



Observational Health Data Sciences and Informatics (OHDSI)

Data integration & data sharing in the
era of “Big Data”

Berlin 2016 July 12-13



OHDSI (pronounced “Odyssey”)

- 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



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's global research community



- >140 collaborators from 20 different countries
- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- Currently 600 million patient records in 52 databases

<http://ohdsi.org/who-we-are/collaborators/>

Why large-scale analysis is needed in healthcare

All health outcomes of interest

The table displays a comprehensive matrix of drug-outcome relationships. The columns represent various medical conditions and body systems, including: Bl ood disorders, Co rra do disor..., Ea rly eye disor..., En do cr..., Eye disor..., Gastroin testina..., Gener al di..., Immune system di..., Infections and infestations, Injury, poisoning and procedur..., Investiga tions, Metabolis m and n..., Musculoskelet al and conne..., Neoplasms benign, malignant and uns..., Nervous system disorders, Pregnancy, puerperium..., Psychiatric disorders, Renal and urinary diso..., Reproductive system and br..., Respiratory, thoracic and me..., Skin and subcutaneo..., Social and medical..., Surgical and medical..., and Vascular disorders.

The rows list various drug classes and specific drugs, including: OTHER INT..., ACID PREP..., MAGNESIU..., PROPULSI..., OTHER BL..., OTHER LA..., OTHER MIN..., ENZYMES, VITAMIN D..., INTERME..., SECOND-G..., HYDRAZIDES, NON-NUC..., ROTA VIRU..., OTHER AL..., ANTI-ESTR..., BENZIMIDA..., BIGUANIDES, IRON IN CO..., VITAMIN K..., MAGNESIUM, RENIN-INHI..., OTHER AN..., BETA BLOC..., ANTIARRH..., OTHER AN..., THIAZIDES..., 2-AMINO-1..., BIOFLAVO..., OTHER AN..., NUCLEOSI..., OTHER AN..., ANTIPROG..., SELECTIVE..., TESTOSTE..., COXIBS, PROPIONIC..., OTHER CE..., BENZOMO..., SELECTIVE..., OTHER AN..., MONOAMIN..., PROLACTI..., OTHER NE..., NON-SELE..., ALDEHYDE..., CARBAMAT..., MELATONI..., PHENOTHI..., SUBSTITUT..., OPIUM ALK..., SYMPATHO..., OTHER AN..., SYMPATHO..., PROPIONIC..., GLUCOCO..., ANTIDOTES, and WATERSOL...

All drugs



Patient-level predictions for personalized evidence requires big data

2 million patients seem excessive or unnecessary?

- Imagine a provider wants to compare her patient with other patients with the same gender (50%), in the same 10-year age group (10%), and with the same comorbidity of Type 2 diabetes (5%)
- Imagine the patient is concerned about the risk of ketoacidosis (0.5%) associated with two alternative treatments they are considering
- With 2 million patients, you'd only expect to observe 25 similar patients with the event, and would only be powered to observe a relative risk > 2.0

Aggregated data across a health system of 1,000 providers may contain 2,000,000 patients



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?

What is the quality of the current evidence from observational analyses?

ORIGINAL CONTRIBUTION

JAMA

Exposure to Oral Bisphosphonates and Risk of Esophageal Cancer

Chris R. Cardwell, PhD

Christian C. Abnet, PhD

Marie M. Cantwell, PhD

Liam J. Murray, MD

Context Use of oral bisphosphonates has increased dramatically and elsewhere. Esophagitis is a known adverse effect of these drugs, and recent reports suggest a link between bisphosphonate use and esophageal cancer; this has not been robustly investigated.

Objective To investigate the association between bisphosphonate use and esophageal cancer.

Design, Setting, and Participants Data were extracted from the UK General Practice Research Database (GPRD) cohort. The incidence of esophageal and gastric cancer per person-years of risk in both the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared.

DISPHOSPHONATES INHIBIT OSTEOCLAST-MEDIATED BONE RESORPTION AND INCREASE BONE MASS. They are used to treat osteoporosis, hypercalcemia of malignancy, and bone metastases. Esophagitis is a known adverse effect of these drugs, and recent reports suggest a link between bisphosphonate use and esophageal cancer; this has not been robustly investigated.

August 2010: "Among patients in the UK General Practice Research Database, the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer"

seemles ground alendronate tablets has been found on biopsy in patients with bisphosphonate-related esophagitis, and follow-up endoscopies have shown that abnormalities remain after the esophagitis heals.⁶ Reflux esophagitis is an established risk factor for esophageal cancer through the Barrett pathway.⁷⁻⁹ It is not known whether bisphosphonate-related esophagitis can also increase esophageal cancer risk. However, the US Food and Drug Administration recently reported 23 cases of esophageal cancer (between 1995 and 2008) in patients using the bisphosphonate alendronate and a further 31 cases in patients using bisphosphonates in Europe.

cohort. The incidence of esophageal and gastric cancer per person-years of risk in both the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared. The incidence of esophageal cancer alone in the bisphosphonate and control groups was compared.

Conclusion Among patients in the UK General Practice Research Database, the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer.

JAMA. 2010;304(6):657-663

Large studies with appropriate comparison groups, adequate follow-up, robust characterization of bisphosphonate use, and prospective data are needed to determine whether the use of oral bisphosphonates increases the risk of esophageal cancer.

BMJ

RESEARCH

Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort

Jane Green, clinical epidemiologist,¹ Gabriela Czanner, statistician,¹ Gillian Reeves, statistical epidemiologist,¹ Joanna Watson, epidemiologist,¹ Lesley Wise, manager, Pharmacoepidemiology Research and Intelligence Unit,² Valerie Beral, professor of cancer epidemiology¹

ABSTRACT

Objective To examine the hypothesis that risk of oesophageal, but not of gastric or colorectal, cancer is increased in users of oral bisphosphonates.

Design Nested case-control analysis within a primary care cohort of about 6 million people in the UK, with prospectively recorded information on prescribing of bisphosphonates.

Setting UK General Practice Research Database cohort. **Participants** Men and women aged 40 years or over—2954 with oesophageal cancer, 2018 with gastric cancer, and 10 641 with colorectal cancer, diagnosed in 1995-2005; five controls per case matched for age, sex, general practice, and observation time.

Main outcome measures Relative risks for incident invasive cancers of the oesophagus, stomach, and colorectum, adjusted for smoking, alcohol, and body mass index.

Conclusions The risk of oesophageal cancer increased with 10 or more prescriptions for oral bisphosphonates and with prescriptions over about a five year period. In Europe and North America, the incidence of oesophageal cancer at age 60-79 is typically 1 per 1000 population over five years, and this is estimated to increase to about 2 per 1000 with five years' use of oral bisphosphonates.

INTRODUCTION

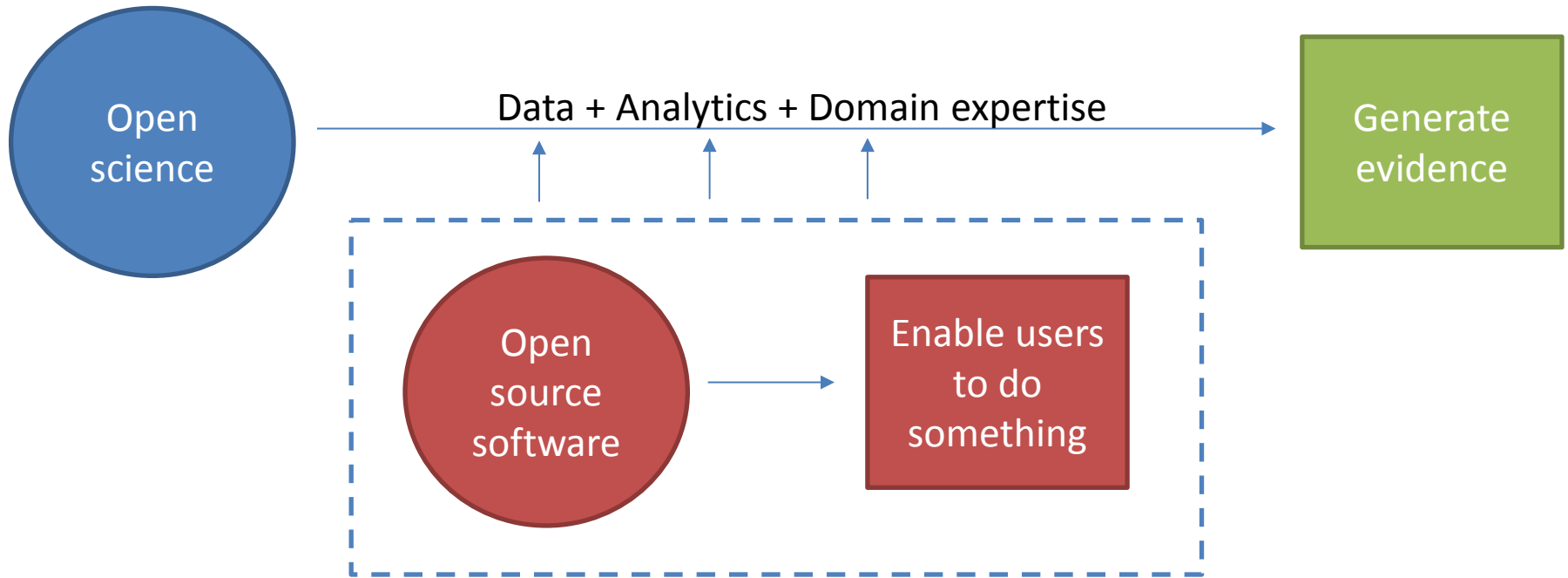
Adverse gastrointestinal effects are common among people who take oral bisphosphonates for the prevention and treatment of osteoporosis; they range from dyspepsia, nausea, and abdominal pain to erosive oesophagitis and oesophageal ulcers.¹ Recent case reports have suggested a possible increase in the risk of oesophageal cancer with use of such bisphosphonate preparations.² We report here on the relation between prospectively recorded prescribing information for

Sept 2010: "In this large nested case-control study within a UK cohort [General Practice Research Database], we found a significantly increased risk of oesophageal cancer in people with previous prescriptions for oral bisphosphonates"

0.87 (0.64 to 1.19) and 0.87 (0.77 to 1.00). The specificity of hospital records are around 95% valid and



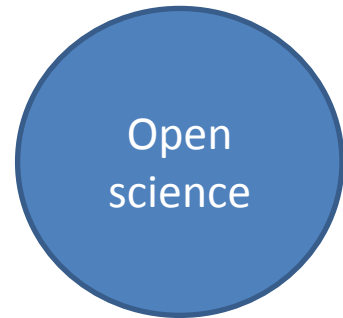
OHDSI's approach to open science



- Open science is about sharing the journey to evidence generation
- Open-source software can be part of the journey, but it's not a final destination
- Open processes can enhance the journey through improved reproducibility of research and expanded adoption of scientific best practices



Standardizing workflows to enable reproducible research



Population-level estimation for comparative effectiveness research:

Is <intervention X> better than <intervention Y> in reducing the risk of <condition Z>?

Generate evidence

Database summary

Cohort definition

Cohort summary

Compare cohorts

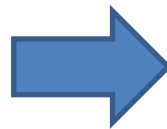
Exposure-outcome summary

Effect estimation & calibration

Compare databases

Defined inputs:

- Target exposure
- Comparator group
- Outcome
- Time-at-risk
- Model specification



Consistent outputs:

- analysis specifications for transparency and reproducibility (protocol + source code)
- only aggregate summary statistics (no patient-level data)
- model diagnostics to evaluate accuracy
- results as evidence to be disseminated
 - static for reporting (e.g. via publication)
 - interactive for exploration (e.g. via app)



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

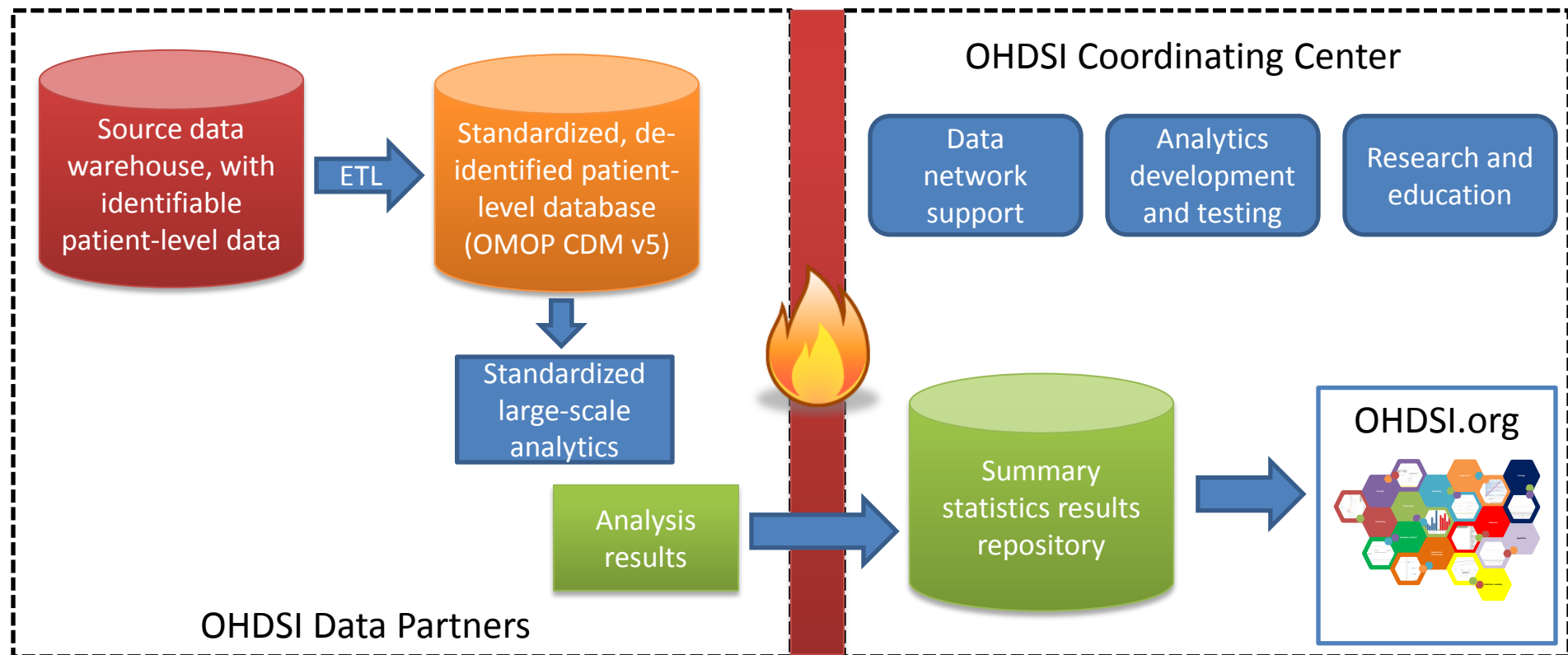


OHDSI Distinguishing Features

- International effort (size & coverage)
 - 43 sources terminologies from around the world
- Open science (depth)
 - Infrastructure serves the science
 - Stack: Terminology, CDM, ETL, QA, Visualization, Novel analytic methods, Clinical research
- Full information model



How OHDSI Works





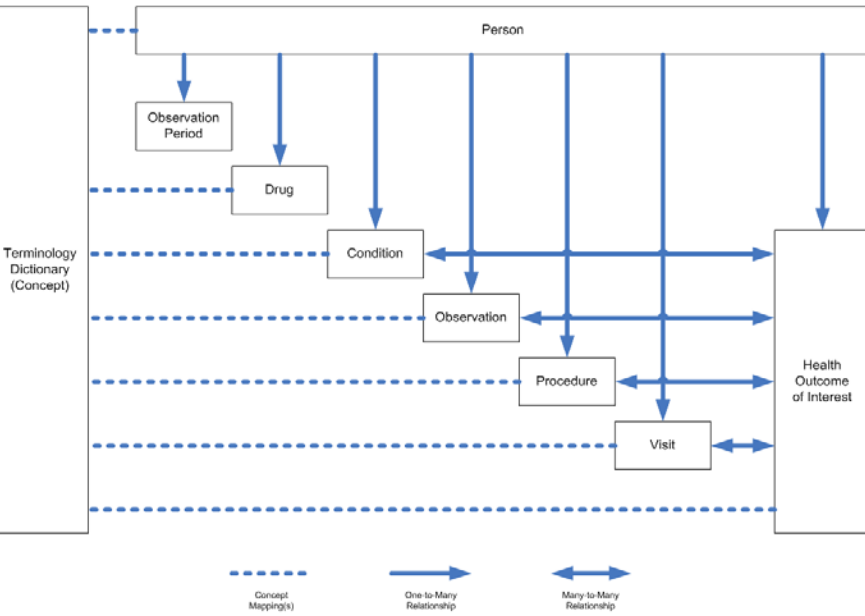
Objectives in OMOP Common Data Model development

- 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 from around the world
- 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

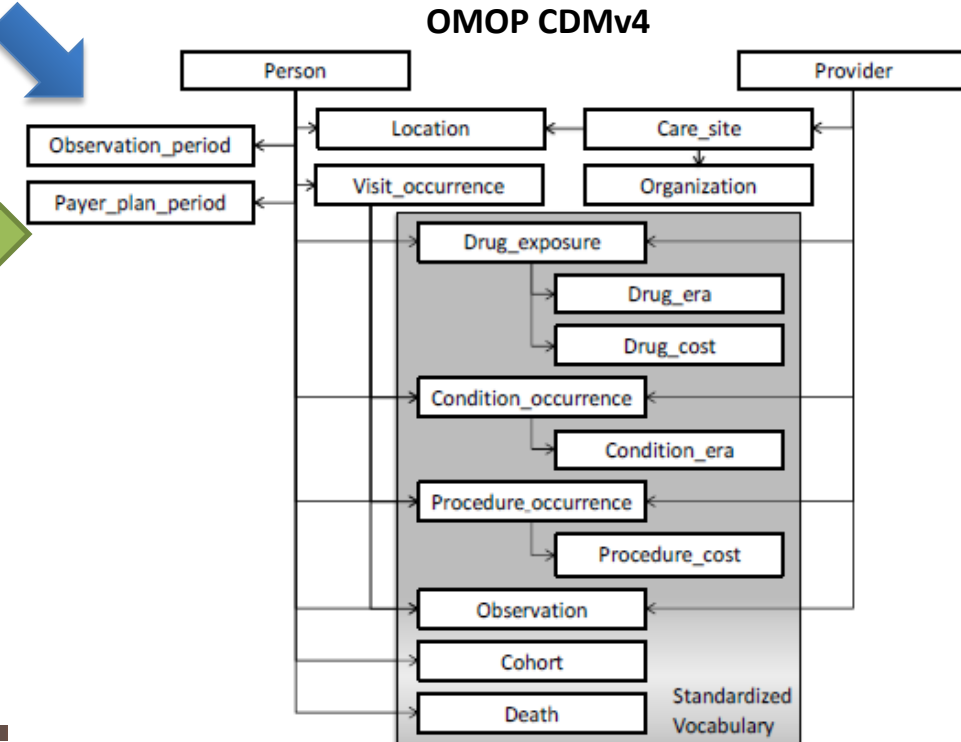
Evolution of the OMOP Common data model

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.

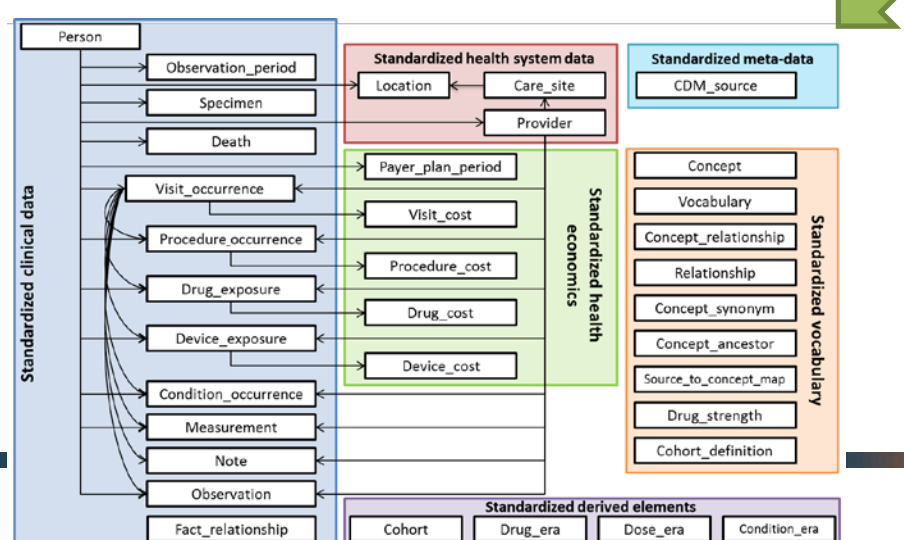
OMOP CDMv2



OMOP CDMv4

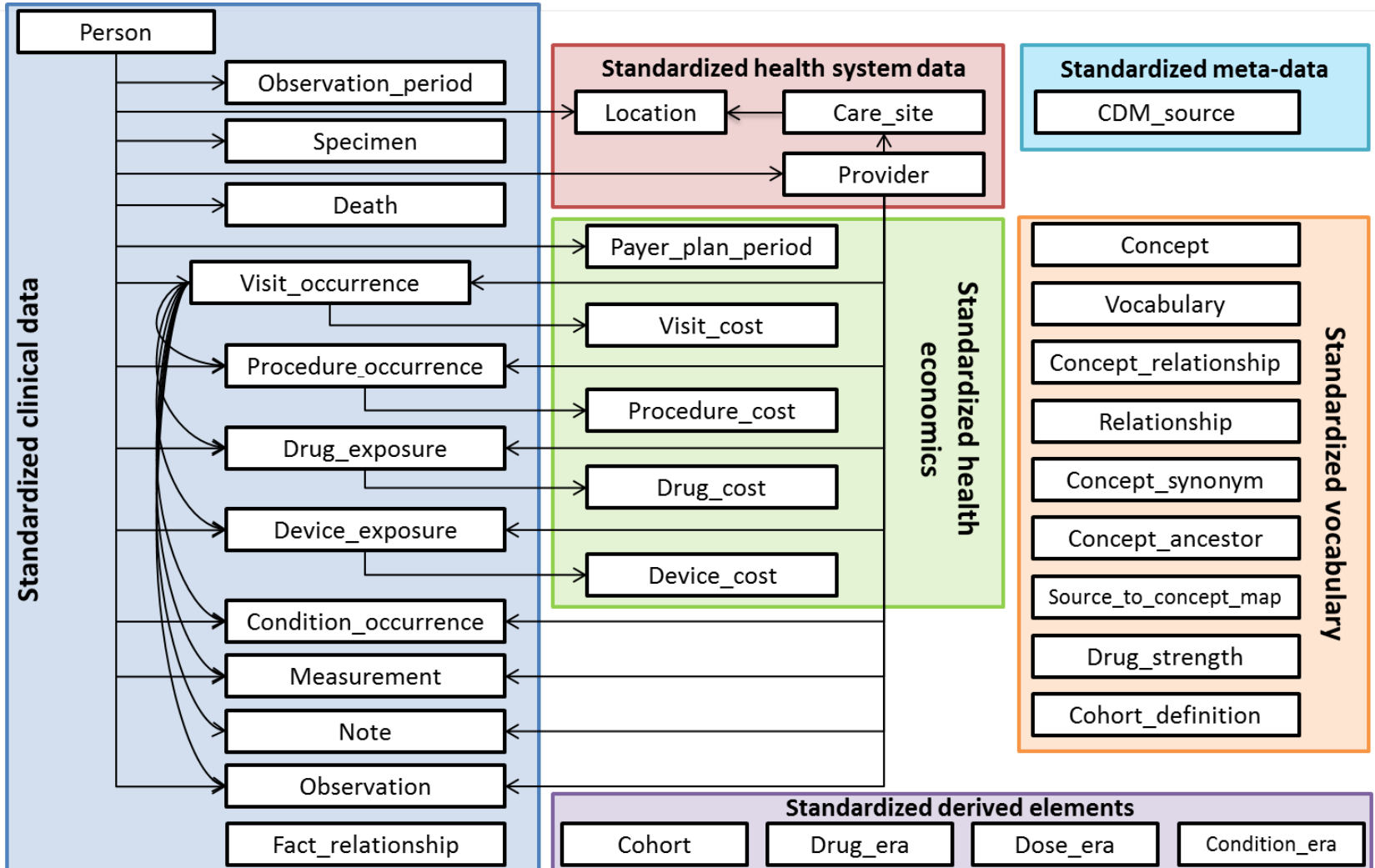


OMOP CDMv5

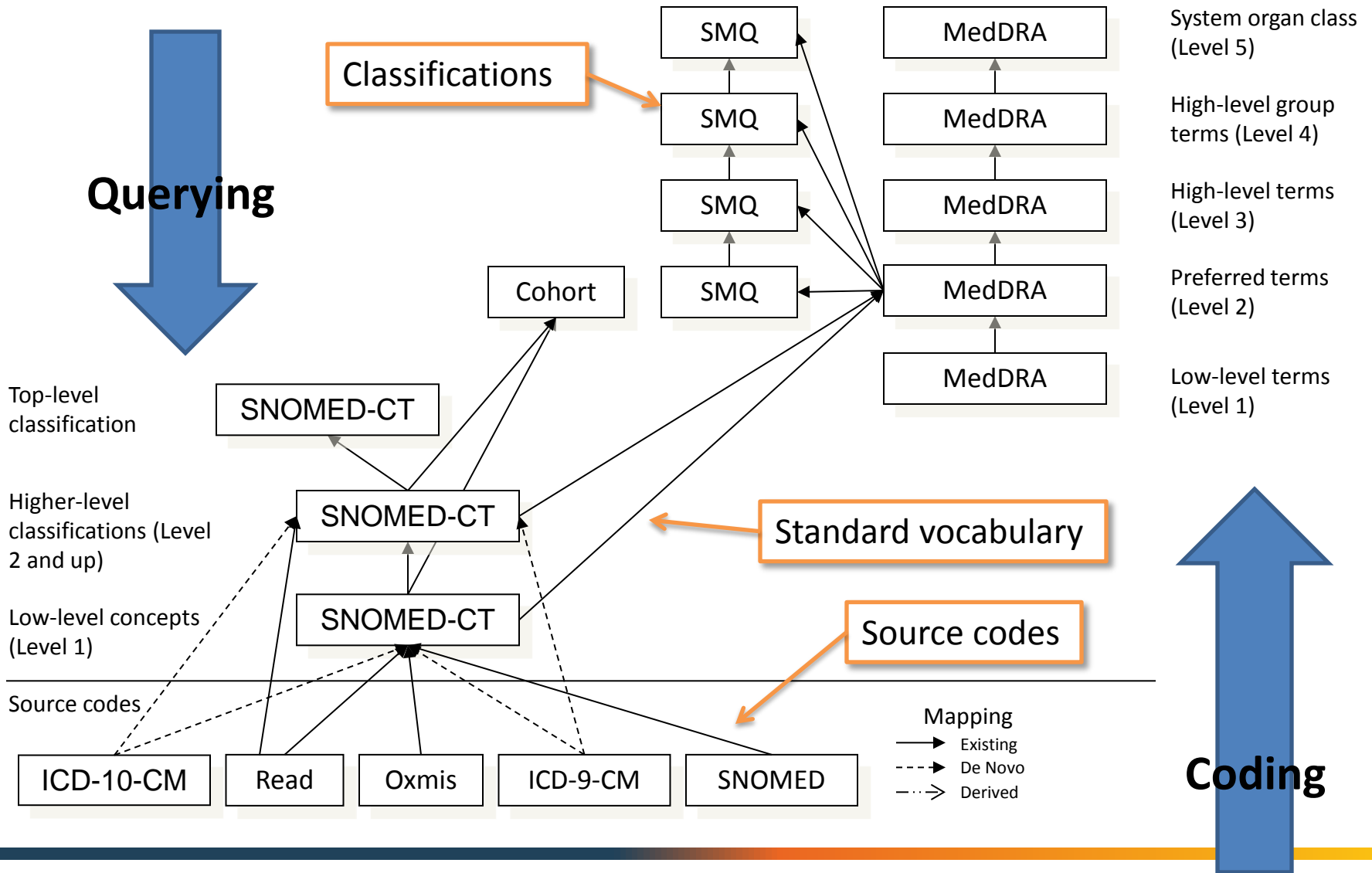




OMOP CDM V5



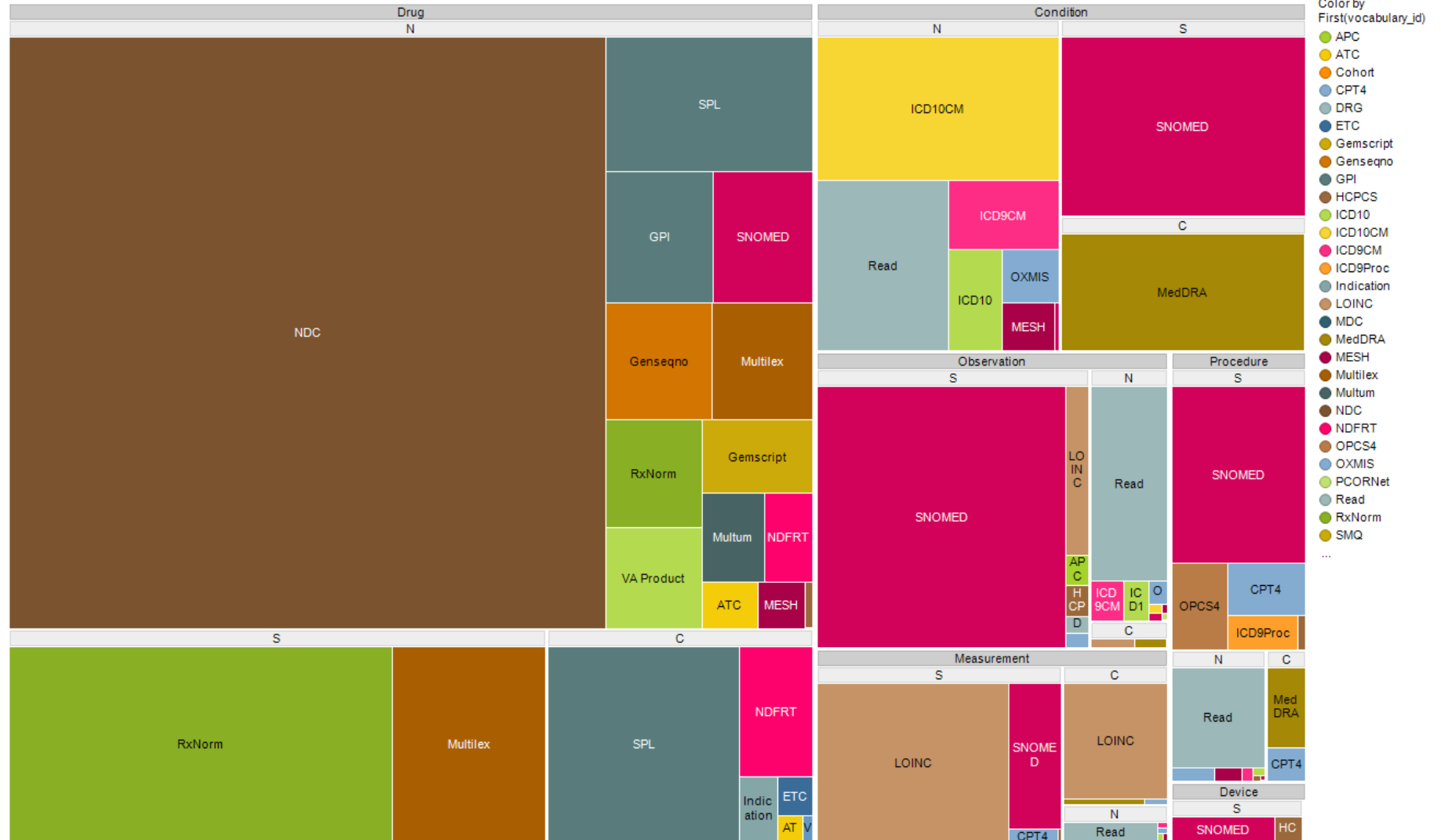
Standardized Vocabularies: Conditions





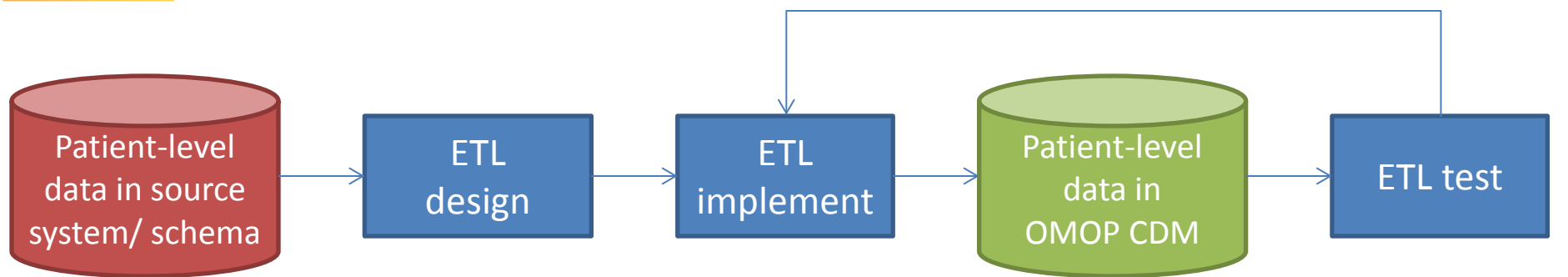
Distribution of Domains in Vocabularies

Breakdown of OHDSI concepts by domain, standard class, and vocabulary





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

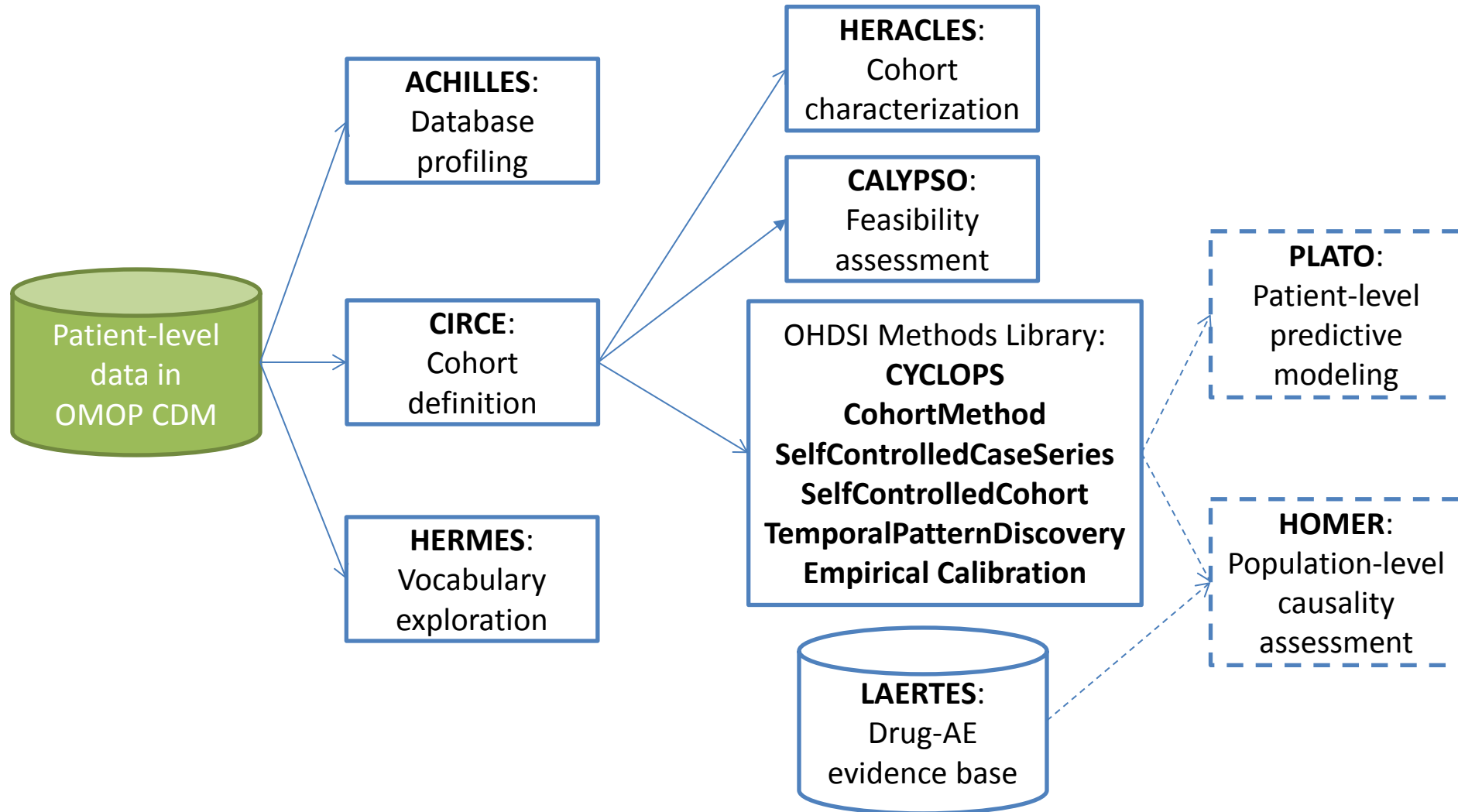
OHDSI Forums:

Public discussions for OMOP CDM Implementers/developers

<http://github.com/OHDSI>



Standardized large-scale analytics tools under development within OHDSI





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

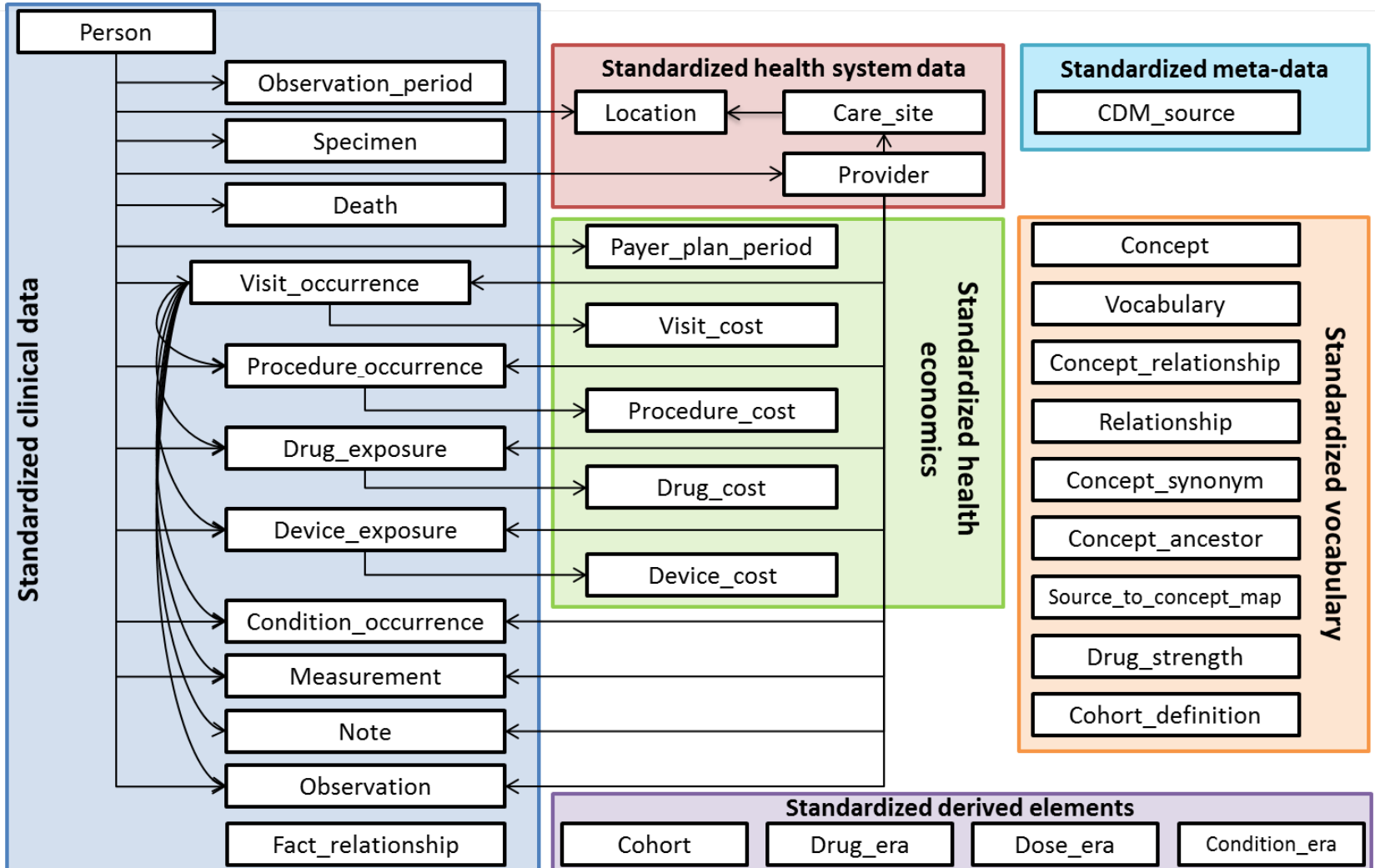


Example

- 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





Where to find the CDM?

 **OHDSI / CommonDataModel**

 Unwatch ▾ 18

Specifications and related files for the Common Data Model — Edit

 26 commits

 1 branch

 1 release

 5 contributors



Branch: **master** ▾

CommonDataModel / +











Merge pull request **#20** from anthonyseena/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



Synthetic Sample Data Set

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



Step 1: What is in your dataset?

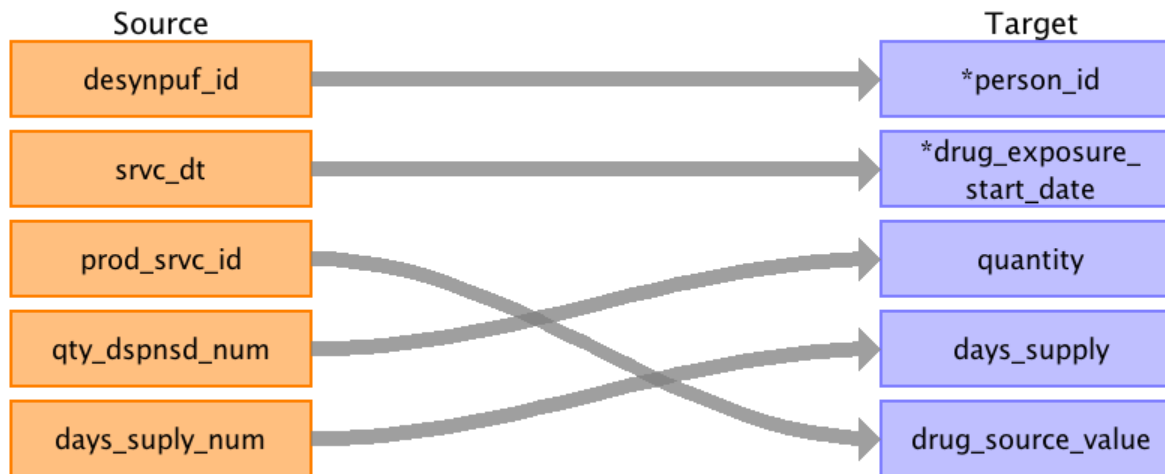
WhiteRabbit

- 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





OHDSI Has Extensive Vocabulary Maps

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 Gemscrip	Gemscrip 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)

Athena

Additional Vocabulary Support

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

Overview Table

The screenshot displays the Usagi application interface, which is used for mapping non-standard vocabularies to standard ones. The interface is divided into three main sections:

- Overview Table:** A table showing a list of source terms and their corresponding target concepts. The table has columns for Status, Source code, Source term, Frequency, Dutch term, Match score, Concept ID, Concept name, Domain, Concept class, Vocabulary, and Concept code.
- Selected Mapping:** A detailed view of a specific mapping, showing the source code (A), source term (General and un...), frequency (25), and target concepts (generalized, 4244571, Generalized, Observation, Qualifier Value, SNOMED, 60132005).
- Search Facility:** A search interface with a query field, filters (Filter by automatically select concepts, Filter by concept class, Filter by domain, Filter invalid concepts, Filter by vocabulary), and a results table. The results table has columns for Score, Synonym, Concept ID, Concept name, Domain, Concept class, Vocabulary, Concept code, Valid start date, Valid end date, and Invalid reason.

The interface also includes a search query field, filters, and a results table. The search query field is currently empty. The filters section includes options for "Filter by automatically select concepts", "Filter by concept class" (set to "Admin Concept"), "Filter by domain" (set to "Condition"), "Filter invalid concepts", and "Filter by vocabulary" (set to "APC"). The results table shows a list of search results with columns for Score, Synonym, Concept ID, Concept name, Domain, Concept class, Vocabulary, Concept code, Valid start date, Valid end date, and Invalid reason.

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
-



Getting Value from Your Data

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



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

OPTUM

Drug Era Report

Drug Prevalence

Treemap **Table**

BLOOD AND BLOOD FORMING ORGANS
ANTITHROMBOTIC AGENTS
VITAMIN K ANTAGONISTS

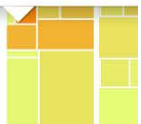


Warfarin

Prevalence: 0.91%

Number of People: **Warfarin**

Length of Era: 193

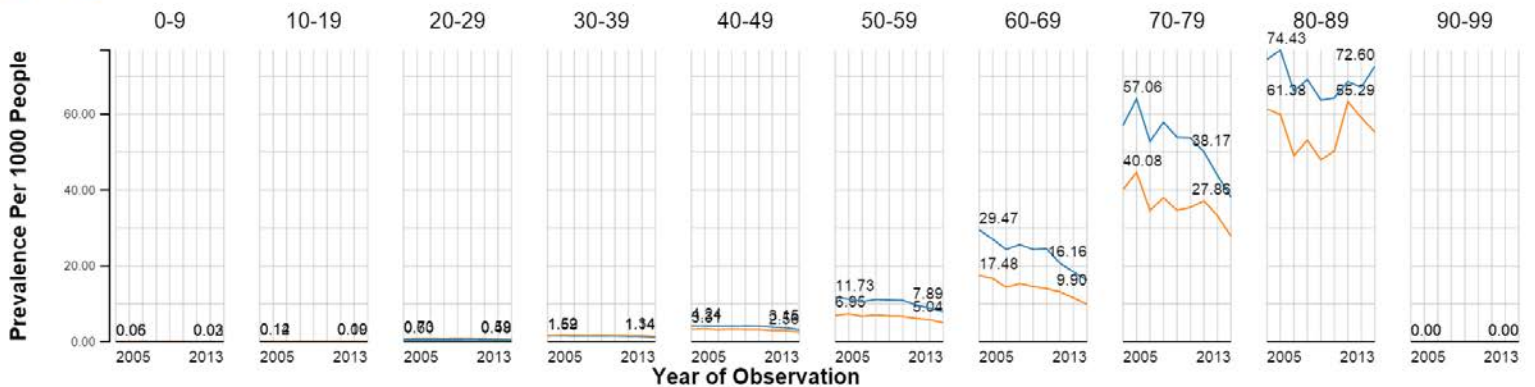


Box Size: Prevalence

Drug Prevalence

MALE FEMALE

Age Decile





ACHILLES Report Types

Achilles Data Sources ▾ Reports ▾

SynPUF 1K
Person

Person Summary

Source name: Demo data - 1K synthetic patients

Number of persons: 1k

Population by Gender

- FEMALE
- MALE

Year

Population

- Dashboard
- Achilles Heel
- Person
- Observation Periods
- Data Density
- Conditions
- Condition Eras
- Observations
- Drug Eras
- Drug Exposures
- Procedures
- Visits
- Death

Year

Population by Ethnicity

- Hispanic or Latino
- No matching concept
- Not Hispanic or Latino

- Dashboard
- Achilles Heel
- Person
- Observation Periods
- Data Density
- Conditions
- Condition Eras
- Observations
- Drug Eras
- Drug Exposures
- Procedures
- Visits
- Death



ACHILLES Heel Helps You Validate Your Data Quality

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



Why Data Quality?

- Fitness for analysis, trust in outputs, completeness of data
- Data transformation: Source -> Target
- Errors in data:
 - Source error (typo in birth year; no pattern)
 - ETL error (has pattern)
 - Mapping error
- Common Data Models allows sharing of data quality rules and creating of data quality tools
- Existence of data quality tools allows sites to quickly implement a starter set of rules



Achilles Heel (your free data quality tool)

- Achilles (step 1 of 2)
 - Pre-computed measures (Achilles.sql)
- Achilles Heel (step 2 of 2)
 - Data quality rules (AchillesHeel.sql)
- Achilles Web
 - Web-based “data viewer”
- Paradigm:
Patient level data -> “something smaller”
(10B rows) (2M rows)



OHDSI / Achilles

Watch 51

Star 29

Code

Issues 25

Pull requests 1

Wiki

Pulse

Graphs

Branch: master

Achilles / inst / sql / sql_server / AchillesHeel_v5.sql

Find file

aaron0browne Add separate vocabulary schema argument

3f28b7d 12

6 contributors



719 lines (668 sloc) | 22.1 KB

Raw

Blame

History




```

1 /*****
2
3 # @file ACHILLESHEEL.SQL
4 #
5 # Copyright 2014 Observational Health Data Sciences and Informatics
6 #
7 # This file is part of ACHILLES

```

Branch: master ▾

Achilles / inst / csv / achilles_rule.csv

 vojtechuser typo from 2.0 fixed to 1.2

1 contributor

29 lines (28 sloc) | 1.85 KB

Raw

Blame

🔍 Search this file...

	rule_id	rule_name	severity	rule_description
1				
2	0	Achilles Heel version 1.2		this rule is not used for data analysis. It communicates th
3	1	multiple checks	error	multiple error checks
4	2	multiple checks	error	distributions where min should not be negative
5	3	multiple checks	warning	death distributions where max should not be positive
6	4	invalid concept_id	error	invalid concept_id
7	5	invalid type concept_id	error	invalid type concept_id
8	6	concept from the wrong vocabulary	error	concepts from wrong vocabulary 12 HL7
9	7	concept from the wrong vocabulary	error	concept from the wrong vocabulary
10	8	concept from the wrong vocabulary; race	error	concept from the wrong vocabulary; race
11	9	concept from the wrong vocabulary; ethnicity	error	concept from the wrong vocabulary; ethnicity
12	10	concept from the wrong vocabulary; place of service	error	concept from the wrong vocabulary; place of service

vojtechuser typo from 2.0 fixed to 1.2

1 contributor

29 lines (28 sloc) | 1.85 KB

Raw Blame

Search this file...

	rule_id	rule_name	severity	rule_description
2	0	Achilles Heel version 1.2		this rule is not used for data analysis. It communicates th
3	1	multiple checks	error	multiple error checks
4	2	multiple checks	error	distributions where min should not be negative
5	3	multiple checks	warning	death distributions where max should not be positive
6	4	invalid concept_id	error	invalid concept_id
7	5	invalid type concept_id	error	invalid type concept_id
8	6	concept from the wrong vocabulary	error	concepts from wrong vocabulary 12 HL7
9	7	concept from the wrong vocabulary	error	concept from the wrong vocabulary
10	8	concept from the wrong vocabulary; race	error	concept from the wrong vocabulary; race
11	9	concept from the wrong vocabulary; ethnicity	error	concept from the wrong vocabulary; ethnicity
12	10	concept from the wrong vocabulary; place of service	error	concept from the wrong vocabulary; place of service

20	18	year of birth is in the future	error	year of birth should not be in the future
21	19	year of birth is prior 1800	warning	year of birth < 1800
22	20	age below 0	error	age < 0
23	21	age too high	error	age > 150
24	22	monthly trend	warning	monthly change > 100%
25	23	monthly trend	warning	monthly change > 100% at concept level
26	24	too high days_supply	warning	days_supply > 180
27	25	too high number of refills	warning	refills > 10
28	26	implausible quantity for drug	warning	quantity > 600





Step 1 Pre-computed analyses

ANALYSIS_ID	ANALYSIS_NAME	STRATUM_1_NAME	STRATUM_2_NAME	STRATUM_3_NAME	STRATUM_4_NAME	STRATUM_5_NAME
0	Source name	NA	NA	NA	NA	NA
1	Number of persons	NA	NA	NA	NA	NA
2	Number of persons by gender	gender_concept_id	NA	NA	NA	NA
3	Number of persons by year of birth	year_of_birth	NA	NA	NA	NA
4	Number of persons by race	race_concept_id	NA	NA	NA	NA
5	Number of persons by ethnicity	ethnicity_concept_id	NA	NA	NA	NA
7	Number of persons with invalid provider_id	NA	NA	NA	NA	NA
8	Number of persons with invalid location_id	NA	NA	NA	NA	NA
9	Number of persons with invalid care_site_id	NA	NA	NA	NA	NA
101	Number of persons by age, with age at first observation period	age	NA	NA	NA	NA
102	Number of persons by gender by age, with age at first observation period	gender_concept_id	age	NA	NA	NA
103	Distribution of age at first observation period	NA	NA	NA	NA	NA
104	Distribution of age at first observation period by gender	gender_concept_id	NA	NA	NA	NA
105	Length of observation (days) of first observation period	NA	NA	NA	NA	NA
106	Length of observation (days) of first observation period by gender	gender_concept_id	NA	NA	NA	NA
107	Length of observation (days) of first observation period by age decile	age decile	NA	NA	NA	NA
108	Number of persons by length of observation period, in 30d increments	Observation period length	NA	NA	NA	NA
109	Number of persons with continuous observation in each year	calendar year	NA	NA	NA	NA
110	Number of persons with continuous observation in each month	calendar month	NA	NA	NA	NA
111	Number of persons by observation period start month	calendar month	NA	NA	NA	NA
112	Number of persons by observation period end month	calendar month	NA	NA	NA	NA
113	Number of persons by number of observation periods	number of observation periods	NA	NA	NA	NA
114	Number of persons with observation period before year-of-birth	NA	NA	NA	NA	NA
115	Number of persons with observation period end < observation period start	NA	NA	NA	NA	NA
116	Number of persons with at least one day of observation in each year	calendar year	gender_concept_id	age decile	NA	NA
117	Number of persons with at least one day of observation in each month	calendar month	NA	NA	NA	NA



Drug quantity by drug ID

ANALYSIS_ID	ANALYSIS_NAME	STRATUM_1_NAME	STRATUM_2_NAME	STRATUM_3_NAME	STRATUM_4_NAME	STRATUM_5_NAME
701	Number of drug exposure records, by drug_concept_id	drug_concept_id	NA	NA	NA	NA
702	Number of persons by drug exposure start month, by drug_concept_id	drug_concept_id	calendar month	NA	NA	NA
703	Number of distinct drug exposure concepts per person	NA	NA	NA	NA	NA
704	Number of persons with at least one drug exposure, by drug_concept_id	drug_concept_id	calendar year	gender_concept_id	age decile	NA
705	Number of drug exposure records, by drug_concept_id by drug_type_concept_id	drug_concept_id	drug_type_concept_id	NA	NA	NA
706	Distribution of age by drug_concept_id	drug_concept_id	gender_concept_id	NA	NA	NA
709	Number of drug exposure records with invalid person_id	NA	NA	NA	NA	NA
710	Number of drug exposure records outside valid observation period	NA	NA	NA	NA	NA
711	Number of drug exposure records with end date < start date	NA	NA	NA	NA	NA
712	Number of drug exposure records with invalid provider_id	NA	NA	NA	NA	NA
713	Number of drug exposure records with invalid visit_id	NA	NA	NA	NA	NA
715	Distribution of days_supply by drug_concept_id	drug_concept_id	NA	NA	NA	NA
716	Distribution of refills by drug_concept_id	drug_concept_id	NA	NA	NA	NA
717	Distribution of quantity by drug_concept_id	drug_concept_id	NA	NA	NA	NA
720	Number of drug exposure records by drug exposure start month	calendar month	NA	NA	NA	NA
800	Number of persons with at least one observation occurrence, by observation_concept_id	observation_concept_id	NA	NA	NA	NA
801	Number of observation occurrence records, by observation_concept_id	observation_concept_id	NA	NA	NA	NA
802	Number of persons by observation occurrence start month, by observation_concept_id	observation_concept_id	calendar month	NA	NA	NA
803	Number of distinct observation occurrence concepts per person	NA	NA	NA	NA	NA
804	Number of persons with at least one observation occurrence, by observation_concept_id	observation_concept_id	calendar year	gender_concept_id	age decile	NA
805	Number of observation occurrence records, by observation_concept_id	observation_concept_id	observation_type_concept_id	NA	NA	NA
806	Distribution of age by observation_concept_id	observation_concept_id	gender_concept_id	NA	NA	NA
807	Number of observation occurrence records, by observation_concept_id	observation_concept_id	unit_concept_id	NA	NA	NA
809	Number of observation records with invalid person_id	NA	NA	NA	NA	NA
810	Number of observation records outside valid observation period	NA	NA	NA	NA	NA
812	Number of observation records with invalid provider_id	NA	NA	NA	NA	NA



What is new? (Achilles Heel v1.2; March 2016)

- Introduction of RULE_ID and rule overview CSV file
- Better reporting of “depth of the error” (number of rows with a given error)
- Support for CDM v5
- Generalizability to other CDMs
 - Separation of model-conformance rules from rules examining “source” data (zombie events)
 - Data measure vs. data quality measure; target model terminology (RxNorm)
- More rules (contribute your favorite DQ rule); non-Achilles efforts (IRIS)



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



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



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

HERMES

warfarin



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)	Geneseqno (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)	Geneseqno (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) RxNorm to Multilex equivalent (OMOP) (2)	Has tradename (RxNorm) (7) 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

Cohort Definition List

Help

Index Population: MiniSentinel replication - warfarin new users

Save

Description:

Expression

Concept Sets

Print Friendly

Raw JSON

Generate

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

a drug era of

Add Filter...

Delete Filter

for the first time in the person's history

era start is:

with age at era start

with observation at least days prior and days after index

Limit primary events to: per person.

Add Additional Filters

Limit cohort expression results to: per person.

Show SQL

Add Options



— People having any of the following: **Add Primary Criteria...** ▼

a condition occurrence of **Delivery** ▼

Add Criterion... ▼

Delete

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

X with age **Between** ▼ 18 and 55

X with a gender of: **X FEMALE** **Add** **Import**

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:

Add Criterion... ▼

a condition occurrence of **Depression** ▼

occurring between **0** ▼ days **Before** ▼ and **180** ▼ days **After** ▼ index

Delete Criteria

and with **At Most** ▼ **0** ▼ occurrences of:

Add Criterion... ▼

a condition occurrence of **Depression** ▼

occurring between **All** ▼ days **Before** ▼ and **0** ▼ days **After** ▼ index

Delete Criteria



Cohort Creation vs Analysis

- Cohorts may be designed and stored and shared
- Choice of tools to visualize and analyze
- Cohort visualization is performed using [Heracles](#)



HERACLES: Characterize the cohorts of interest

OHDSI Heracles

«Back

Refresh

Truven MDCD (APS) ▼

Heracles Runner

Cohort Specific

Condition

Condition Eras

Conditions by Index

Dashboard

Data Density

Death

Drug Eras

Drug Exposures

Drugs by Index

Heracles Heel

Drug Exposures

Drugs by Index

Heracles Heel

Measurements

Observation Periods

Observations

Person

Procedures

Procedures by Index

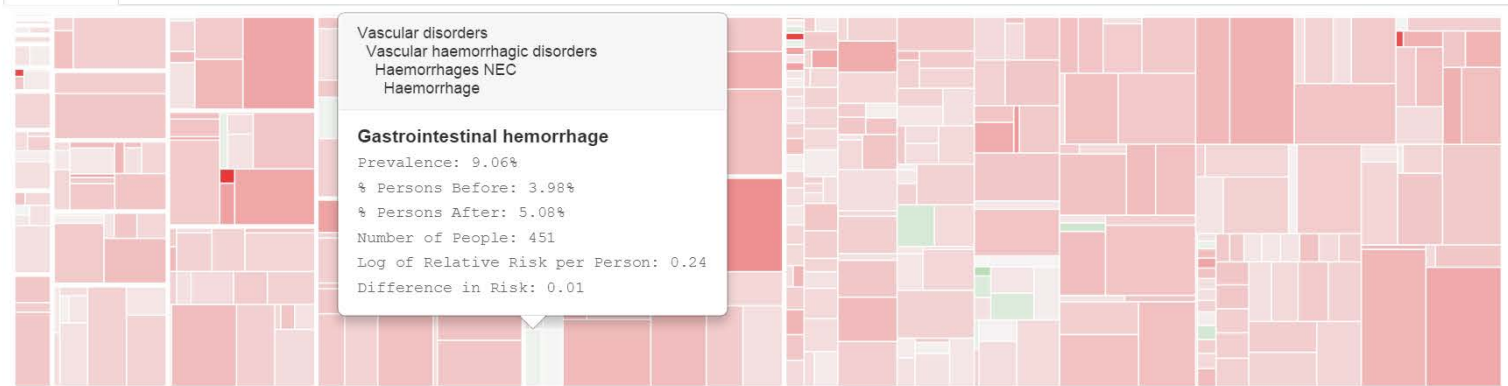
Visits

Matching Population: MiniSentinel replication - warfarin new users

Condition Prevalence

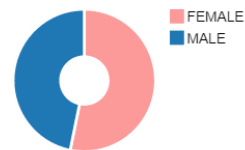
Treemap

Table

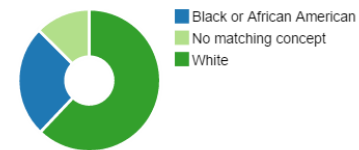


Box Size: Prevalence, Color: Log of Relative Risk (Red to Green = Negative to Positive), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

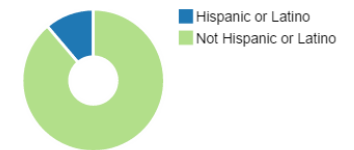
Population by Gender ↓



Population by Race ↓



Population by Ethnicity ↓





HERACLES

Heracles

Analysis Viewer

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

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

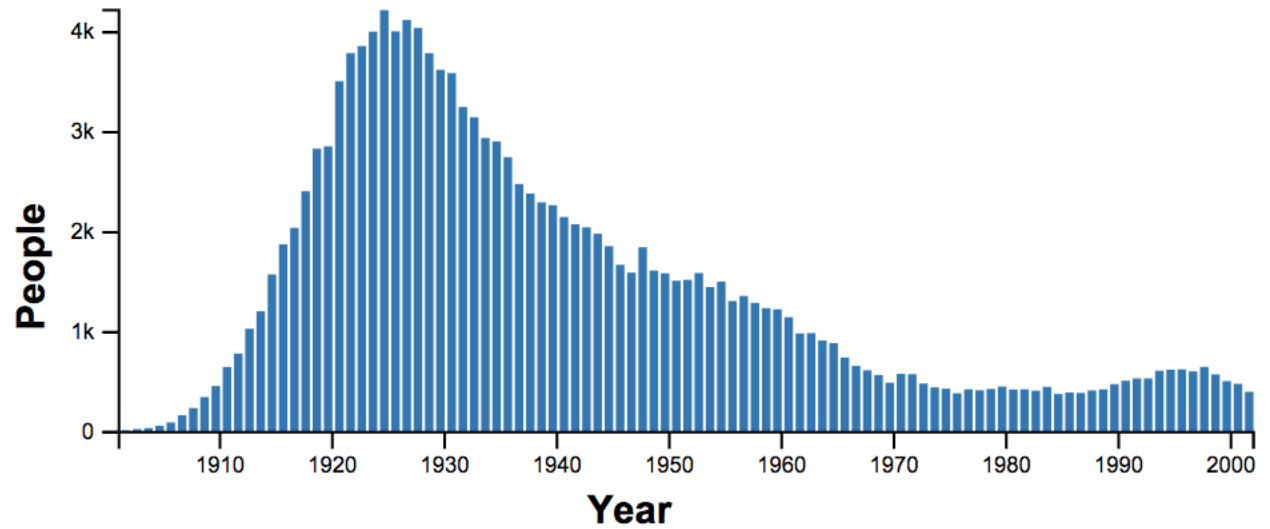


Alzheimers

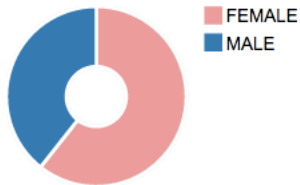
Source: INPC

Number of Persons:
145,246

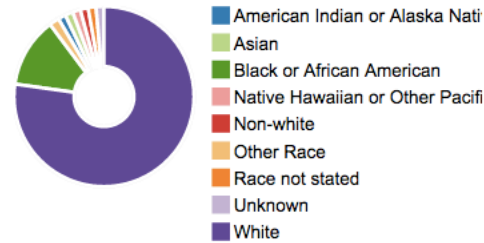
Year of Birth



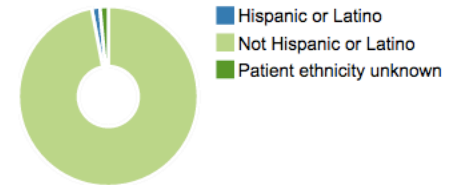
Population by Gender



Population by Race



Population by Ethnicity





OHDSI Heracles

«Back

Refresh

Heracles Runner

Dashboard

Cohort Specific

Heracles Heel

Person

Observation Periods

Data Density

Condition

Condition Eras

Observations

Drug Eras

Drug Exposures

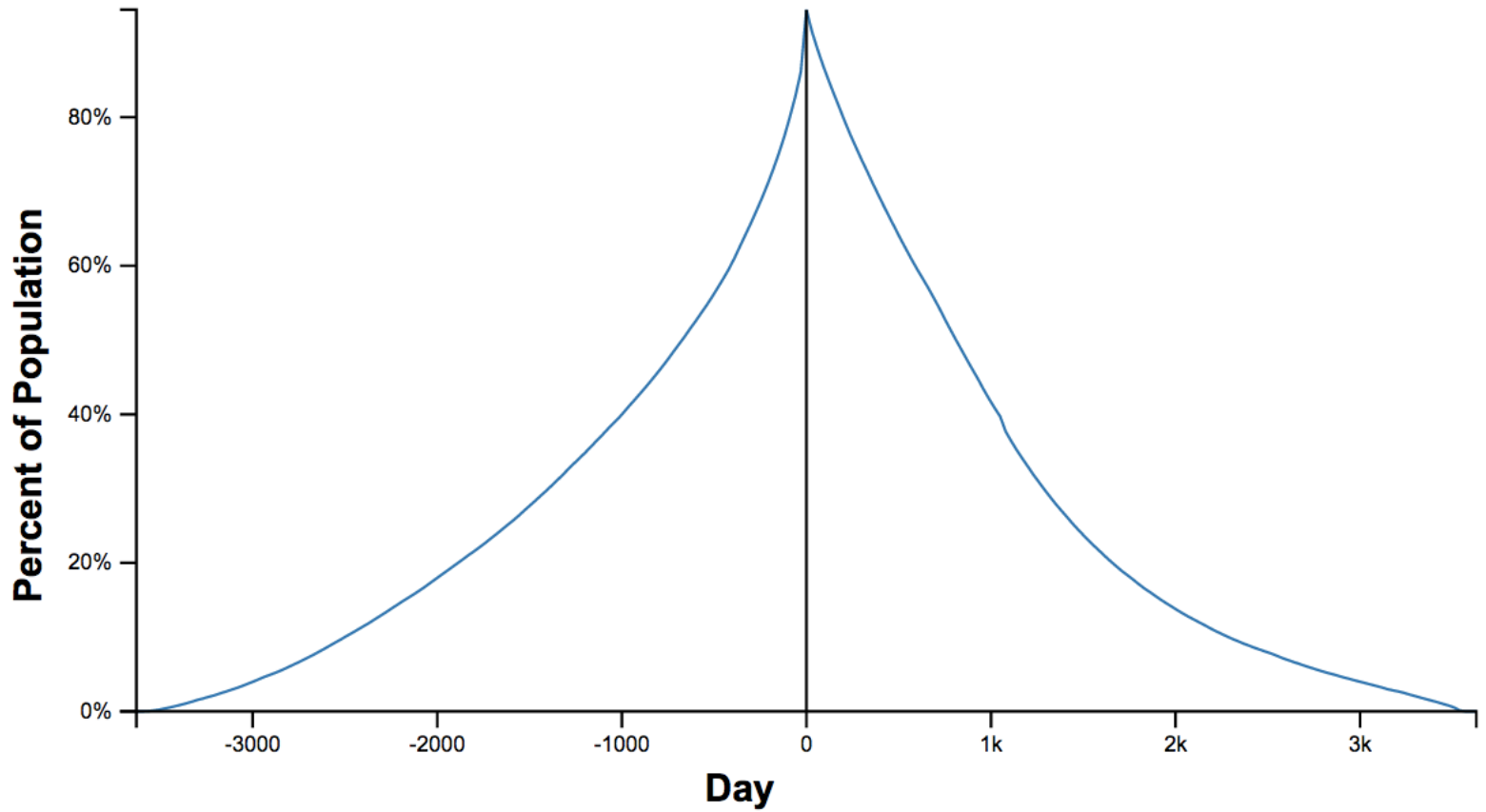
Procedures

Visits

Death

Alzheimers

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



Alzheimers

Condition Prevalence

Treemap

Table

Search:

Show / hide columns

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)

Previous 2 3 4 5 ... 9 Next

Conditions

Condition Prevalence

Treemap

Table

Search:

Show / hide columns

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)

Previous 2 3 4 5 ... 10 Next



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: Impact of Study Inclusion Criteria in Clinical Trials

Index Rule
Inclusion Rules
Concept Sets
Results

Source	Name	Dialect	
<input type="radio"/>	TRUVENCCA	Truven CCAE (APS)	pdw Generate
<input type="radio"/>	TRUVENMDCR	Truven MDCR (APS)	pdw Generate
<input type="radio"/>	TRUVENMDCD	Truven MDCD (APS)	pdw Generate
<input checked="" type="radio"/>	OPTUM	Optum (APS)	pdw Generate
<input type="radio"/>	CPRD	CPRD (APS)	pdw Generate
<input type="radio"/>	PREMIER	Premier (APS)	pdw Generate
<input type="radio"/>	JMDC	JMDC (APS)	pdw Generate
<input type="radio"/>	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 valve or mitral stenosis		94.99%	2.23%
6. No dialysis in last 30 days		98.97%	0.39%
7. No history of kidney transplant		99.61%	0.06%
8. Not at long-term care visit		97.29%	0.70%

Population Visualization



Open-source large-scale analytics through R (and C, CUDA)

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

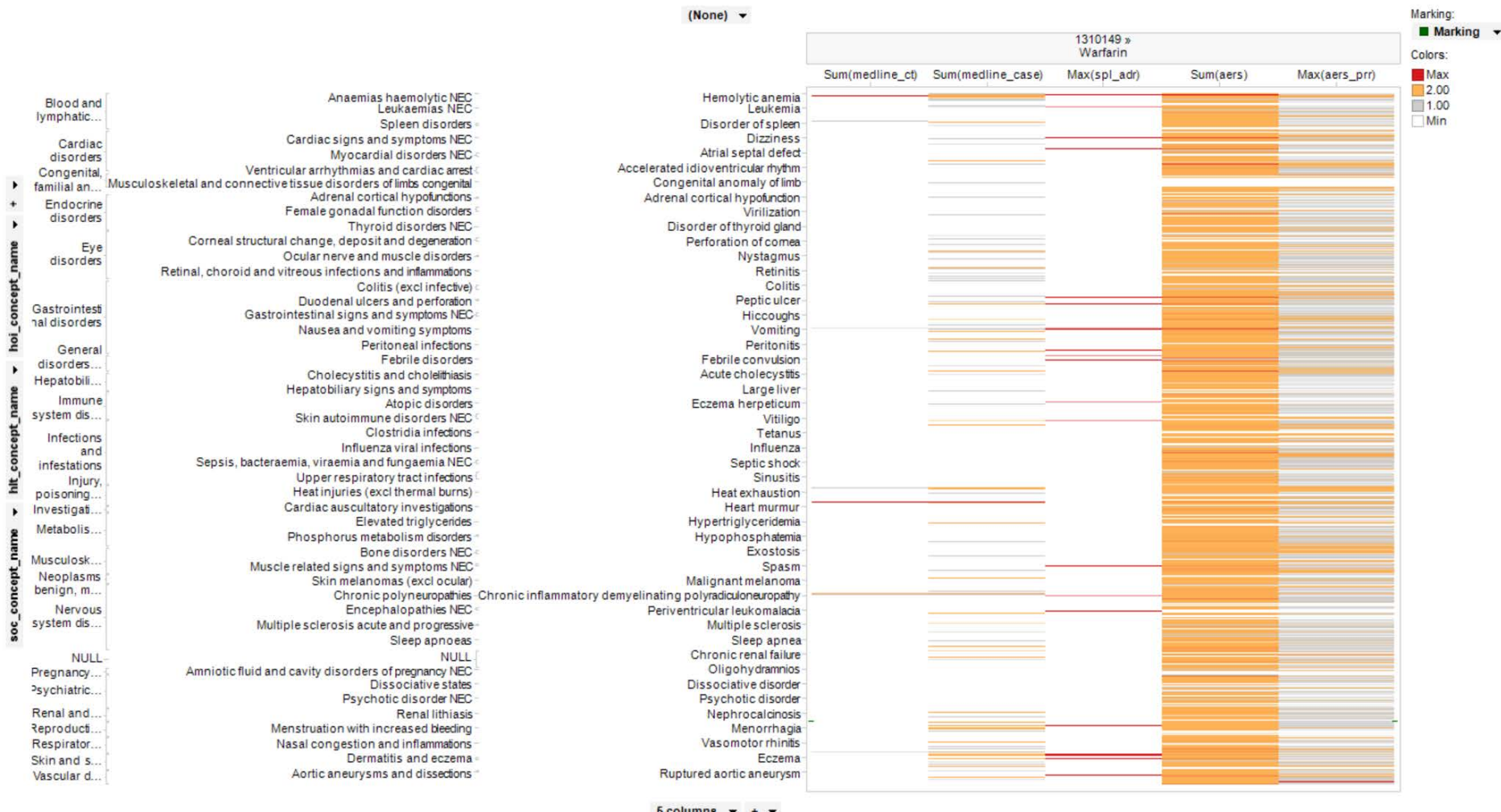
NeedsCompilation yes

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’

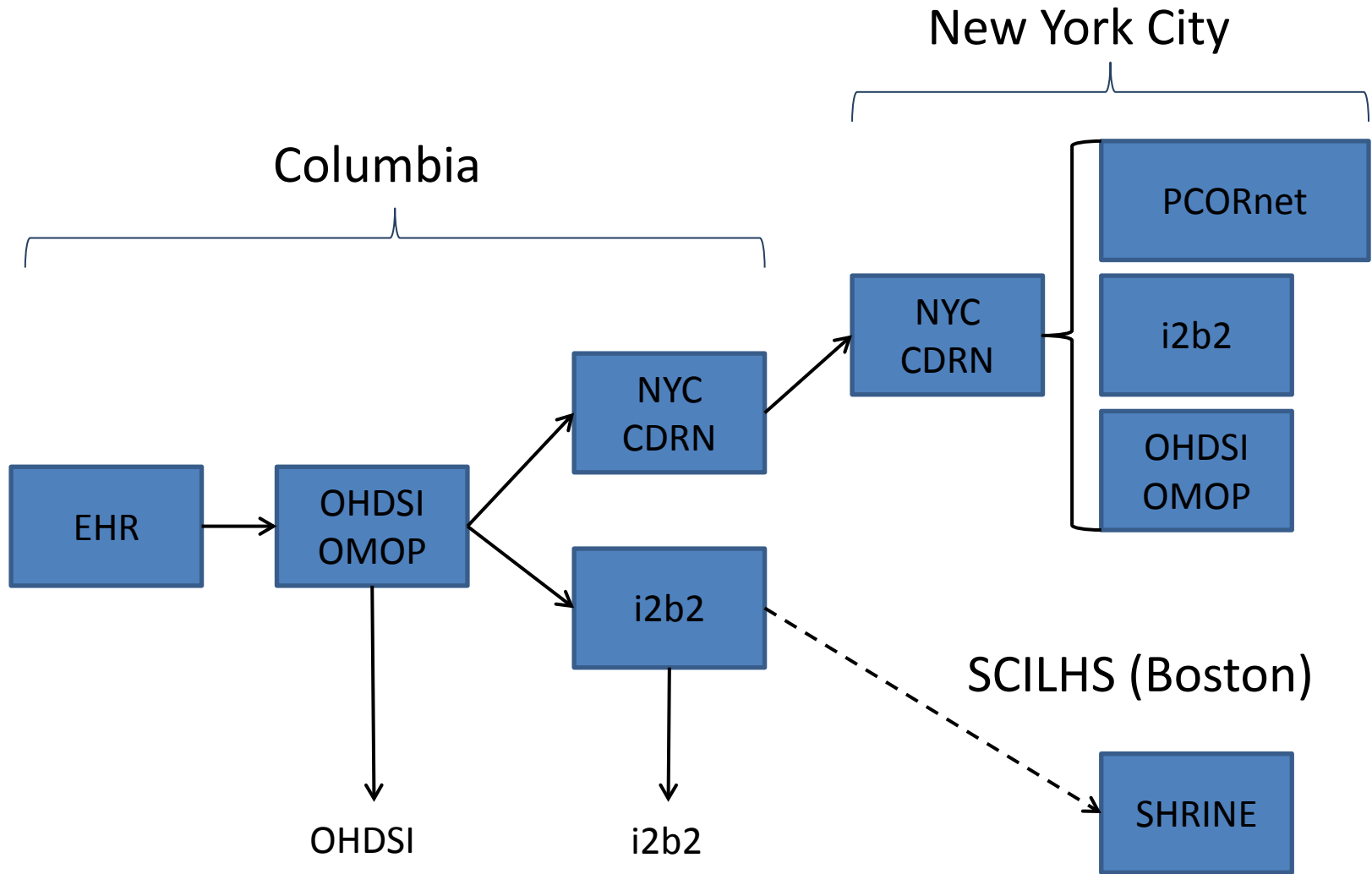
LAERTES: Summarizing evidence from existing data sources: literature, labeling, spontaneous reporting

LAERTES Evidence Map





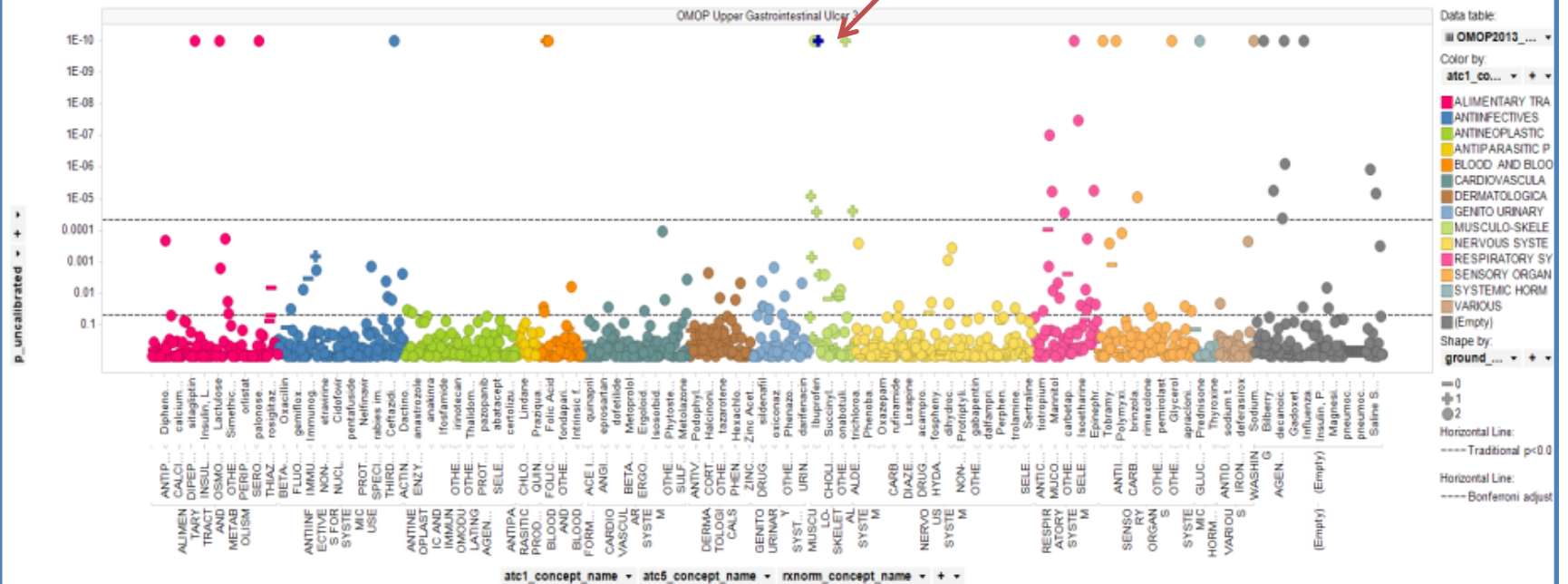
Columbia data network approach



OHDSI answers questions

Naproxen has one of the most significant associations with GI bleed, along with other NSAIDs

Medication-Wide Association Study (MWAS) for specificity and analogy



MWAS Details

rxnorm_concept_name	atc5_concept_name	atc3_concept_name	atc1_concept_name	num_persons	num_outcomes_exp...	irr	irr1b95	irr95	p_uncalibrated
Naproxen	ANTIINFLAMMATORY P...	TOPICAL PRODUCTS ...	MUSCULO-SKELETAL SYSTEM	341843	3783	1.38	1.33	1.43	0.0000

Explore all drugs for a given outcome



OHDSI in Action

- **Generate evidence**
 - Randomized trial is the gold standard
 - Observational research seen as supporting



Observational Data & Clinical Trials

- Sample size calculations
 - Do we have enough patients to carry out a trial?
- Recruitment
 - Find patients or their clinicians from EHRs
- Pragmatic trials: recruitment and data collection
 - ADAPTABLE aspirin trial
- ...
- Complementary causal evidence (future)
 - New methods to handle confounding and ascertain causes from retrospective observational databases



Characterization

- Today we carry out RCTs without clear knowledge of actual practice
- There will be no RCTs without an observational precursor
 - It will be required to characterize a population using large-scale observational data before designing an RCT
 - Disease burden
 - Actual treatment practice
 - Time on therapy
 - Course and complication rate
 - Done now somewhat through literature and pilot studies



Treatment Pathways

Global stakeholders

Public

Academics

Industry

Regulator

Evidence

RCT, Obs

Conduits

Social media

Lay press

Literature

Guidelines

Advertising

Formulary

Labels

Inputs

Indication

Feasibility

Cost

Preference

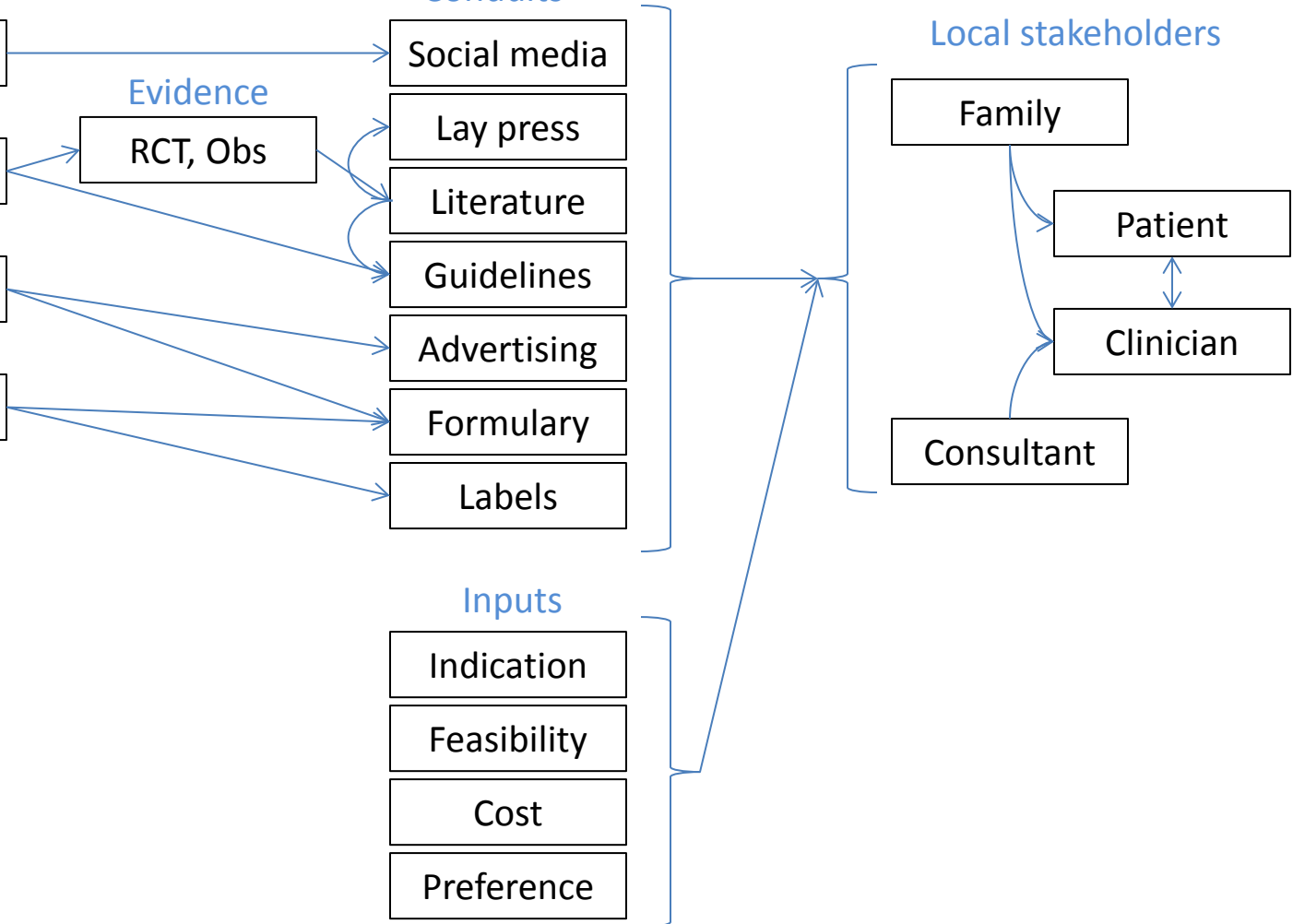
Local stakeholders

Family

Patient

Clinician

Consultant





Network process

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



OHDSI in action: Chronic disease treatment pathways

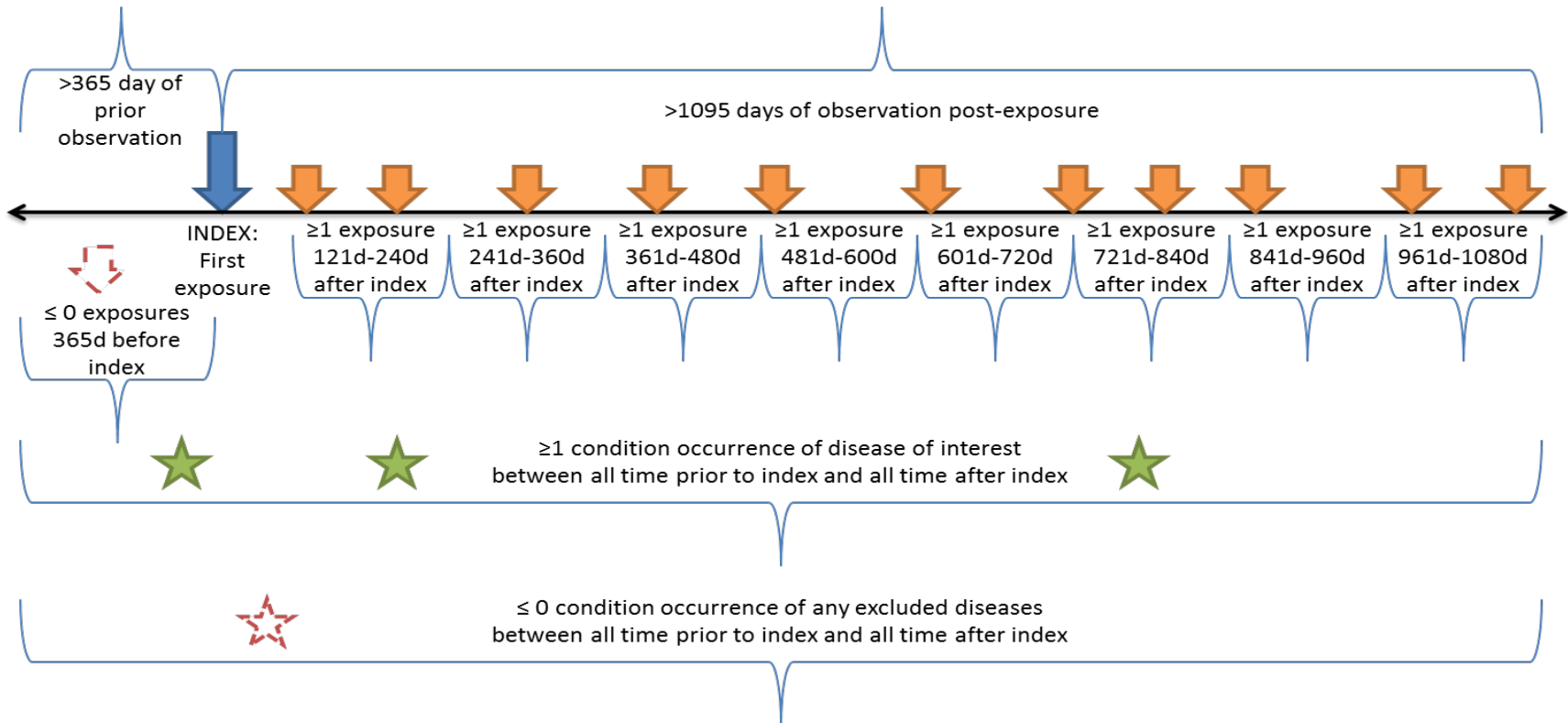
- Conceived at AMIA 15Nov2014
 - Protocol written, code written and tested at 2 sites 30Nov2014
 - Analysis submitted to OHDSI network 2Dec2014
 - Results submitted for 7 databases 5Dec2014
-



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)

Treatment pathway event flow





OHDSI participating data partners

Abbreviation	Name	Description	Population, millions
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



Characterizing treatment pathways at scale using the OHDSI network

George Hripcsak^{a,b,c,1}, Patrick B. Ryan^{c,d}, Jon D. Duke^{c,e}, Nigam H. Shah^{c,f}, Rae Woong Park^{c,g}, Vojtech Huser^{c,h}, Marc A. Suchard^{c,i,j,k}, Martijn J. Schuemie^{c,d}, Frank J. DeFalco^{c,d}, Adler Perotte^{a,c}, Juan M. Banda^{c,l}, Christian G. Reich^{c,l}, Lisa M. Schilling^{c,m}, Michael E. Matheny^{c,n,o}, Daniella Meeker^{c,p,q}, Nicole Pratt^{c,r}, and David Madigan^{c,s}

^aDepartment of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032; ^bMedical Informatics Services, New York-Presbyterian Hospital, New York, NY 10032; ^cObservational Health Data Sciences and Informatics, New York, NY 10032; ^dEpidemiology Analytics, Janssen Research and Development, Titusville, NJ 08560; ^eCenter for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN 46205; ^fCenter for Biomedical Informatics Research, Stanford University, CA 94305; ^gDepartment of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea, 443-380; ^hLister Hill National Center for Biomedical Communications (National Library of Medicine), National Institutes of Health, Bethesda, MD 20894; ⁱDepartment of Biomathematics, University of California, Los Angeles, CA 90095; ^jDepartment of Biostatistics, University of California, Los Angeles, CA 90095; ^kDepartment of Human Genetics, University of California, Los Angeles, CA 90095; ^lReal World Evidence Solutions, IMS Health, Burlington, MA 01809; ^mDepartment of Medicine, University of Colorado School of Medicine, Aurora, CO 80045; ⁿDepartment of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212; ^oGeriatric Research, Education and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, TN 37212; ^pDepartment of Preventive Medicine, University of Southern California, Los Angeles, CA 90089; ^qDepartment of Pediatrics, University of Southern California, Los Angeles, CA 90089; ^rDivision of Health Sciences, University of South Australia, Adelaide, SA, Australia 5001; and ^sDepartment of Statistics, Columbia University, New York, NY 10027

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 5, 2016 (received for review June 14, 2015)

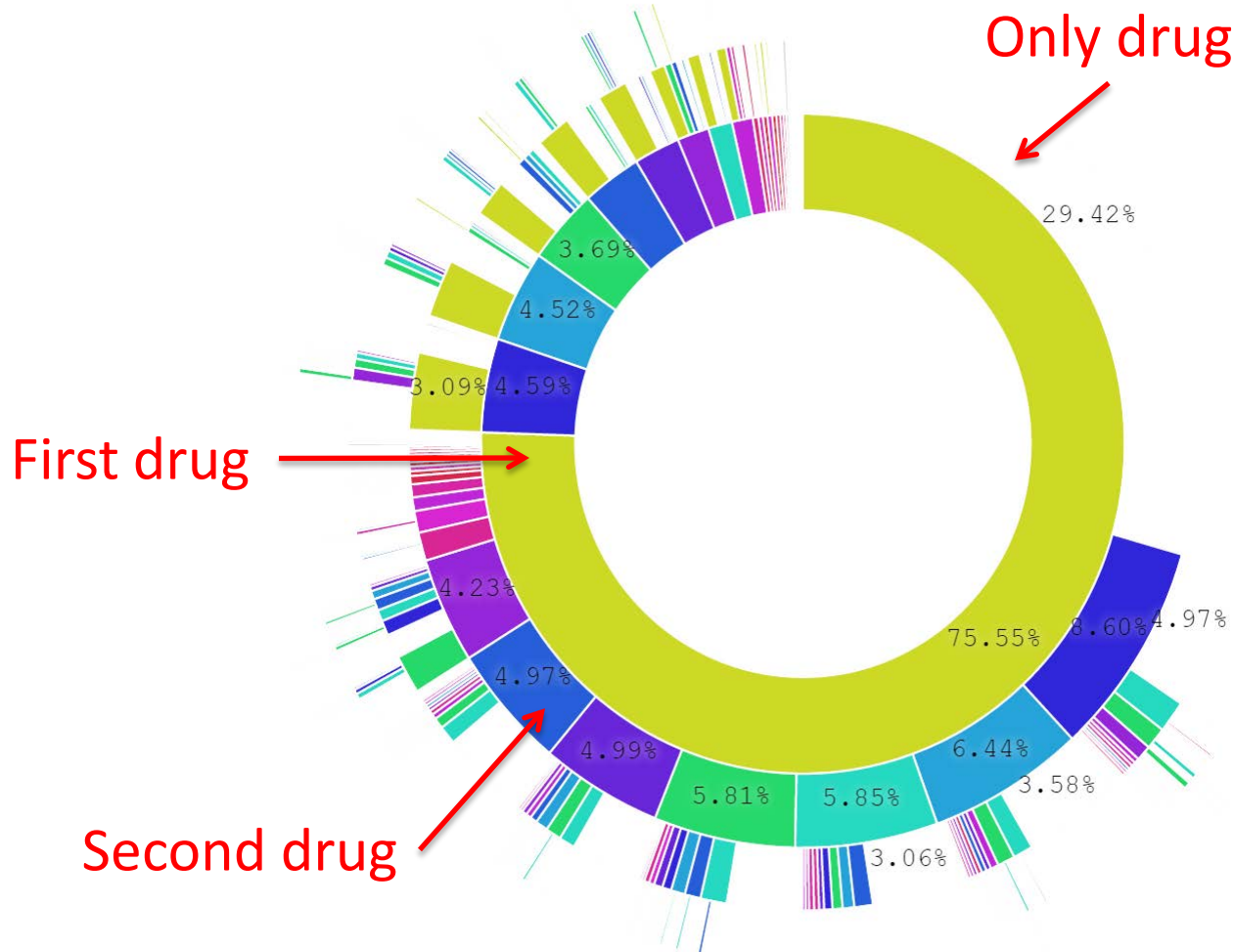
Observational research promises to complement experimental research by providing large, diverse populations that would be infeasible for an experiment. Observational research can test its own clinical hypotheses, and observational studies also can contribute to the design of experiments and inform the generalizability of experimental research. Understanding the diversity of populations

Without sufficiently broad databases available in the first stage, randomized trials are designed without explicit knowledge of actual disease status and treatment practice. Literature reviews are restricted to the population choices of previous investigations, and pilot studies usually are limited in scope. By exploiting the ClinicalTrials.gov national trial registry (9) and electronic health



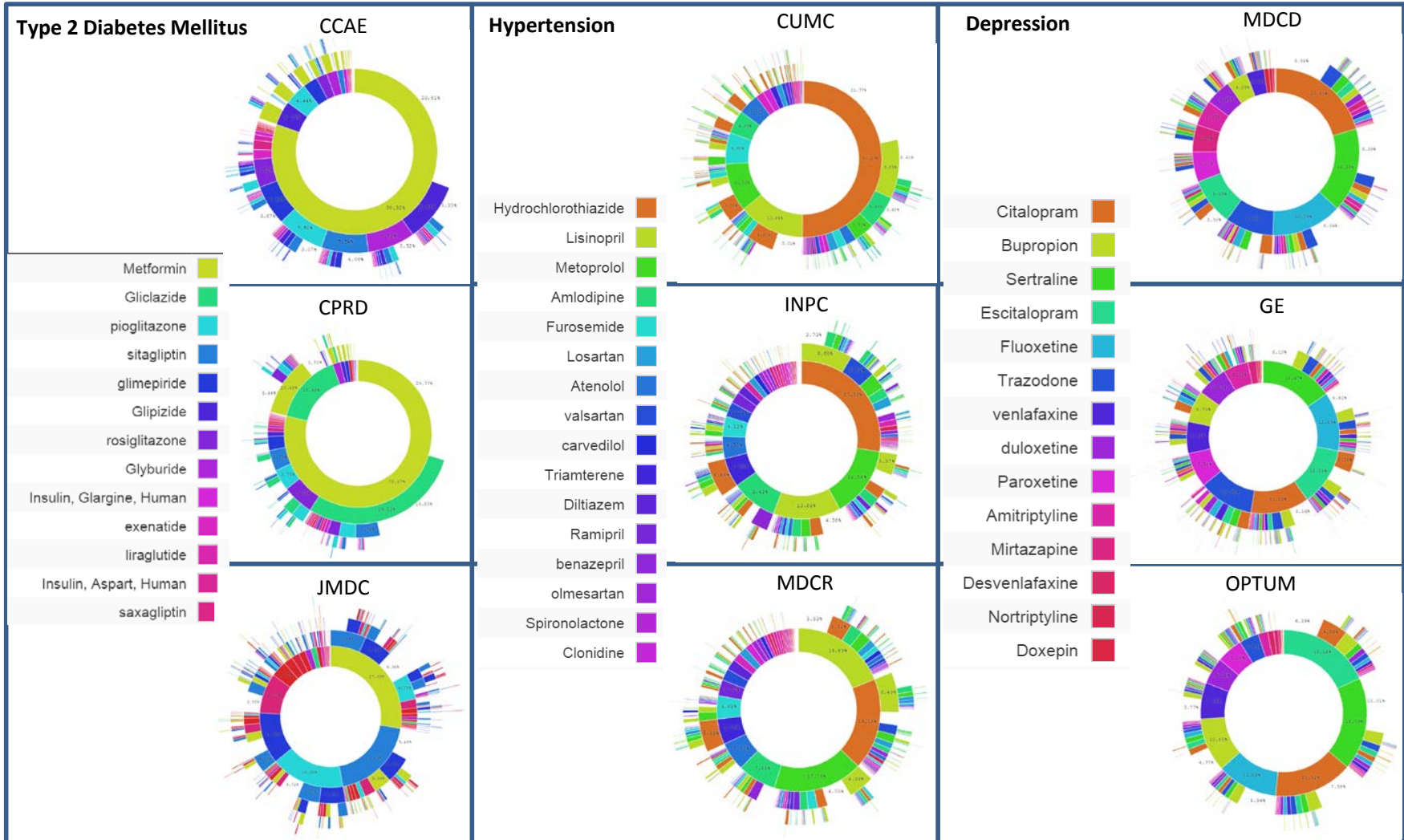
Treatment pathways for diabetes

T2DM : All databases



Metformin	
pioglitazone	
sitagliptin	
Glipizide	
glimepiride	
Gliclazide	
Glyburide	
rosiglitazone	
Insulin, Glargine, Human	
exenatide	
Insulin, Aspart, Human	
liraglutide	
saxagliptin	
Insulin, Lispro, Human	
Glucose	
Insulin, Isophane, Human	

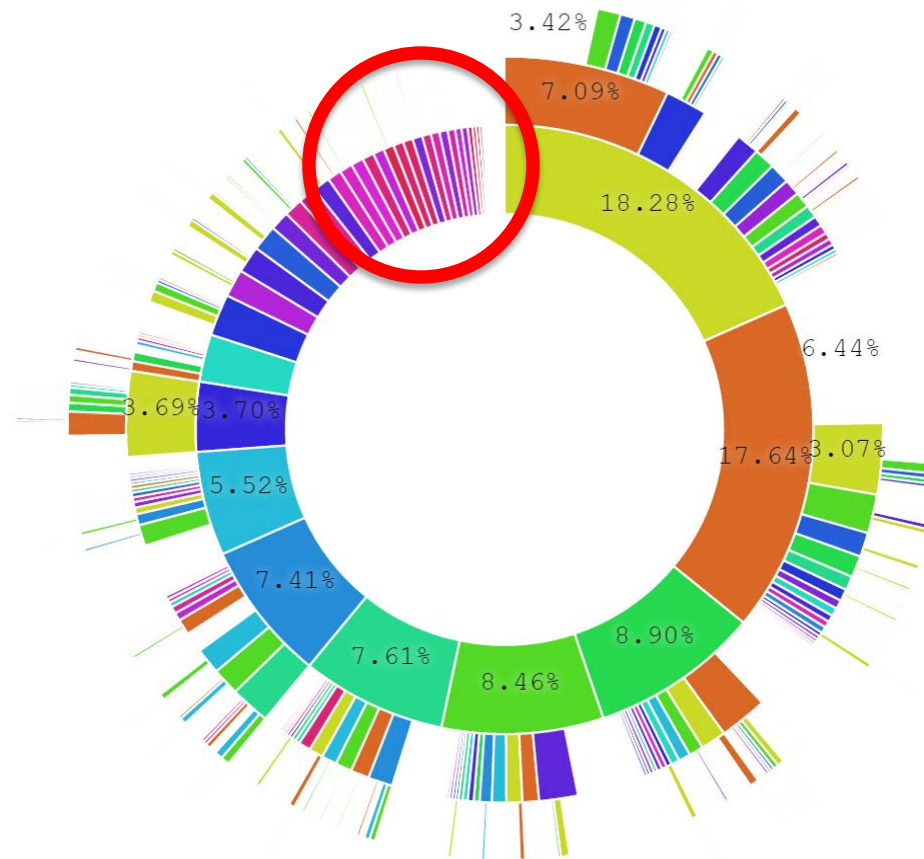
Population-level heterogeneity across systems, and patient-level heterogeneity within systems





Patient-level heterogeneity

HTN: All databases

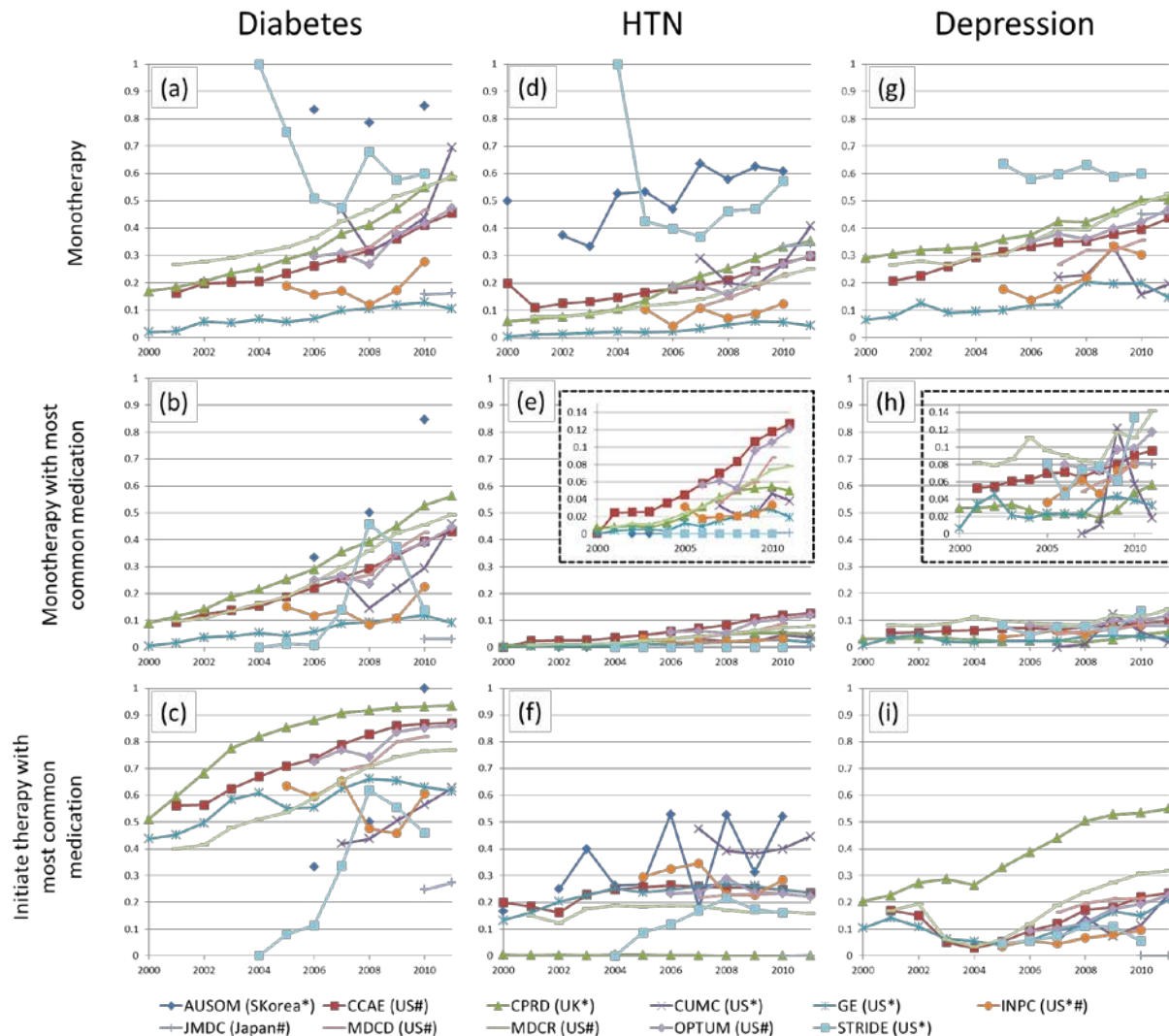


- Lisinopril
- Hydrochlorothiazide
- Amlodipine
- Metoprolol
- Atenolol
- Furosemide
- Ramipril
- Bendroflumethiazide
- Losartan
- valsartan
- Triamterene
- olmesartan
- benazepril
- Diltiazem
- carvedilol
- Bisoprolol
- Doxazosin
- Enalapril

25% of HTN patients (10% of others) have a unique path despite 250M pop



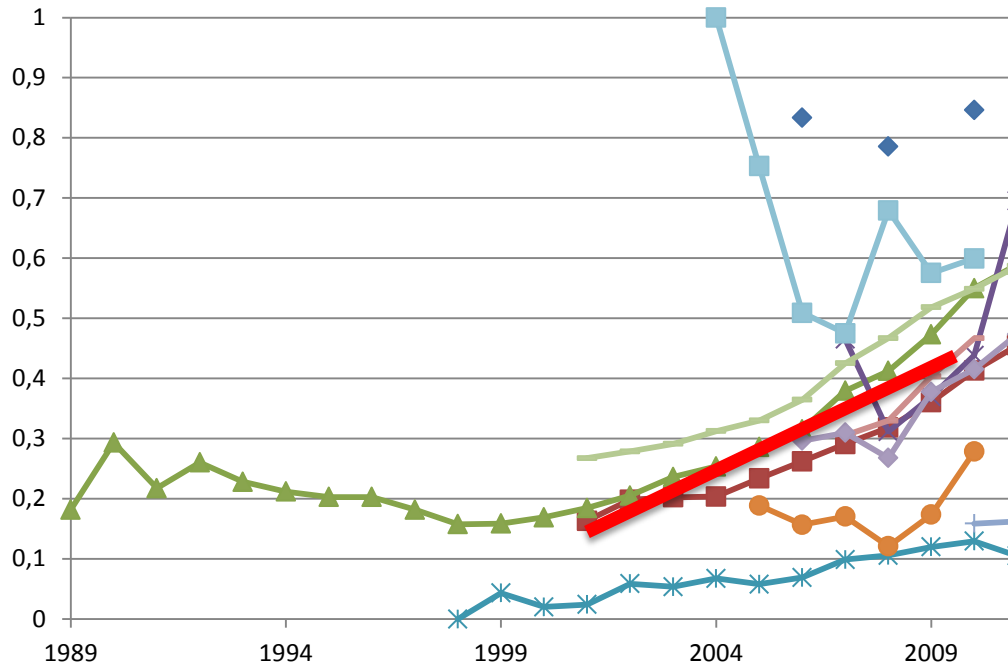
Medication-use metrics by data source





Monotherapy – diabetes

General upward trend in monotherapy



◆ AUSOM (SKorea*)

■ CCAIE (US#)

▲ CPRD (UK*)

✕ CUMC (US*)

* GE (US*)

● INPC (US*#)

+ JMDC (Japan#)

— MDCD (US#)

— MDCR (US#)

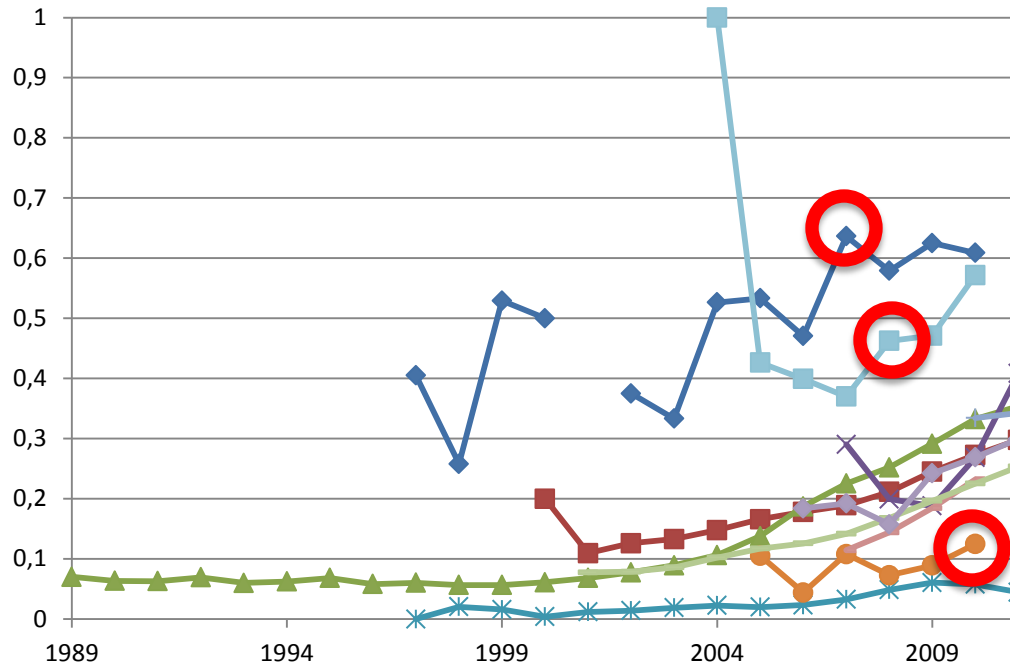
◇ OPTUM (US#)

■ STRIDE (US*)



Monotherapy – HTN

Academic
medical
centers
differ from
general
practices



◆ AUSOM (SKorea*)

■ CCAIE (US#)

▲ CPRD (UK*)

× CUMC (US*)

* GE (US*)

● INPC (US*#)

+ JMDC (Japan#)

— MDCD (US#)

— MDCR (US#)

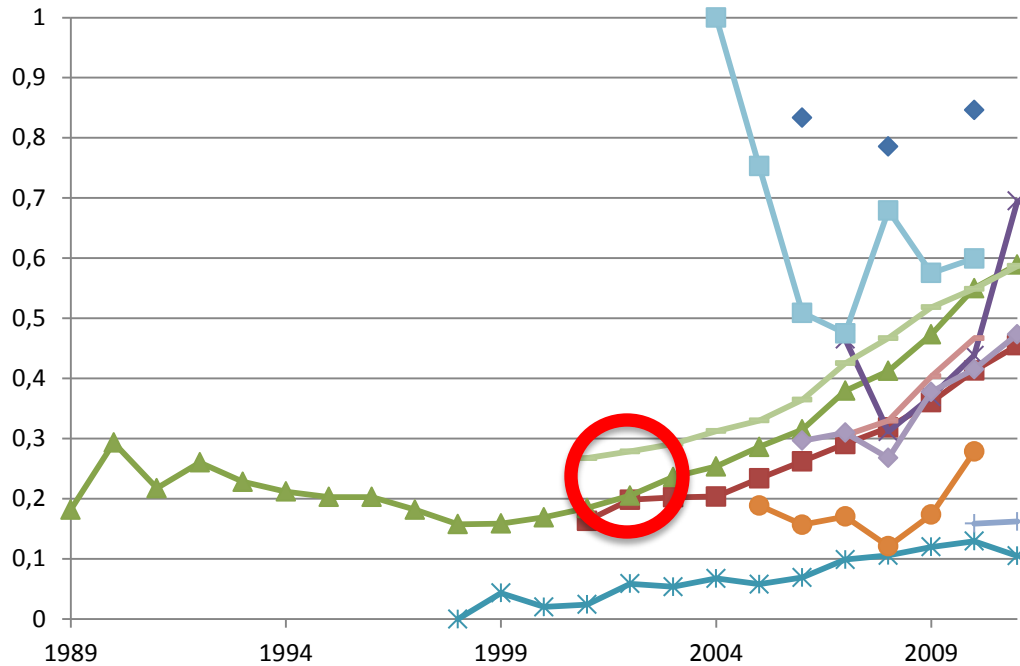
◇ OPTUM (US#)

■ STRIDE (US*)



Monotherapy – diabetes

General practices, whether EHR or claims, have similar profiles



◆ AUSOM (SKorea*)

■ CCAIE (US#)

▲ CPRD (UK*)

✕ CUMC (US*)

* GE (US*)

● INPC (US*#)

+ JMDC (Japan#)

— MDCD (US#)

— MDCR (US#)

◇ OPTUM (US#)

■ STRIDE (US*)



Privacy

- Patient privacy
 - Keep data within institutional firewall
 - De-identify the database removing identifiers and potentially shifting dates
 - US: Safe Harbor and Statistical Determination of Low Risk of Re-identification
- Business privacy
 - Public display of uncorrected error rates
 - Retained object
 - Public display of competitive strengths and weaknesses
 - Pool data



Conclusions: Treatment pathways

- General progress toward more consistent therapy over time and across locations
- Differ by country
- Differ by practice type
- Not differ so much by data type (claims, EHR)
- Differ by disease
 - Even before guidelines published
 - Disease differences and literature
- Huge proportion of unique pathways



Conclusions: Network research

- It is feasible to encode the world population in a single data model
 - Over 600,000,000 records by voluntary effort (682,000,000)
 - Generating evidence is feasible
 - Stakeholders willing to share results
 - Able to accommodate vast differences in privacy and research regulation
-



Join the journey

<http://ohdsi.org>