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.

http://ohdsi.org
Why large-scale analysis is needed in healthcare

All health outcomes of interest
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.

http://ohdsi.org
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
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**
- Data quality assessment
- Common Data Model evaluation
- ATHENA for standardized vocabularies
- Phenotype evaluation
- Empirical calibration
- LAERTES for evidence synthesis
- Evaluation framework and benchmarking

**Open-source analytics development**
- WhiteRabbit for CDM ETL
- Usagi for vocabulary mapping
- HERMES for vocabulary exploration
- ACHILLES for database profiling
- CIRCE for cohort definition
- CALYPSO for feasibility assessment
- HERACLES for cohort characterization
- CohortMethod
- SelfControlledCaseSeries
- SelfControlledCohort
- TemporalPatternDiscovery
- PatientLevelPrediction
- APHRODITE for predictive phenotyping

**Clinical applications**
- Chronic disease therapy pathways
- HOMER for causality assessment
- 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

- **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

Source data warehouse, with identifiable patient-level data

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

Standardized large-scale analytics

Analysis results

Summary statistics results repository

OHDSI Coordinating Center

Data network support
Analytics development and testing
Research and education

OHDSI Data Partners
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
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.

http://omop.org/CDM
One model, multiple use cases

Standardized health system data
- Observation_period
- Specimen
- Death
- Visit_occurrence
- Procedure_occurrence
- Drug_exposure
- Device_exposure
- Condition_occurrence
- Measurement
- Note
- Observation
- Fact_relationship

Standardized derived elements
- Cohort
- Cohort_attribute
- Condition_era
- Drug_era
- Dose_era

Standardized economics
- Payer_plan_period
- Visit_cost
- Procedure_cost
- Drug_cost
- Device_cost

Standardized health
- Location
- Care_site
- Provider

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 meta-data
- CDM_source

Standardized clinical data
- Drug_exposure
- Condition_occurrence
- Procedure_occurrence
- Visit_occurrence
- Observation_period
- Death

Standardized health system data
- Location
- Care_site
- Provider

Standardized derived elements
- Cohort
- Cohort_attribute
- Condition_era
- Drug_era
- Dose_era

Standardized economics
- Payer_plan_period
- Visit_cost
- Procedure_cost
- Drug_cost
- Device_cost

Standardized health
- Observation_period
- Specimen
- Death
- Visit_occurrence
- Procedure_occurrence
- Drug_exposure
- Device_exposure
- Condition_occurrence
- Measurement
- Note
- Observation
- Fact_relationship

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 meta-data
- CDM_source

Standardized clinical data
- Drug_exposure
- Condition_occurrence
- Procedure_occurrence
- Visit_occurrence
- Observation_period
- Death

Standardized health system data
- Location
- Care_site
- Provider

Standardized derived elements
- Cohort
- Cohort_attribute
- Condition_era
- Drug_era
- Dose_era

Standardized economics
- Payer_plan_period
- Visit_cost
- Procedure_cost
- Drug_cost
- Device_cost

Standardized health
- Observation_period
- Specimen
- Death
- Visit_occurrence
- Procedure_occurrence
- Drug_exposure
- Device_exposure
- Condition_occurrence
- Measurement
- Note
- Observation
- Fact_relationship

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 meta-data
- CDM_source
Standardized Vocabularies: Conditions

Classifications

SNOMED-CT

MedDRA

SMQ

Source codes

ICD-10-CM
Read
Oxmis
ICD-9-CM
SNOMED

Cohort

System organ class (Level 5)
High-level group terms (Level 4)
High-level terms (Level 3)
Preferred terms (Level 2)
Low-level terms (Level 1)

Mapping
- Existing
- De Novo
- Derived
Preparing your data for analysis

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

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 tools built to help

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
Data Evidence sharing paradigms

- **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

Patient-level data in OMOP CDM

- 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

CALYPSO: Feasibility assessment

HERACLES: Cohort characterization

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
HERMES: Explore the standardized vocabularies to define exposures, outcomes, and covariates.
CIRCE: Define cohorts of interest

Index Population: MiniSentinel replication - warfarin new users

Description:

Expression

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.

Add Additional Filters

Limit cohort expression results to: All Events per person.

Show SQL Add Options
CALYPSO: Conduct feasibility assessment to evaluate the impact of study inclusion criteria
HERACLES: Characterize the cohorts of interest
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
Where to find the CDM?

Specifications and related files for the Common Data Model — Edit

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>Reordered the folder structure</td>
<td>5 months ago</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>Reordered the folder structure</td>
<td>5 months ago</td>
</tr>
<tr>
<td>Sql Server</td>
<td>Reordered the folder structure</td>
<td>5 months ago</td>
</tr>
<tr>
<td>Version4 To Version5 Conversion Improvement</td>
<td>Improvements to scripts, documentation and inclusion of DRG conversion.</td>
<td>13 days ago</td>
</tr>
<tr>
<td>Version4</td>
<td>changes after V4 testing</td>
<td>5 months ago</td>
</tr>
<tr>
<td>LICENSE</td>
<td>Initial commit</td>
<td>10 months ago</td>
</tr>
<tr>
<td>OMOP CDM v5.pdf</td>
<td>Added PDF file</td>
<td>10 months ago</td>
</tr>
<tr>
<td>README.md</td>
<td>Initial commit</td>
<td>10 months ago</td>
</tr>
</tbody>
</table>
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.

![Diagram showing mapping of fields from Source to Target]

Source
- desynpuf_id
- srcv_dt
- prod_srcv_id
- qty_dspsnsd_num
- days_suply_num

Target
- *person_id
- *drug_exposure_start_date
- quantity
- days_supply
- drug_source_value
**OHDSI Has Extensive Vocabulary Maps**

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

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

**Overview Table**

**Selected Mapping**

**Search Facility**
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
- CIRCE
  Cohort Creation
- HERMES
  OMOP Vocabulary Explorer
- HERACLES
  Cohort Characterization
- ACHILLES
  Dataset Characterization
- CALYPSO
  Clinical Trial Feasibility
OHDSI Web Tools

ATHENA
OMOP Vocabulary Loader

CIRCE
Cohort Creation

CIRCE:
Define Patient Cohorts

HERMES:
Explore the OMOP Vocabulary

HERMES
OMOP Vocabulary Explorer

HERACLES:
Explore Cohort Level Data

HERACLES
Cohort Characterization

ACHILLES:
Explore Population Level Data

ACHILLES
Dataset Characterization

CALYPSO:
Clinical Trial Feasibility

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 Report Types
ACHILLES Heel Helps You Validate Your Data Quality

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

Cohort End Date:
Max Observation Date →

Send
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
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 Alzheimers and other organic dementias
Alzheimers

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

Percent of Population

Day
### Alzheimers

#### Condition Prevalence

**Treemap** | **Table**
--- | ---

**SNOMED**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Person Count</th>
<th>Prevalence</th>
<th>Records per Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressive disorder</td>
<td>69,014</td>
<td>40.63%</td>
<td>35.99</td>
</tr>
<tr>
<td>Recurrent major depressive episodes</td>
<td>13,080</td>
<td>9.01%</td>
<td>54.40</td>
</tr>
<tr>
<td>Senile dementia with depression</td>
<td>7,975</td>
<td>5.49%</td>
<td>23.21</td>
</tr>
<tr>
<td>Single major depressive episode</td>
<td>7,702</td>
<td>5.30%</td>
<td>14.58</td>
</tr>
<tr>
<td>Recurrent major depressive episodes</td>
<td>6,891</td>
<td>4.74%</td>
<td>30.04</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 46 entries (filtered from 9,887 total entries)

---

### Conditions

#### Condition Prevalence

**Treemap** | **Table**
--- | ---

**SNOMED**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Person Count</th>
<th>Prevalence</th>
<th>Records per Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressive disorder</td>
<td>487,695</td>
<td>4.08%</td>
<td>16.47</td>
</tr>
<tr>
<td>Manic-depressive psychosis</td>
<td>143,826</td>
<td>1.20%</td>
<td>38.26</td>
</tr>
<tr>
<td>Recurrent major depressive episodes, moderate</td>
<td>113,236</td>
<td>0.96%</td>
<td>41.18</td>
</tr>
<tr>
<td>Single major depressive episode</td>
<td>60,295</td>
<td>0.51%</td>
<td>11.62</td>
</tr>
<tr>
<td>Single major depressive episode, moderate</td>
<td>51,822</td>
<td>0.43%</td>
<td>24.18</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Name</th>
<th>Dialect</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUVENCCAE</td>
<td>Truven CCAE (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>TRUVENMDCR</td>
<td>Truven MDCR (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>TRUVENMDCD</td>
<td>Truven MDCD (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>OPTUM</td>
<td>Optum (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>CPRD</td>
<td>CPRD (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>PREMIER</td>
<td>Premier (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>JMDC</td>
<td>JMDC (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>NHANES</td>
<td>NHANES (APS)</td>
<td>pdw</td>
<td>Generate</td>
</tr>
<tr>
<td>VOCAB</td>
<td>Default Vocabulary</td>
<td>sql server</td>
<td>Generate</td>
</tr>
<tr>
<td>LAERTEES</td>
<td>Laertes</td>
<td>postgresql</td>
<td>Generate</td>
</tr>
</tbody>
</table>

**Summary Statistics:**

- **Match Rate:** 18.15%
- **Matching Persons:** 12061
- **Total Persons:** 66443

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

**Population Visualization**

[Visualization showing population distribution]
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
• Code it, run it locally, debug it (minimize others’ work)
• Publish it: [https://github.com/ohdsi](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

<table>
<thead>
<tr>
<th>Disease</th>
<th>Medication classes</th>
<th>Diagnosis</th>
<th>Exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension (“HTN”)</td>
<td>antihypertensives, diuretics, peripheral vasodilators, beta blocking agents, calcium channel blockers, agents acting on the renin-angiotensin system (all ATC)</td>
<td>hyperpiesis (SNOMED)</td>
<td>pregnancy observations (SNOMED)</td>
</tr>
<tr>
<td>Diabetes mellitus, Type 2 (“Diabetes”)</td>
<td>drugs used in diabetes (ATC), diabetic therapy (FDB)</td>
<td>diabetes mellitus (SNOMED)</td>
<td>pregnancy observations (SNOMED), type 1 diabetes mellitus (MedDRA)</td>
</tr>
<tr>
<td>Depression</td>
<td>antidepressants (ATC), antidepressants (FDB)</td>
<td>depressive disorder (SNOMED)</td>
<td>pregnancy observations (SNOMED), bipolar I disorder (SNOMED), schizophrenia (SNOMED)</td>
</tr>
</tbody>
</table>
Protocol
<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Description</th>
<th>Size (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUSOM</td>
<td>Ajou University School of Medicine</td>
<td>South Korea; inpatient hospital EHR</td>
<td>2</td>
</tr>
<tr>
<td>CCAE</td>
<td>MarketScan Commercial Claims and Encounters</td>
<td>US private-payer claims</td>
<td>119</td>
</tr>
<tr>
<td>CPRD</td>
<td>UK Clinical Practice Research Datalink</td>
<td>UK; EHR from general practice</td>
<td>11</td>
</tr>
<tr>
<td>CUMC</td>
<td>Columbia University Medical Center</td>
<td>US; inpatient EHR</td>
<td>4</td>
</tr>
<tr>
<td>GE</td>
<td>GE Centricity</td>
<td>US; outpatient EHR</td>
<td>33</td>
</tr>
<tr>
<td>INPC</td>
<td>Regenstrief Institute, Indiana Network for Patient Care</td>
<td>US; integrated health exchange</td>
<td>15</td>
</tr>
<tr>
<td>JMDC</td>
<td>Japan Medical Data Center</td>
<td>Japan; private-payer claims</td>
<td>3</td>
</tr>
<tr>
<td>MDCD</td>
<td>MarketScan Medicaid Multi-State</td>
<td>US; public-payer claims</td>
<td>17</td>
</tr>
<tr>
<td>MDCR</td>
<td>MarketScan Medicare Supplemental and Coordination of Benefits</td>
<td>US; private and public-payer claims</td>
<td>9</td>
</tr>
<tr>
<td>OPTUM</td>
<td>Optum ClinFormatics</td>
<td>US; private-payer claims</td>
<td>40</td>
</tr>
<tr>
<td>STRIDE</td>
<td>Stanford Translational Research Integrated Database Environment</td>
<td>US; inpatient EHR</td>
<td>2</td>
</tr>
<tr>
<td>HKU</td>
<td>Hong Kong University</td>
<td>Hong Kong; EHR</td>
<td>1</td>
</tr>
</tbody>
</table>
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
Open-Source Big Data Analytics in Healthcare

Discussion