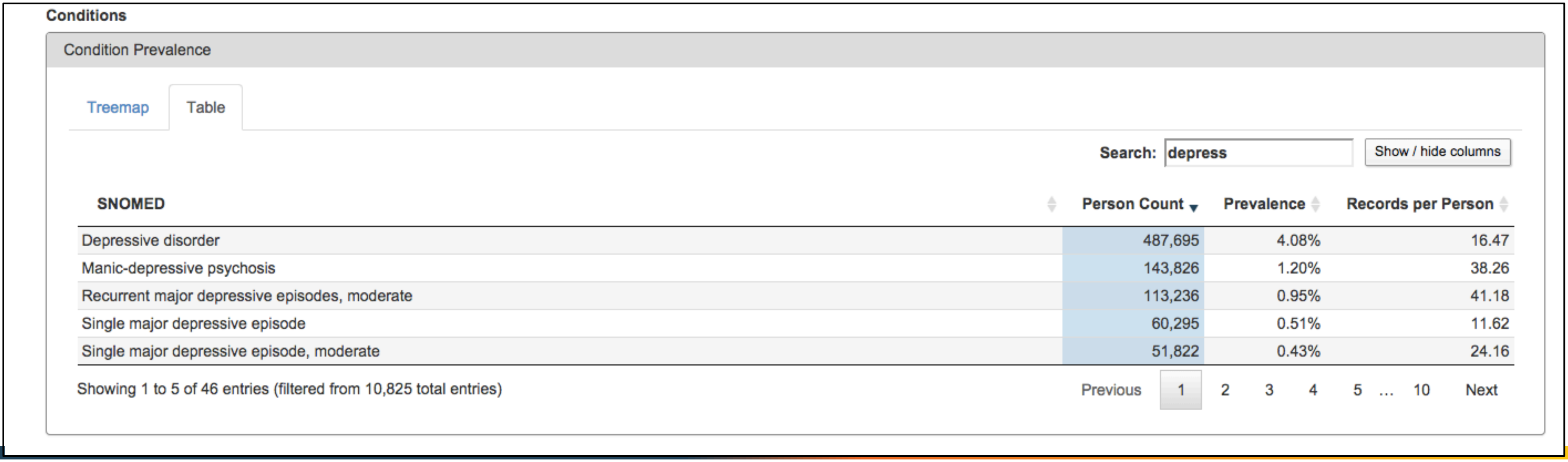
Alzheimers

Condition Prevalence

Search: depre

SNOMED	Person Count	Prevalence	Records per Person
Depressive disorder	59,014	40.63%	35.99
Recurrent major depressive episodes\ moderate	13,080	9.01%	54.40
Senile dementia with depression	7,975	5.49%	23.21
Single major depressive episode	7,702	5.30%	14.58
Recurrent major depressive episodes	6,891	4.74%	30.04

Showing 1 to 5 of 45 entries (filtered from 9,887 total entries)



Conditions

Condition Prevalence

Search: depress

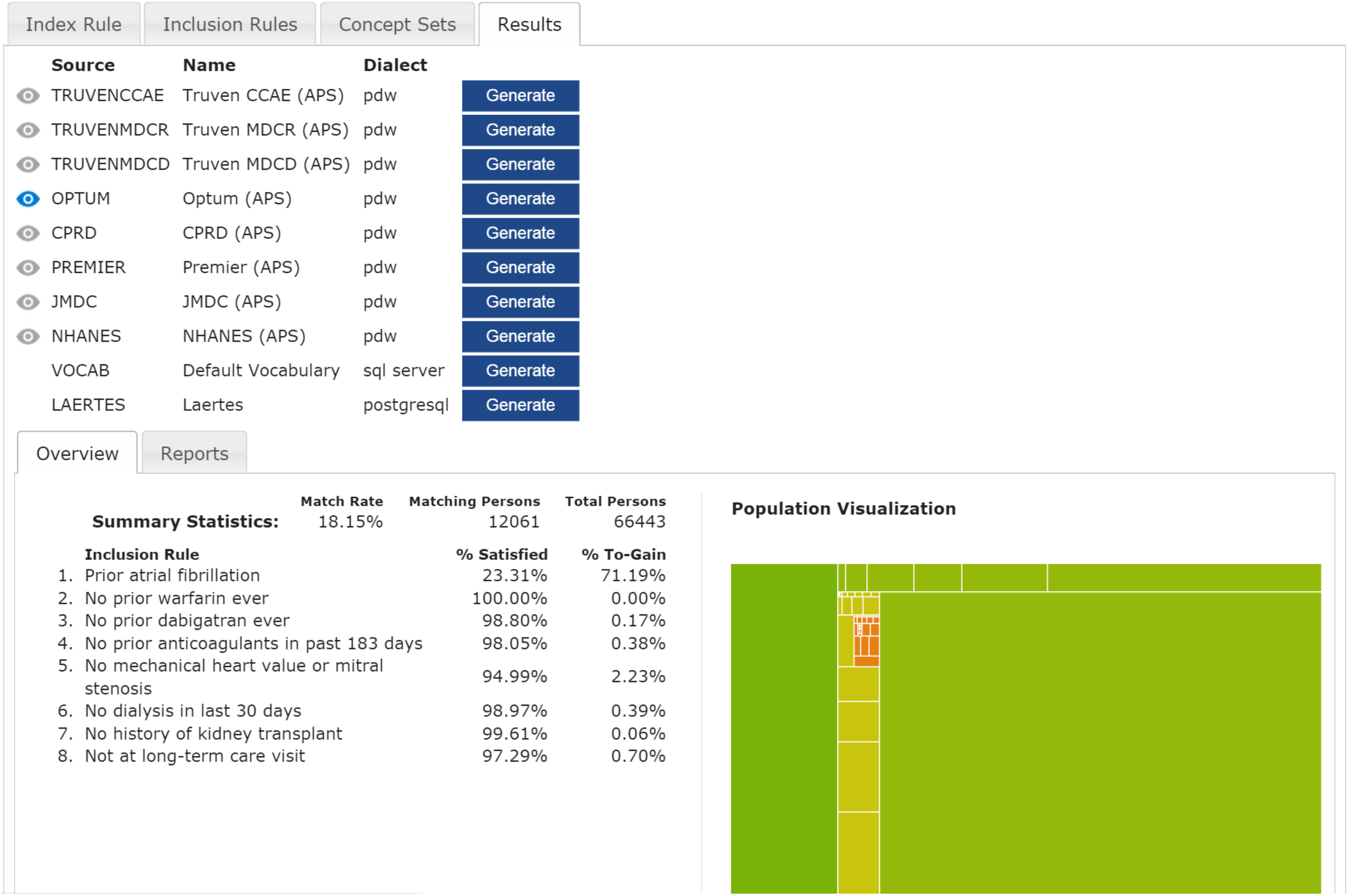
SNOMED	Person Count	Prevalence	Records per Person
Depressive disorder	487,695	4.08%	16.47
Manic-depressive psychosis	143,826	1.20%	38.26
Recurrent major depressive episodes, moderate	113,236	0.95%	41.18
Single major depressive episode	60,295	0.51%	11.62
Single major depressive episode, moderate	51,822	0.43%	24.16

Showing 1 to 5 of 46 entries (filtered from 10,825 total entries)

HERACLES Parameters

- Can limit to specific analyses (e.g., just procedures)
- Can target specific concepts (e.g., a drug class, a particular condition)
- Can window on cohort-specific date ranges

CALYPSO: Integrating Cohorts with Clinical Trial Recruitment



	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

	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

Part III. Network-based Research

Network-based Research

- International network of researchers
  - Data holders
  - Standards developers
  - Methods developers
  - Clinical researchers
- Large-scale collaborative research
  - Larger sample sizes
  - More diverse population
  - Greater expertise

Open-source process

- Join the collaborative
- Propose a study to the open collaborative
- Write protocol
  - http://www.ohdsi.org/web/wiki/doku.php?id=research:studies
- Code it, run it locally, debug it (minimize others' work)
- Publish it: https://github.com/ohdsi
- Each node voluntarily executes on their CDM
- Centrally share results
- Collaboratively explore results and jointly publish findings

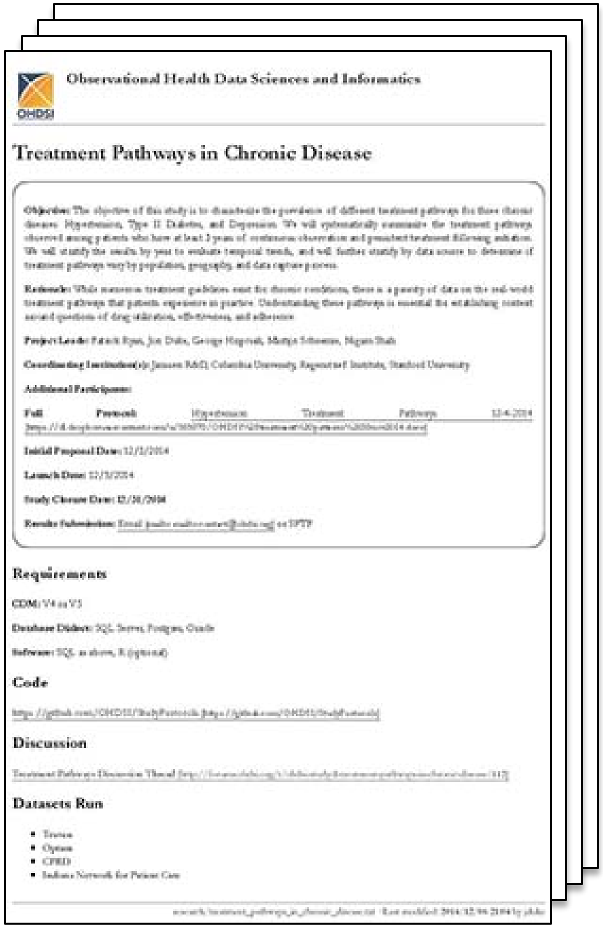
OHDSI in action: Chronic disease treatment pathways

- Conceived at AMIA 15Nov2014
- Protocol written, code written and tested at 2 sites 30Nov2014
- Analysis submitted to OHDSI network 2Dec2014
- Results submitted for 7 databases by 5Dec2014

Condition definitions

Disease	Medication classes	Diagnosis	Exclusions
Hypertension (“HTN”)	antihypertensives, diuretics, peripheral vasodilators, beta blocking agents, calcium channel blockers, agents acting on the renin-angiotensin system (all ATC)	hyperpiesis (SNOMED)	pregnancy observations (SNOMED)
Diabetes mellitus, Type 2 (“Diabetes”)	drugs used in diabetes (ATC), diabetic therapy (FDB)	diabetes mellitus (SNOMED)	pregnancy observations (SNOMED), type 1 diabetes mellitus (MedDRA)
Depression	antidepressants (ATC), antidepressants (FDB)	depressive disorder (SNOMED)	pregnancy observations (SNOMED), bipolar I disorder (SNOMED), schizophrenia (SNOMED)

Protocol



Observational Health Data Sciences and Informatics

Treatment Pathways in Chronic Disease

Objective: The objective of this study is to characterize the prevalence of different treatment pathways for three chronic diseases: Hypertension, Type II Diabetes, and Depression. We will systematically summarize the treatment pathways observed among patients who have at least 3 years of continuous observation and persistent treatment following initiation. We will stratify the results by year to evaluate temporal trends, and will further stratify by data source to determine if treatment pathways vary by population, geography, and data capture process.

Rationale: While numerous treatment guidelines exist for chronic conditions, there is a paucity of data on the real-world treatment pathways that patients experience in practice. Understanding these pathways is essential for establishing context around questions of drug utilization, effectiveness, and adherence.

Project Leads: Patrick Ryan, Jon Duke, George Hripcsak, Martijn Schuemie, Nigam Shah

Coordinating Institution(s): Janssen R&D, Columbia University, Regenstrief Institute, Stanford University

Additional Participants:

Full Protocol: Hypertension Treatment Pathways 12-4-2014 (https://dl.dropboxusercontent.com/u/[illegible]/OHDSI%20Treatment%20Pathways%20[illegible]2014.docx)

Initial Proposal Date: 12/1/2014

Launch Date: 12/5/2014

Study Closure Date: 12/31/2014

Results Submission: Email (mailto:contact@ohdsi.org) or SFTP

Requirements

CDM: V4 or V5

Database Dialect: SQL Server, Postgres, Oracle

Software: SQL as above, R (optional)

Code

https://github.com/OHDSI/StudyProtocols (https://github.com/OHDSI/StudyProtocols)

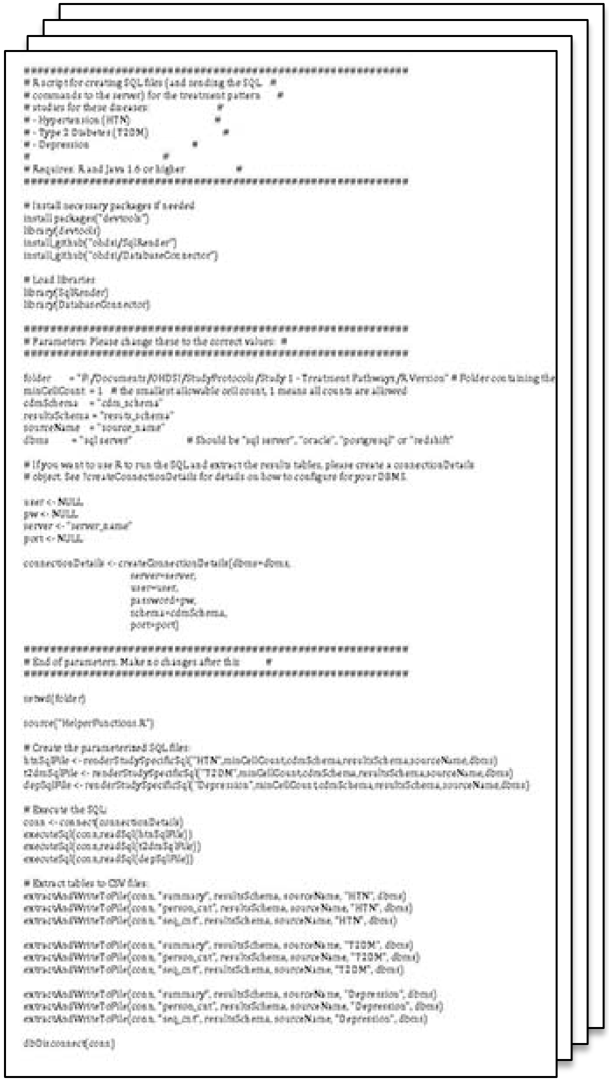
Discussion

Treatment Pathways Discussion Thread (http://forums.ohdsi.org/t/ohdsi-study-1-treatment-pathways-in-chronic-disease/117)

Datasets Run

- Truven
- Optum
- CPRD
- Indiana Network for Patient Care

research/treatment_pathways_in_chronic_disease.txt · Last modified: 2014/12/08 21:04 by jduke



```
####################################################################
# R script for creating SQL files (and sending the SQL   #
# commands to the server) for the treatment pattern      #
# studies for these diseases:                          #
# - Hypertension (HTN)                                #
# - Type 2 Diabetes (T2DM)                             #
# - Depression                                      #
#                                                #
# Requires: R and Java 1.6 or higher                  #
####################################################################

# Install necessary packages if needed
install.packages("devtools")
library(devtools)
install_github("ohdsi/SqlRender")
install_github("ohdsi/DatabaseConnector")

# Load libraries
library(SqlRender)
library(DatabaseConnector)

####################################################################
# Parameters: Please change these to the correct values:  #
####################################################################

folder        = "F:/Documents/OHDSI/StudyProtocols/Study 1 - Treatment Pathways/R Version" # Folder containing the
minCellCount  = 1   # the smallest allowable cell count, 1 means all counts are allowed
cdmSchema     = "cdm_schema"
resultsSchema = "results_schema"
sourceName    = "source_name"
dbms          = "sql server"               # Should be "sql server", "oracle", "postgresql" or "redshift"

# If you want to use R to run the SQL and extract the results tables, please create a connectionDetails
# object. See ?createConnectionDetails for details on how to configure for your DBMS.

user <- NULL
pw <- NULL
server <- "server_name"
port <- NULL

connectionDetails <- createConnectionDetails(dbms=dbms,
                                server=server,
                                user=user,
                                password=pw,
                                schema=cdmSchema,
                                port=port)

####################################################################
# End of parameters. Make no changes after this          #
####################################################################

setwd(folder)

source("HelperFunctions.R")

# Create the parameterized SQL files:
htnSqlFile <- renderStudySpecificSql("HTN",minCellCount,cdmSchema,resultsSchema,sourceName,dbms)
t2dmSqlFile <- renderStudySpecificSql("T2DM",minCellCount,cdmSchema,resultsSchema,sourceName,dbms)
depSqlFile <- renderStudySpecificSql("Depression",minCellCount,cdmSchema,resultsSchema,sourceName,dbms)

# Execute the SQL:
conn <- connect(connectionDetails)
executeSql(conn,readSql(htnSqlFile))
executeSql(conn,readSql(t2dmSqlFile))
executeSql(conn,readSql(depSqlFile))

# Extract tables to CSV files:
extractAndWriteToFile(conn, "summary", resultsSchema, sourceName, "HTN", dbms)
extractAndWriteToFile(conn, "person_cnt", resultsSchema, sourceName, "HTN", dbms)
extractAndWriteToFile(conn, "seq_cnt", resultsSchema, sourceName, "HTN", dbms)

extractAndWriteToFile(conn, "summary", resultsSchema, sourceName, "T2DM", dbms)
extractAndWriteToFile(conn, "person_cnt", resultsSchema, sourceName, "T2DM", dbms)
extractAndWriteToFile(conn, "seq_cnt", resultsSchema, sourceName, "T2DM", dbms)

extractAndWriteToFile(conn, "summary", resultsSchema, sourceName, "Depression", dbms)
extractAndWriteToFile(conn, "person_cnt", resultsSchema, sourceName, "Depression", dbms)
extractAndWriteToFile(conn, "seq_cnt", resultsSchema, sourceName, "Depression", dbms)

dbDisconnect(conn)
```

OHDSI participating data partners

Code	Name	Description	Size (M)
AUSOM	Ajou University School of Medicine	South Korea; inpatient hospital EHR	2
CCAE	MarketScan Commercial Claims and Encounters	US private-payer claims	119
CPRD	UK Clinical Practice Research Datalink	UK; EHR from general practice	11
CUMC	Columbia University Medical Center	US; inpatient EHR	4
GE	GE Centricity	US; outpatient EHR	33
INPC	Regenstrief Institute, Indiana Network for Patient Care	US; integrated health exchange	15
JMDC	Japan Medical Data Center	Japan; private-payer claims	3
MDCD	MarketScan Medicaid Multi-State	US; public-payer claims	17
MDCR	MarketScan Medicare Supplemental and Coordination of Benefits	US; private and public-payer claims	9
OPTUM	Optum ClinFormatics	US; private-payer claims	40
STRIDE	Stanford Translational Research Integrated Database Environment	US; inpatient EHR	2
HKU	Hong Kong University	Hong Kong; EHR	1

Medication-use metrics

- Define generic metrics to be used on all medications
  - Monotherapy: patients who used exactly one medication in the three-year window (one at a time and no changes)
  - Monotherapy with common medication: patients whose monotherapy was the most common mono-med for that condition
  - Start with common medication: patients who started with the most common starting med for that condition

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Open-Source Big Data Analytics in Healthcare

Discussion