OHDSI / OMOP Introduction

for clinical investigators

OMOP Team, IQVIA
Jan 2021

Training series plan

+ Session 1 : Course Introduction
  - OMOP CDM and vocabulary overview, OMOP conversion, data quality, examples of previous research and use cases, introducing ATLAS and OHDSI tools

+ Session 2: OMOP CDM/Vocabulary Tutorial
  - Concept, Concept mapping, Hierarchy, Ancestors, and OMOP CDM

+ Session 3: Cohort and Cohort Characterization
  - Concept sets, cohort definition, and cohort characterization

+ Session 4: Treatment Pathways and Incident Rates
  - Treatment pathways, Incident rates, and Characterization using R

Table of contents

Ground Rules

+ This session will be recorded
+ Please make sure your microphones are muted
+ Type your questions in the chat or bring them to the Q& A session
+ Turn off your camera

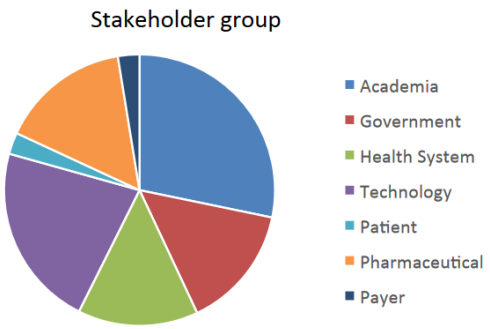
OHDSI overview

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

What OHDSI is:

- ✓ **Open Source**
- ✓ **Community**
- ✓ **Data**



Why Choose OHDSI/OMOP:

- ✓ **Fast, reliable** studies across a series of datasets and data types
- ✓ **Reduced cost of ownership** including understanding coding schemes, writing statistical programs across databases or developing software
- ✓ **Expanded data access** via the OHDSI network and remote multi-center database studies

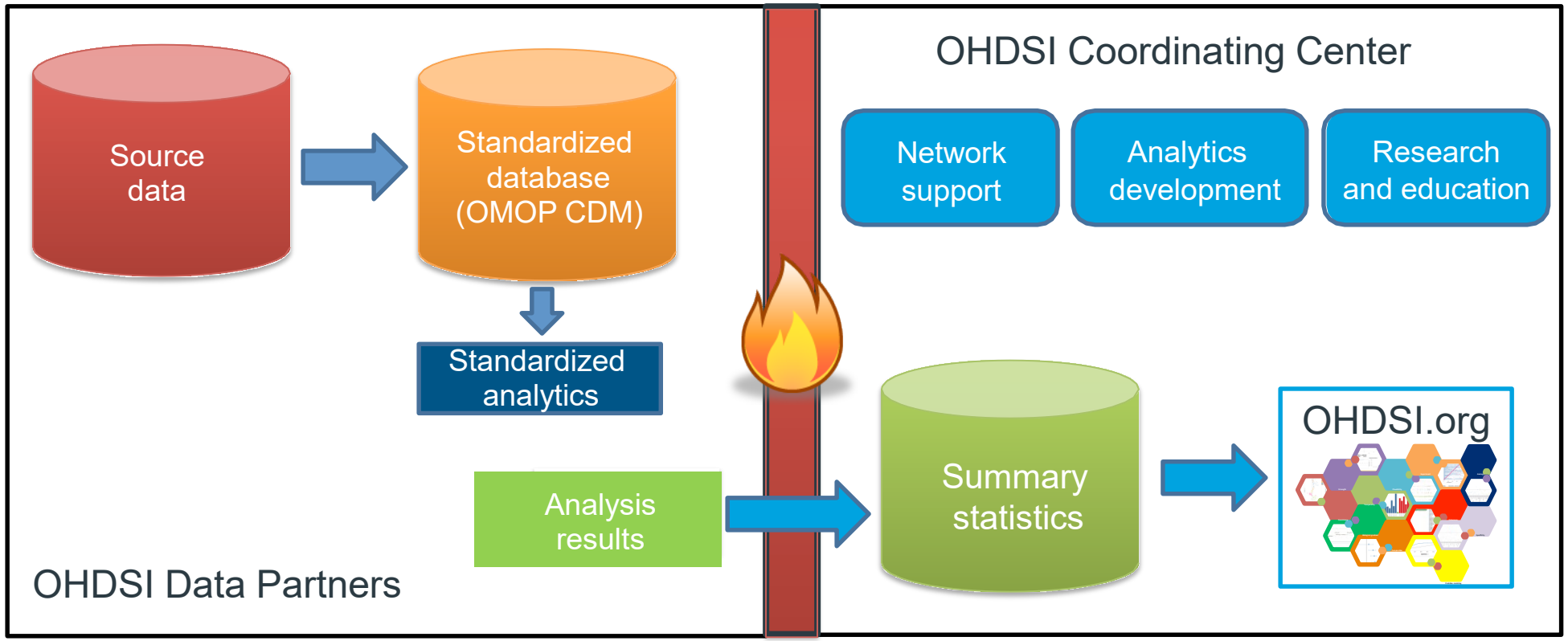
OHDSI Collaborators:

- 2,770 users
- 25 workgroups
- 18,700 posts on 3,250 topics

OHDSI Network:

- >150+ databases
- 21 countries
- 2.1B patient records, 369M ex-US

Keep data local and only share results



OMOP to OHDSI

OMOP Investigators

OHDSI

Columbia University

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 centre housed at Columbia University

Public, open

Not pharma funded

International

History of OMOP and OHDSI

2009

OMOP Experiment #1

- FDAAA calls for establishing Risk Identification and Analysis System for drug surveillance
- OMOP Experiment creates a framework for evaluating 14 methods of epidemiological designs
- 10 data sources, claims and EHRs, 200M+ lives

2011

OMOP Experiment #2/European OMOP

- Focused on a subset of data (4 claims, 1 EHR) and 7 methods
- Replicating experiment findings on European databases

2013

End of OMOP Experiment

- Last of the OMOP Meetings present findings of empirical experiments
- Artifacts include the OMOP CDM, vocabularies and validated methods for analyzing real world data

2014

Launch of OHDSI

- OHDSI's first face-to-face meeting at Columbia University

2015

First OHDSI Symposium/Network Study Published

- Community begins open source work under OHDSI brand
- First global network study characterizing treatment pathways

First Hackathon at Columbia University

2016

Formation of China Chapter

- To use data science and informatics methods to promote health and medical data research in China

Formation of Korea Chapter

2017

Formation of European Chapter

- Led by the coordinating center at the Erasmus University Medical Center in Rotterdam

FDA Adoption (FDA BEST Launch)

2018

EHDEN Initiation (Europe)

- Started under the Innovative Medicines Initiative (IMI) that will drive the adoption of the OMOP-CDM in Europe

FEEDER-NET Initiation (Korea)

First European Symposium

2019

EMA Adoption

Formation of Australia, Japan and Singapore Chapters

2020

NMPA Adoption

Global Acceptance

- OHDSI grows to >152 databases, 18 countries, 2.1B patient records, 369M ex-US
- Regional chapters in US, Europe, China, South Korea + Asia-Pacific, Latin America
- Offering regional symposia

OMOP and OHDSI - recap

OMOP

Consists of

- **OMOP Common Data Model** (CDM)
- **Standardized vocabularies**
- **Standardized analytics** (computationally efficient and reusable analytics)

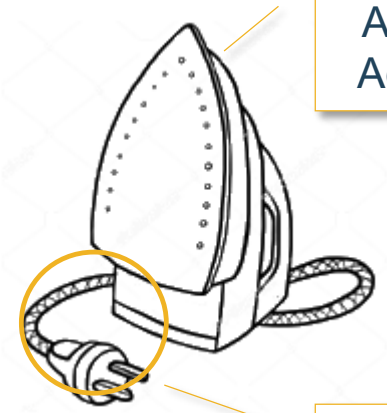
OHDSI

- **OHDSI** is the organization that **owns OMOP**
  - **Open science** community for all levels of stakeholder
  - Generates evidence to promote better health decisions and patient care

Why OHDSI?

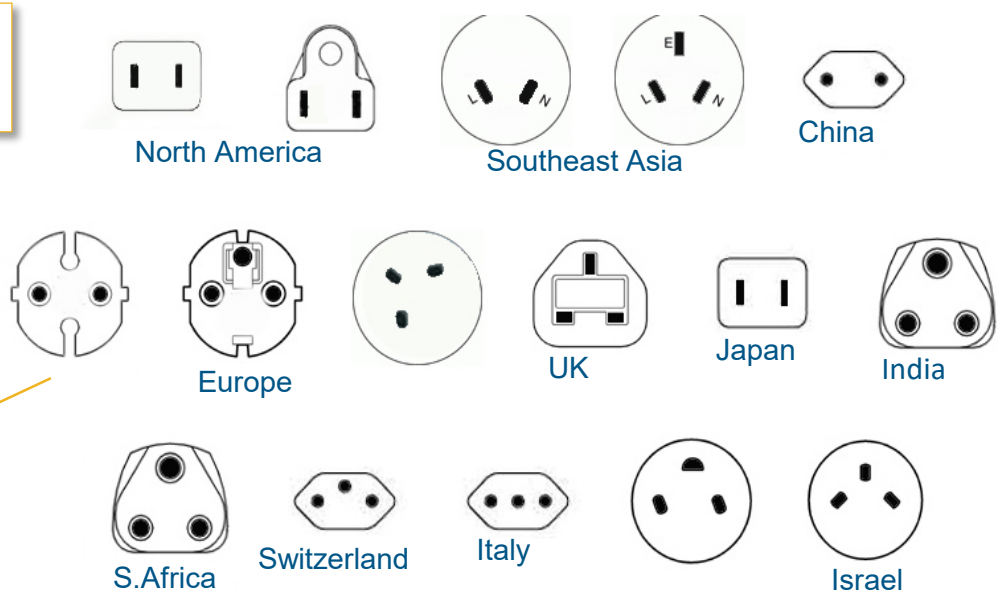
Current Approach: “One Study – One Script”

"What's the adherence to my drug in the data assets I own?"

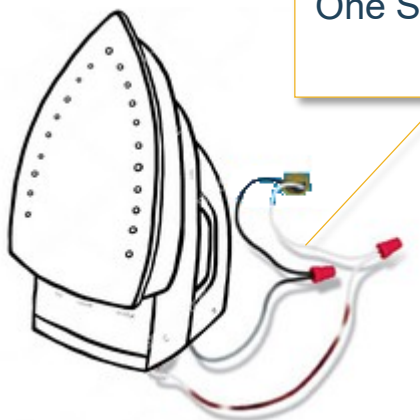


Analytical method: Adherence to Drug

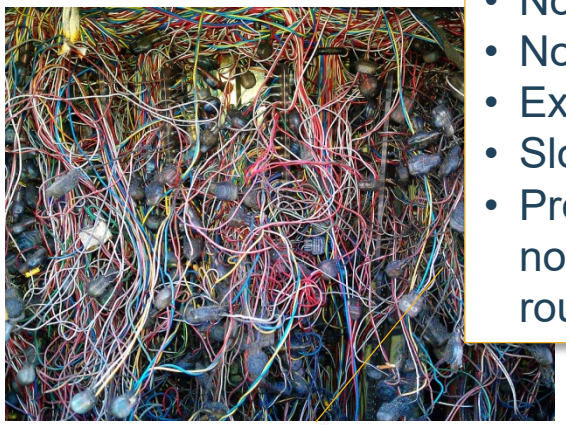
Application to data



Current solution:

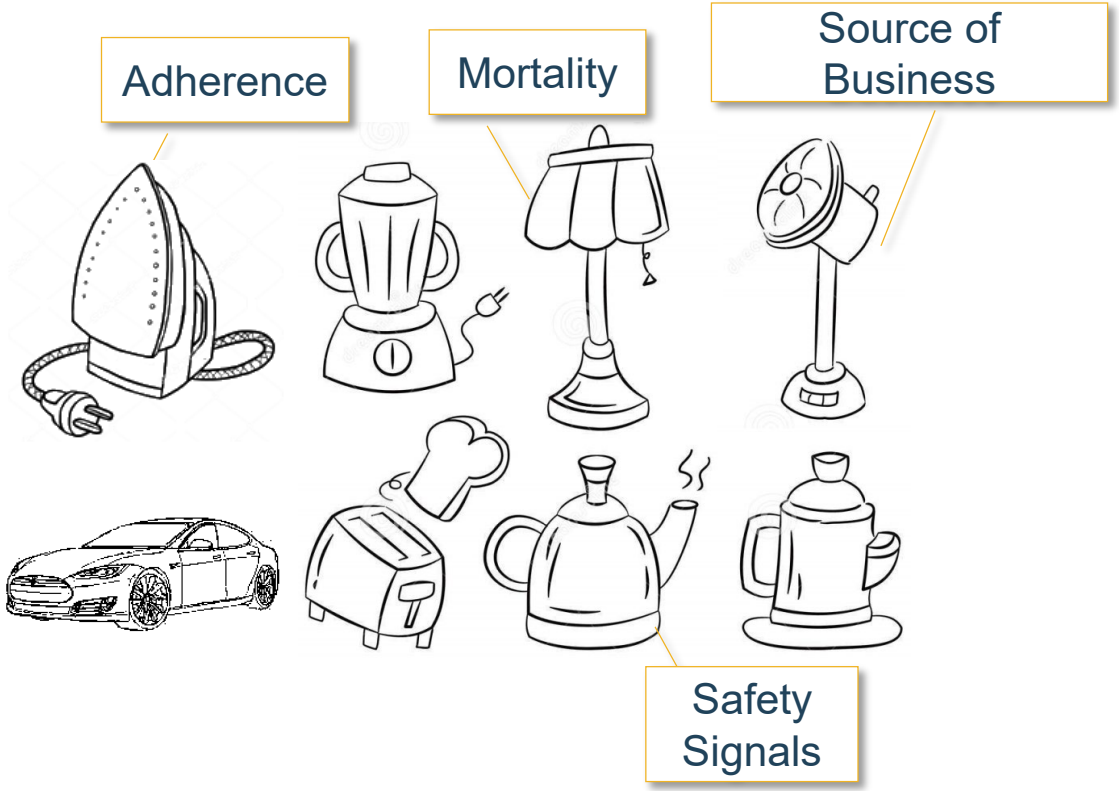


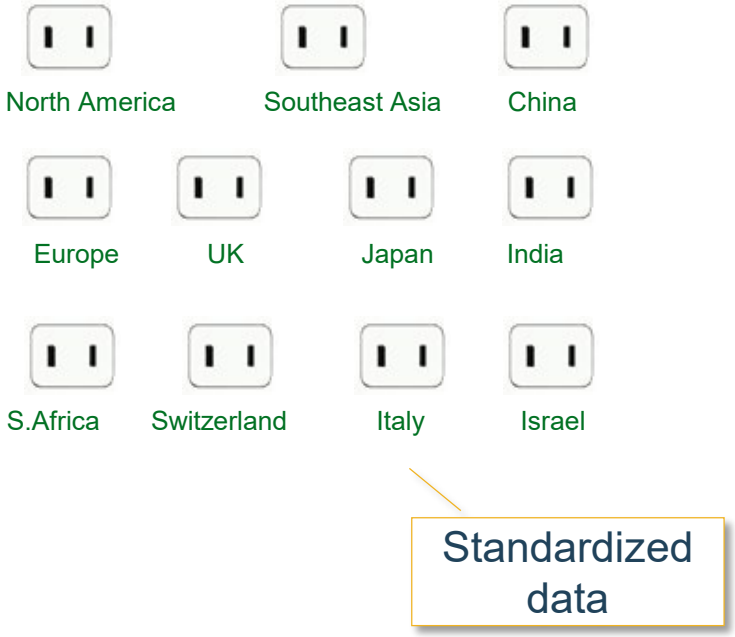
One SAS or R script for each study



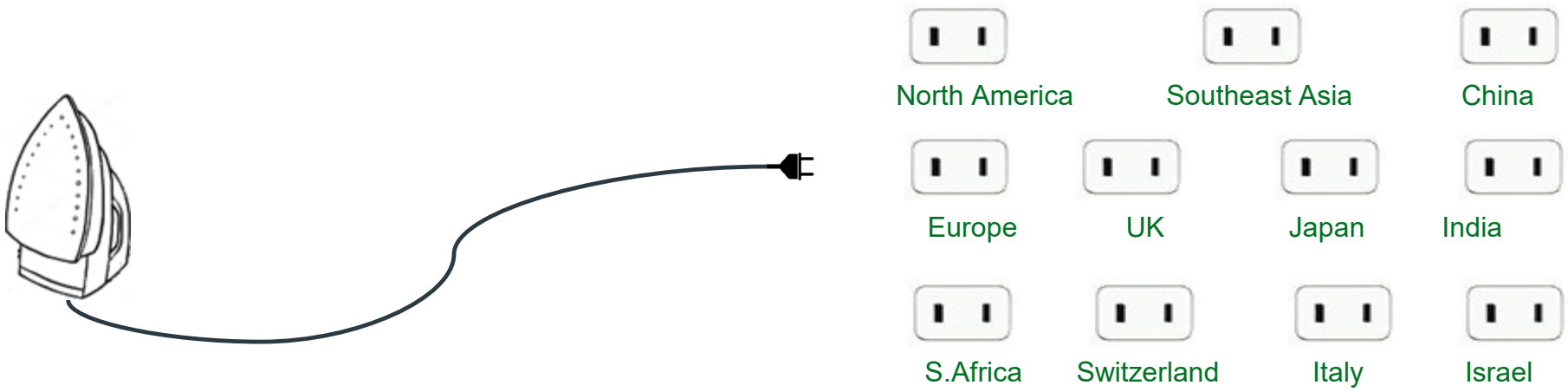
- Not scalable
- Not transparent
- Expensive
- Slow
- Prohibitive to non-expert routine use

Solution: Data Standardization Enables Systematic Research



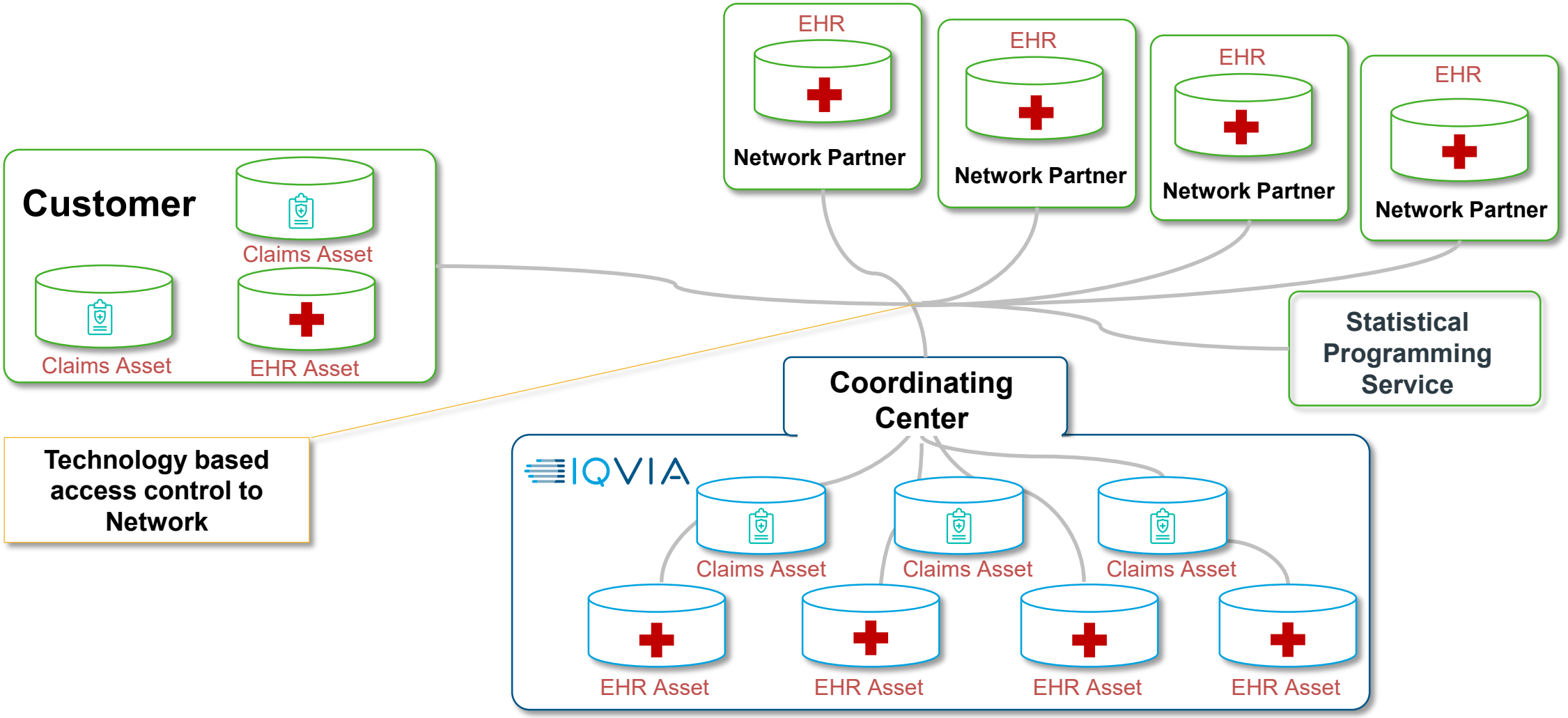


Analytics can be remote



Analytics can be behind firewall

IQVIA Research Network - Structure and participants



Benefits of using OMOP

Standardized data model
- OMOP CDM v5.3

Standard coding system
- SNOMED, RxNorm, RxNorm extension, LOINC

Systematic data quality
- Achilles
- Data Quality Dashboard

Standardized tools / methodology
- Atlas
- Validated methodology

Productized Analytics
- Simple to complex

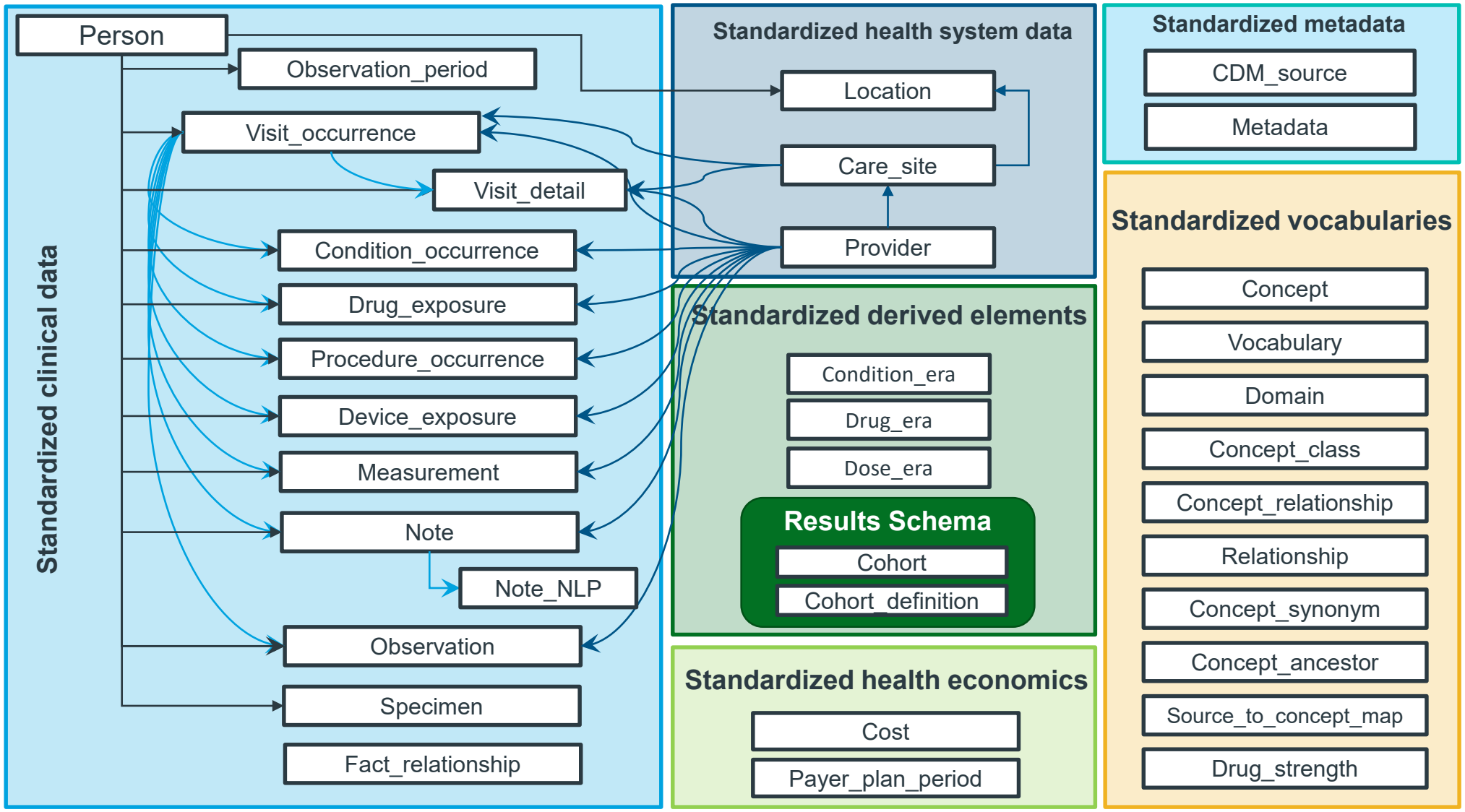
OHDSI/OMOP

Faster and more reliable studies across a series of datasets and data types

Reduced cost of ownership including understanding coding schemes, writing statistical programs across databases or developing software

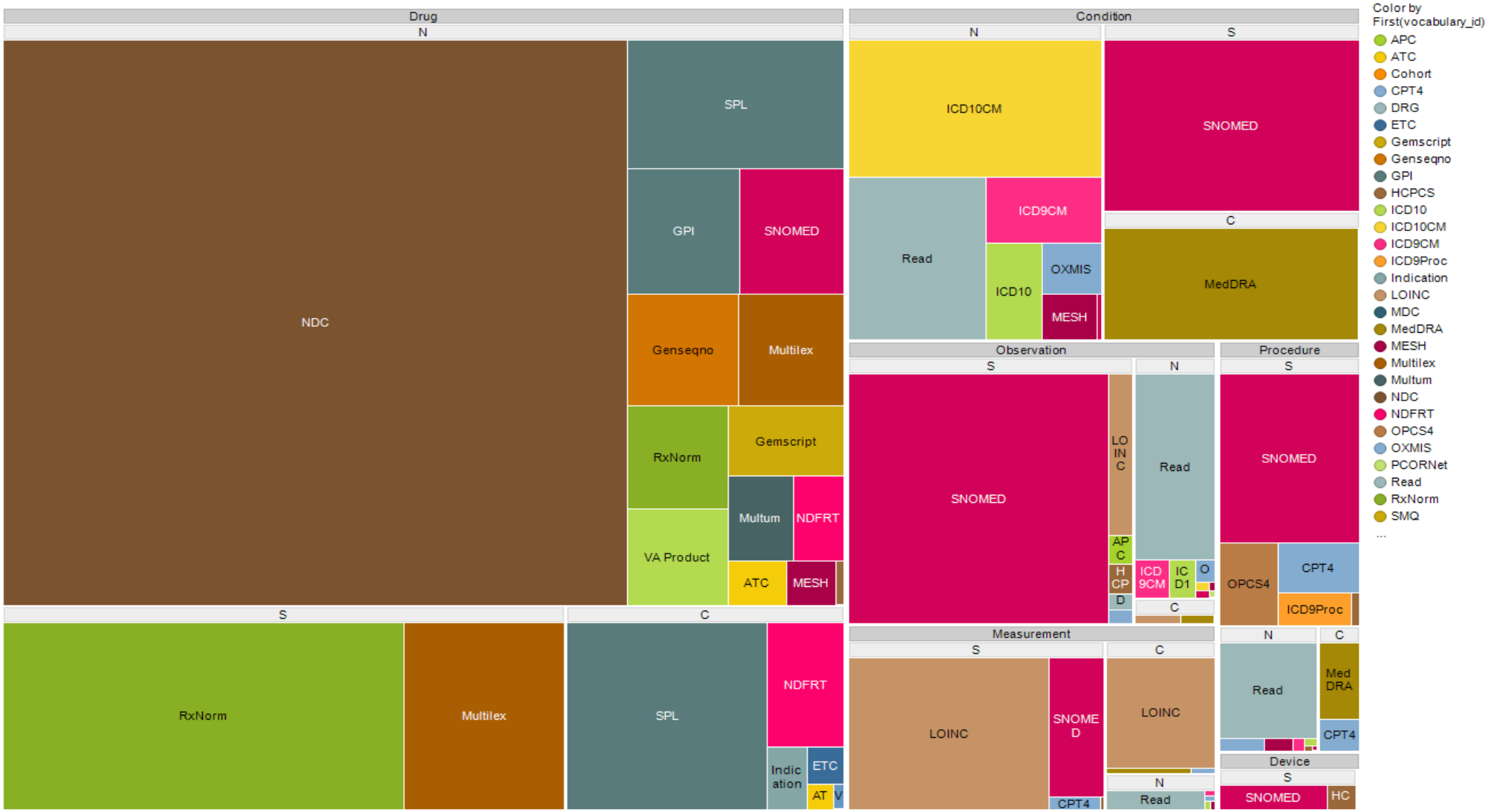
Expanded data access via the OHDSI network and remote multi-center database studies

OMOP Common Data Model v5.3.1



Translation to vocabularies

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



Benefits of using OMOP are far more than one-script fits all

CDM benefits

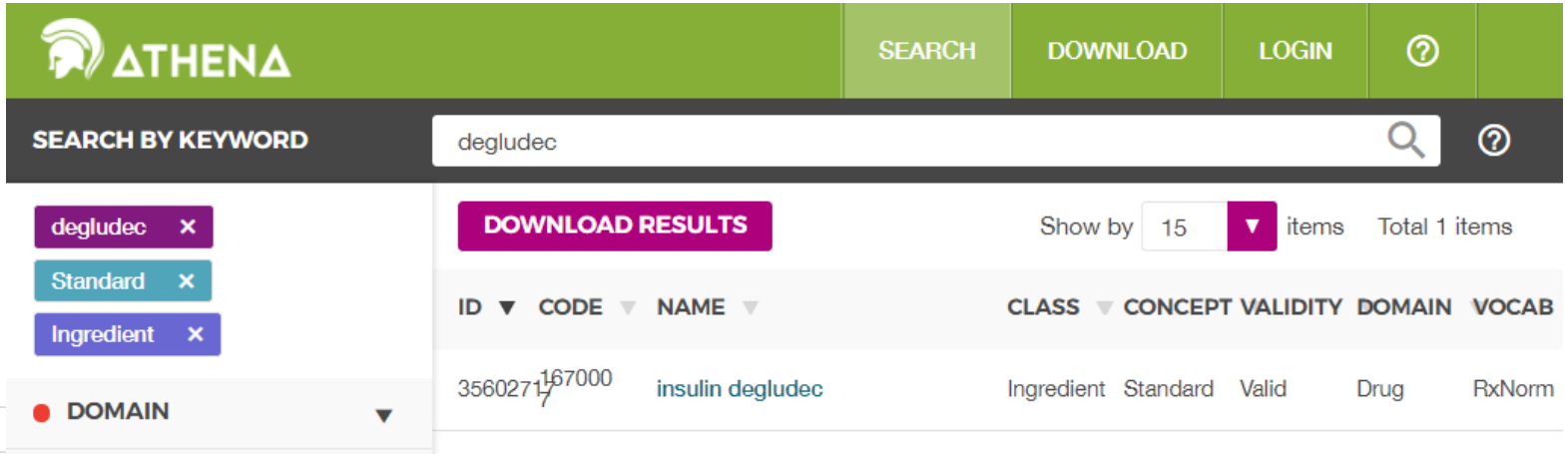
- One script fits all
- No switching between dialects
- Modular table structure and consistent field names for easy querying
- Hierarchical standard vocabularies

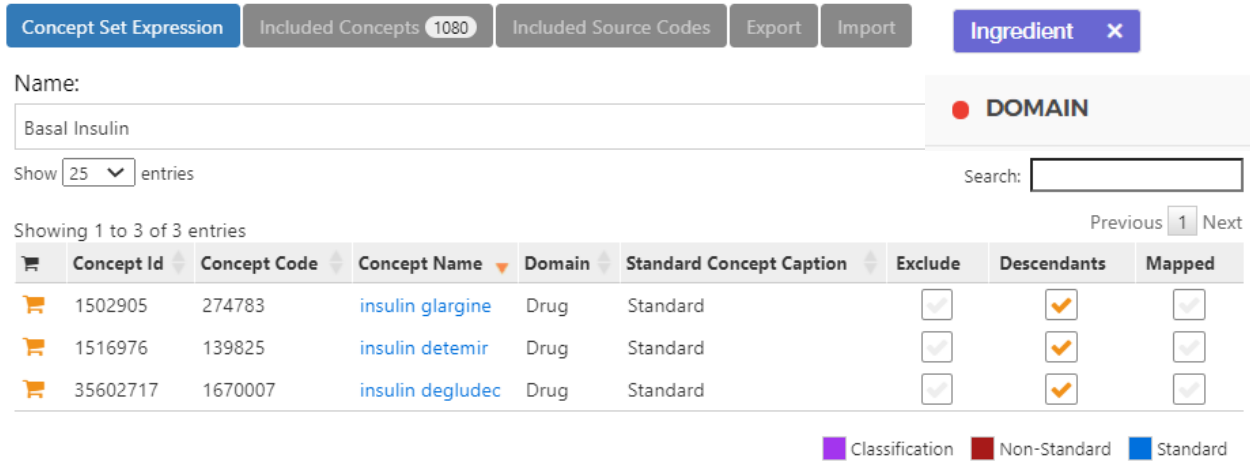
Standardized tools

- Community phenotype definitions
- Comprehensive ecosystem of tools
- High parameterization gives flexibility
- No need to re-code complex analytics

Standard vocabularies have been chosen for efficiency

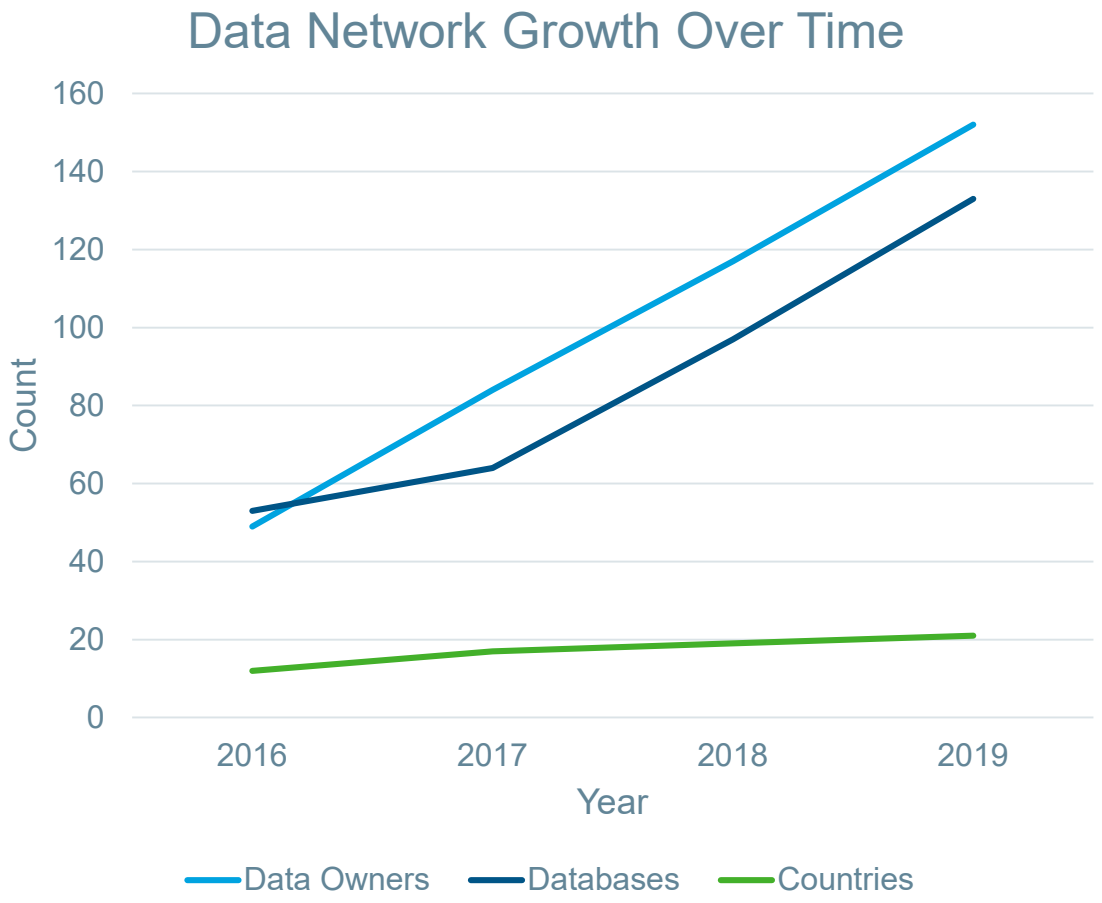
- Hierarchical vocabularies mean one parent concept can capture hundreds of codes
- This top down approach is the most efficient way of building concept sets
- Concept sets can still be specified bottom-up using individual source codes

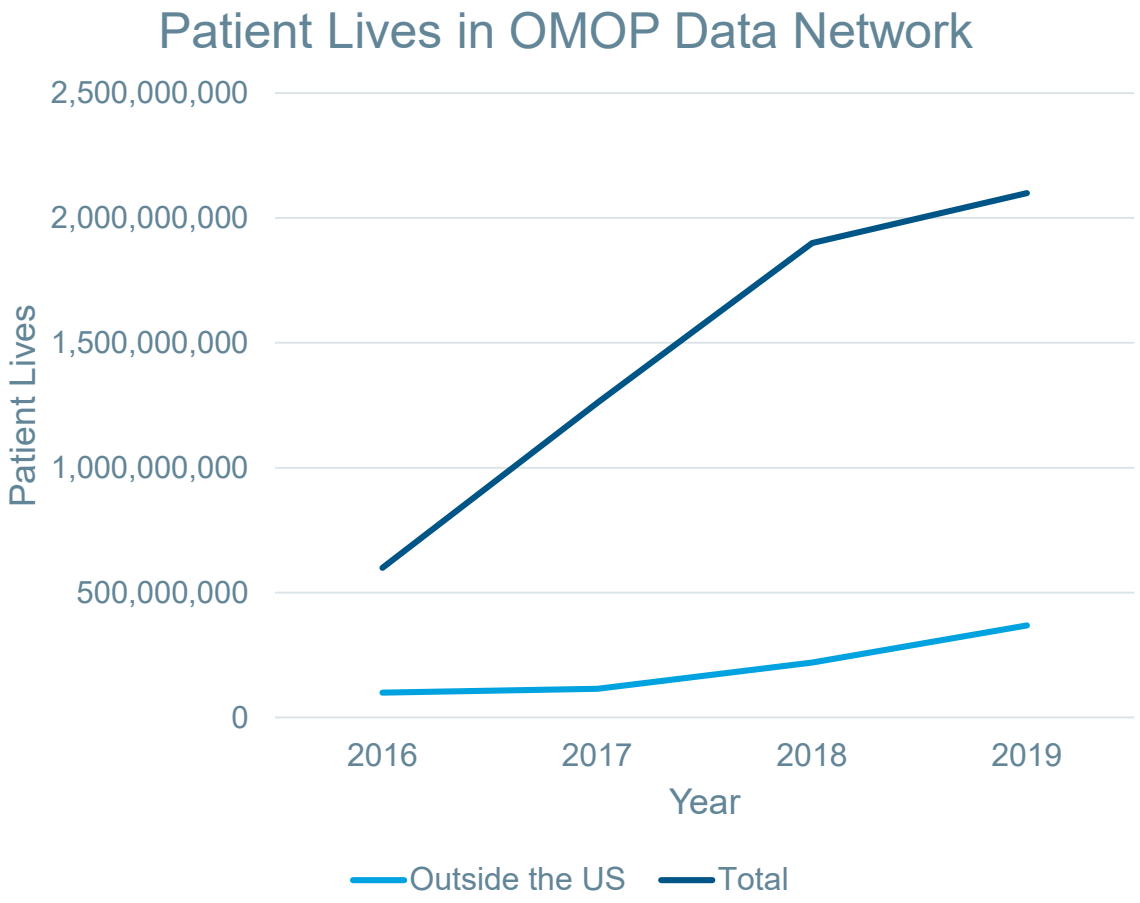




OHDSI adoptions

OHDSI Community Adoption Over the Years





Total lives is not unique due to issues with deduplication of US data

NIH Adopts OMOP CDM for National COVID-19 Surveillance

NIH National Center for Advancing Translational Sciences

NATIONAL CENTER FOR DATA TO HEALTH

National COVID Cohort Collaborative

Overview:
Consortia of distributed clinical data networks (PCORnet, OHDSI, ACT/i2b2, TriNetX)

Goal:
Improve the efficiency and accessibility of analyses with COVID-19 clinical data, expand ability to analyze and understand COVID, and demonstrate a novel approach for collaborative pandemic data sharing

Program Workstreams

Data Partnership and Governance

Develop partnerships with organizations and their IRBs (single IRB review offered at Johns Hopkins University) and execute a common data use agreement (DUA) for contributing to and accessing the COVID-19 dataset. Establish a Data Access Committee for reviewing access requests.

Data Ingestion and Harmonization

Ingest limited data sets that are available in their native data formats, such as PCORnet, ACT, and OMOP. Harmonize the data sets into a common data model (CDM) based on the OMOP v5.3.1 standard.

Phenotype and Data Acquisition

Establish a common COVID-19 phenotype that will define the data pull for the limited data set. Create a "white glove" service to obtain data from each site by building easily adaptable scripts for each clinical data model. Ingest data into a secure location, per approved institutional agreements.

Collaborative Analytics

Work collaboratively to generate insights related to COVID-19 from the harmonized limited data set. Experts in artificial intelligence (AI), machine learning (ML), and other technologies will assist in reviewing and iterating on portal architecture to ensure fit-for-purpose implementation.

EMA Guide on Methodological Standards in Pharmacoepidemiology Rev.8

- **Section 4.6 – Research Networks for multi-database studies**
  - Use of a common data model (CDM) implies that local formats are translated into a predefined, common data structure, which allows launching a similar data extraction and analysis script across several databases.
  - The main advantage of a general CDM is that it can be used for virtually any study involving that database.

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

European Network of Centres for Pharmacoepidemiology and Pharmacovigilance

EMA/95098/2010 Rev.8

The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology
(Revision 8)

KEYWORDS	*methodological standards, pharmacoepidemiology, pharmacovigilance, ENCePP, research, guidance*

**From The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP)*

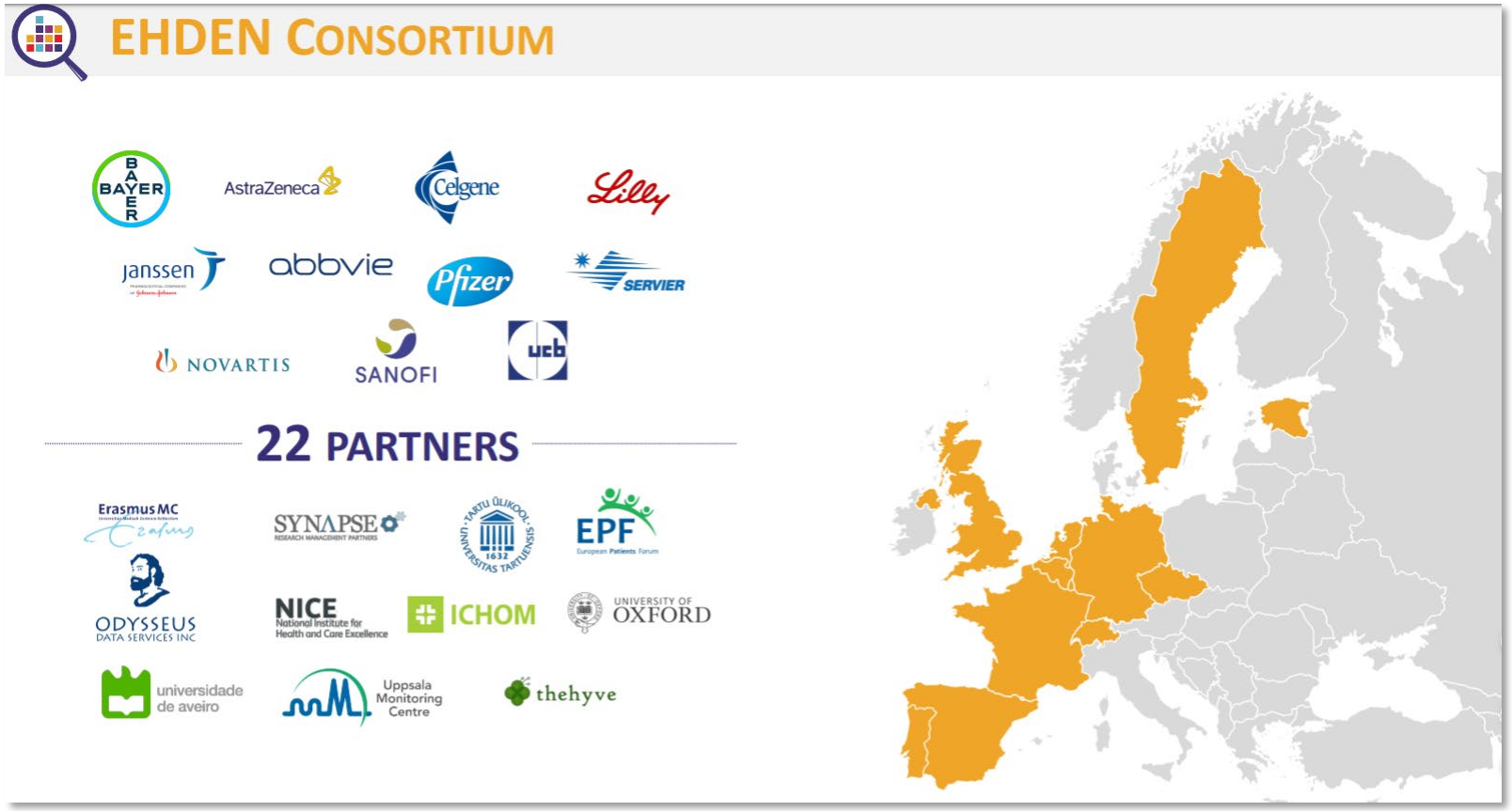
EHDEN

Vision

The European Health Data & Evidence Network (EHDEN) aspires to be the trusted observational research ecosystem to enable better health decisions, outcomes and care

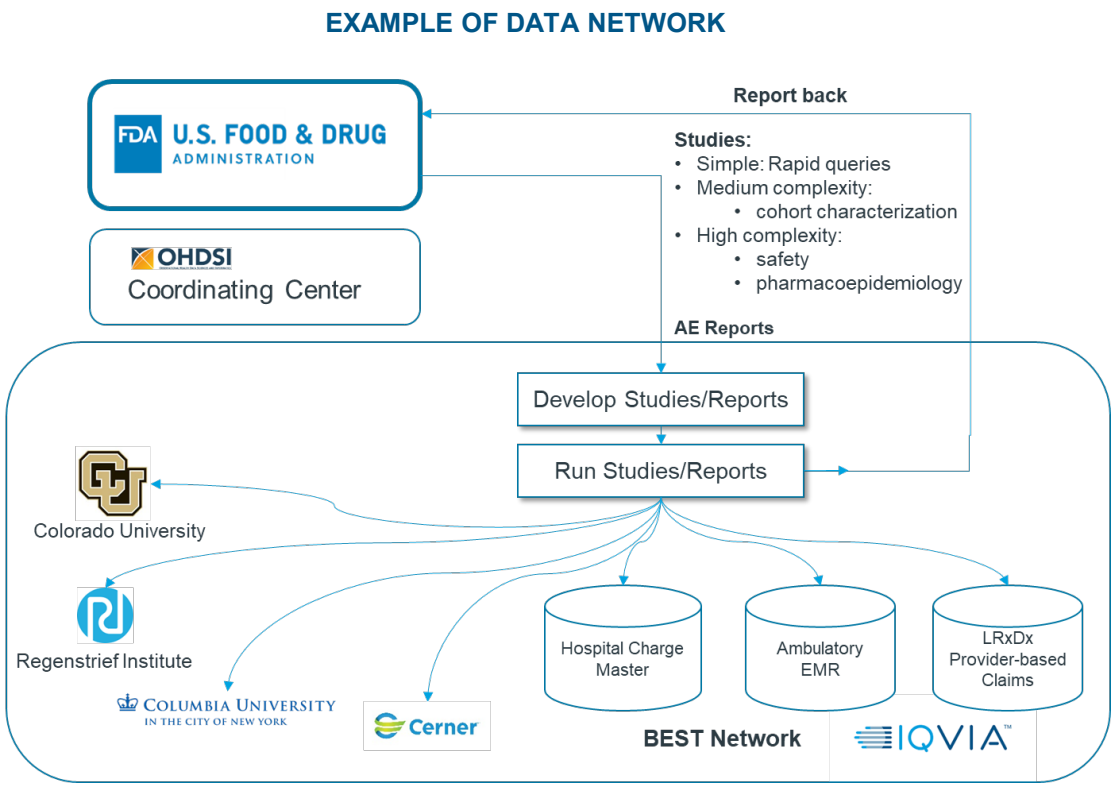
Mission

Our mission is to provide a new paradigm for the discovery and analysis of health data in Europe, by building a large-scale, federated network of data sources standardized to a common data model



**EHDEN = European Health Data & Evidence Network*

FDA BEST – Overview



Network Overview

- Started in September 2017
- Today's largest distributed network of clinical data
- Collaborative research model, guided by efforts across the OHDSI community and US FDA
- Iterative sponsored studies facilitated by IQVIA and the global network of data partners

Benefits to Participating Sites

- Access to large, diverse patient populations
- Maintain direct control of your site's clinical data, share only aggregate data
- Access to IQVIA data enrichment programs to enhance site data (e.g. NLP tools, linkage services)
- Ability for researchers to externally validate single-center findings

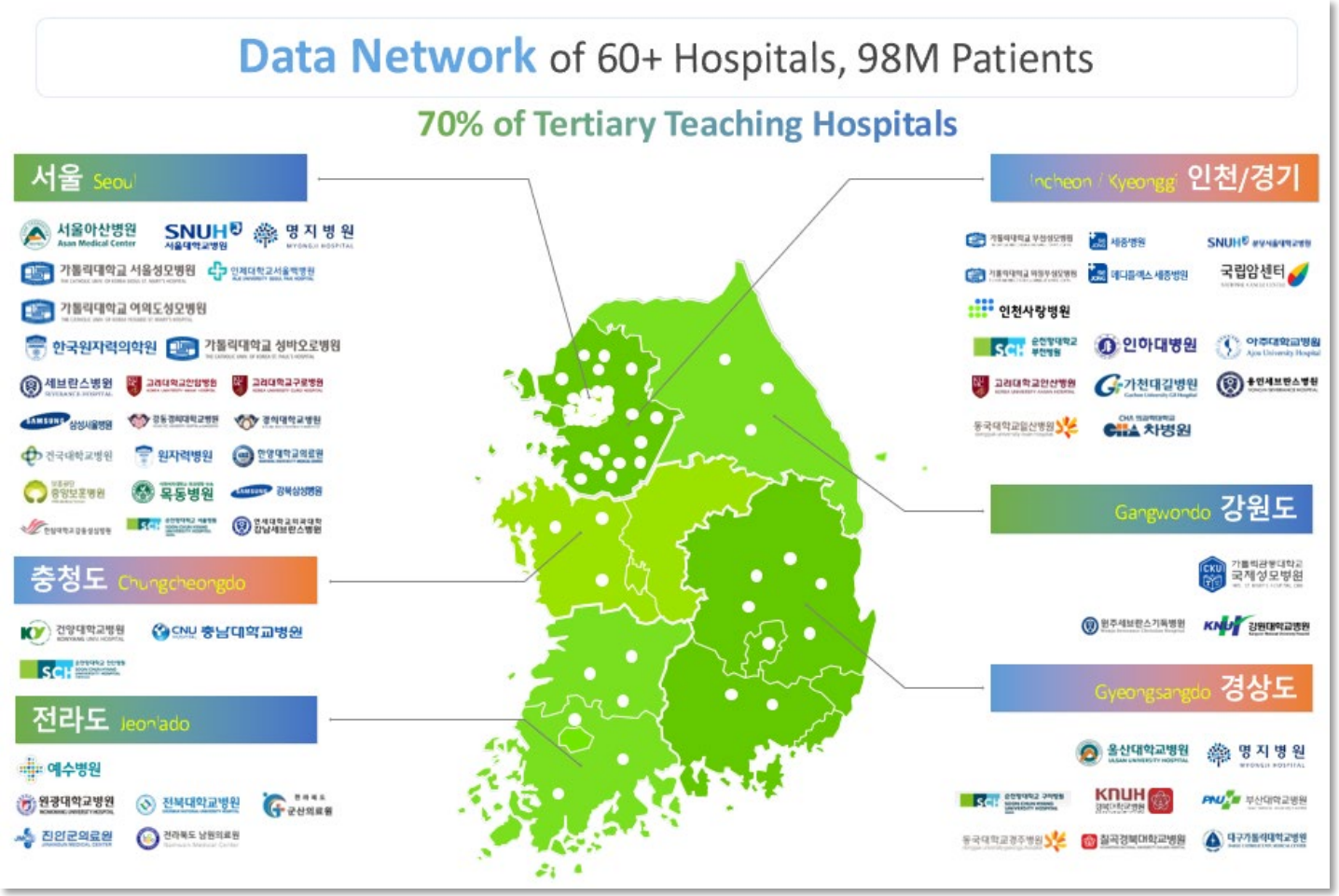
Korean Government Initiatives

Grants from Ministry of Industry 2018:

- Vocabulary and deidentification: 4 year
- Sophistication of FEEDERNET and incorporating more hospitals (+21 hospitals), 4 year
- 3 vertical services on FEEDERNET for companies, 3 years for each project
- 2 other vertical services on FEEDERNET for companies, 3 years for each project

Grants from Ministry of Health 2018:

- 12 projects for various clinical research using CDM, 3 year for each project
- 10 projects on security and vocabulary on CDM, 3 years for each project



China's First Two Guides on RWE & RWD – Released in 2020

- 1st guide was released in Jan 2020, introducing the definition, data source requirement, design, and evaluation of using RWE for drug effectiveness study and safety monitoring.

- 2nd guide was released in Aug 2020, focusing on the details and importance of the source, safety, curation, quality assurance and maintenance of RWD, so that reliable RWE could be produced – see graph on the right

国家药品监督管理局药品审评中心
CENTER FOR DRUG EVALUATION, NMPA
NMPA CDE

此页面上的内容需要较新版本的 Adobe Flash Player。

Get ADOBE® FLASH® PLAYER

当前位置：新闻中心>>工作动态>>通知公告>>新闻正文

关于公开征求《用于产生真实世界证据的真实世界数据指导原则（征求意见稿）》意见的通知

发布日期：20200803

为进一步指导和规范申办者利用真实世界数据生成真实世界证据支持药物研发，我中心组织起草了《用于产生真实世界证据的真实世界数据指导原则（征求意见稿）》，现在中心网站予以公示，以广泛听取各界意见和建议，欢迎各界提出宝贵意见和建议，并请及时反馈给我们。

征求意见时限为自发布之日起2个月。

您的反馈意见请发到以下联系人的邮箱：

联系人：高丽丽、赵骏

联系方式：gaoll@cde.org.cn，zhaojun@cde.org.cn

感谢您的参与和大力支持。

国家药品监督管理局药品审评中心

2020年8月3日

附件 1：	《用于产生真实世界证据的真实世界数据指导原则（征求意见稿）》.docx
附件 2：	《用于产生真实世界证据的真实世界数据指导原则（征求意见稿）》起草说明.doc

**From Center for Drug Evaluation (CDE), National Medical Products Administration (NMPA)*

CDM & OHDSI Citations in the 2nd Guide

CDM Introduction in Guide:

- Under multidisciplinary collaboration, CDM was created with standardized structure, format and vocabulary, to achieve multi-center data integration and collaboration.

References in Guide:

- EMA. A Common Data Model for Europe – Why? Which? How? https://www.ema.europa.eu/en/events/common-data-model-europe-why-which-how
- OHDSI – Observational Health Data Sciences and Informatics, https://www.ohdsi.org

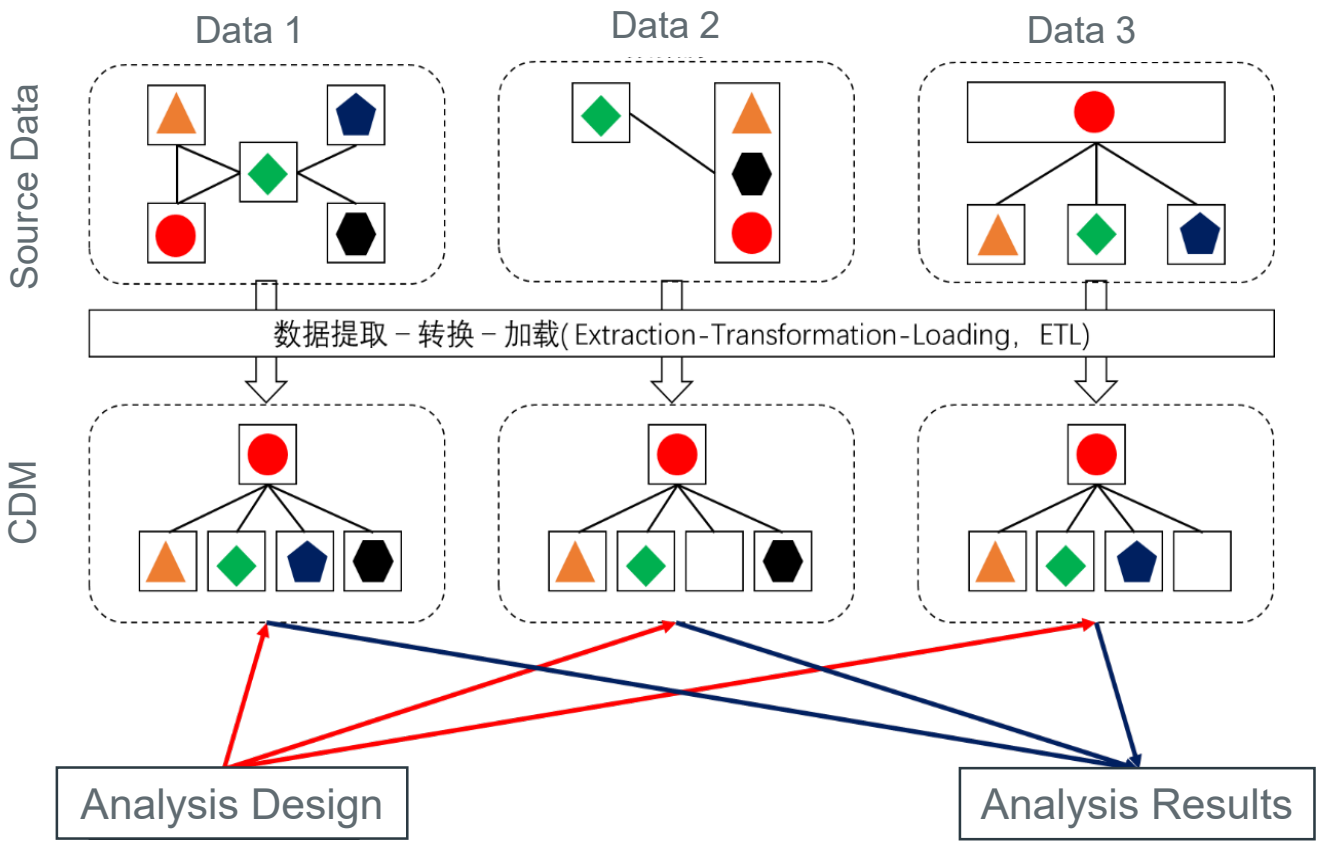
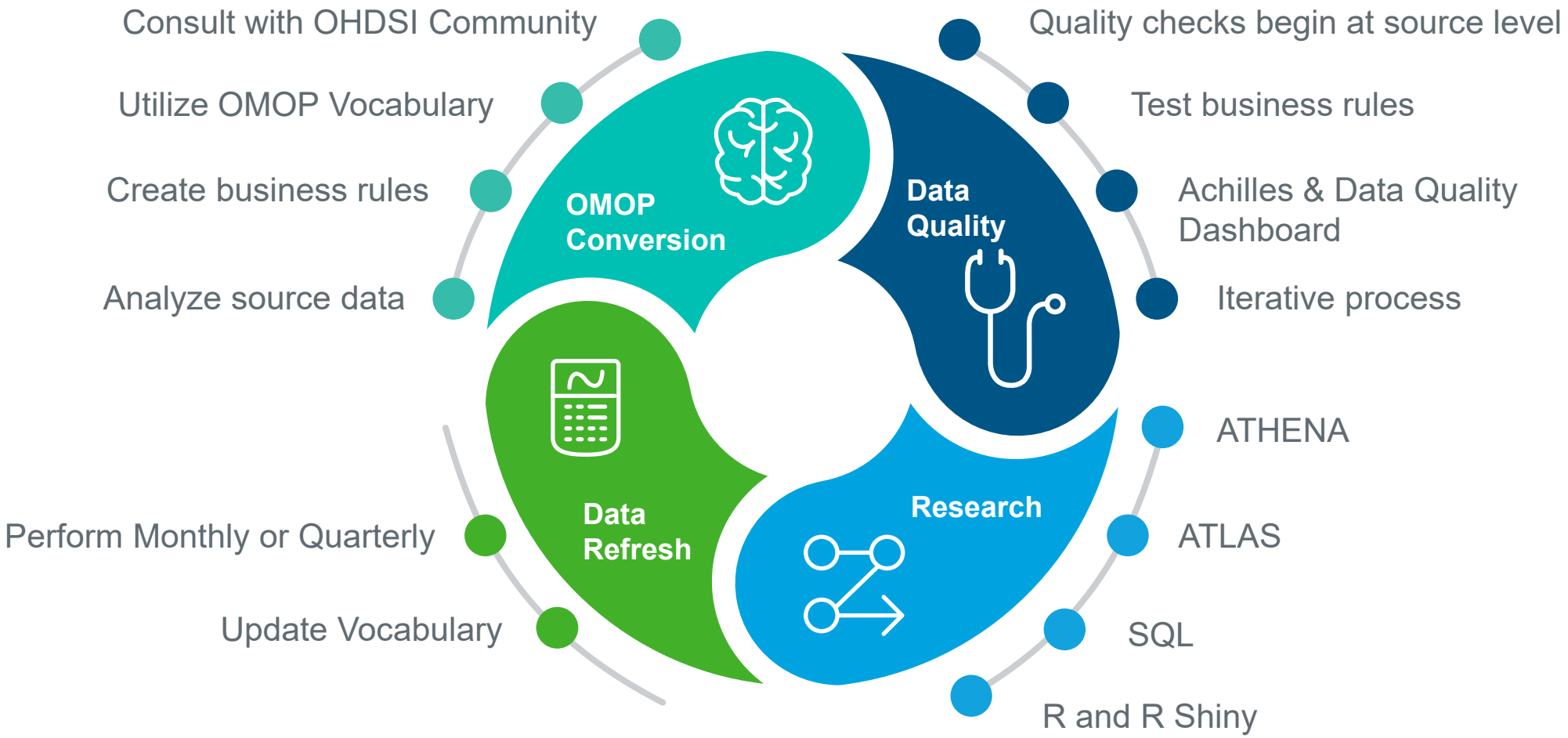


Fig 2 in Guide – Diagram on Converting Source Data to CDM

**Section 4 – Real World Data Curation*

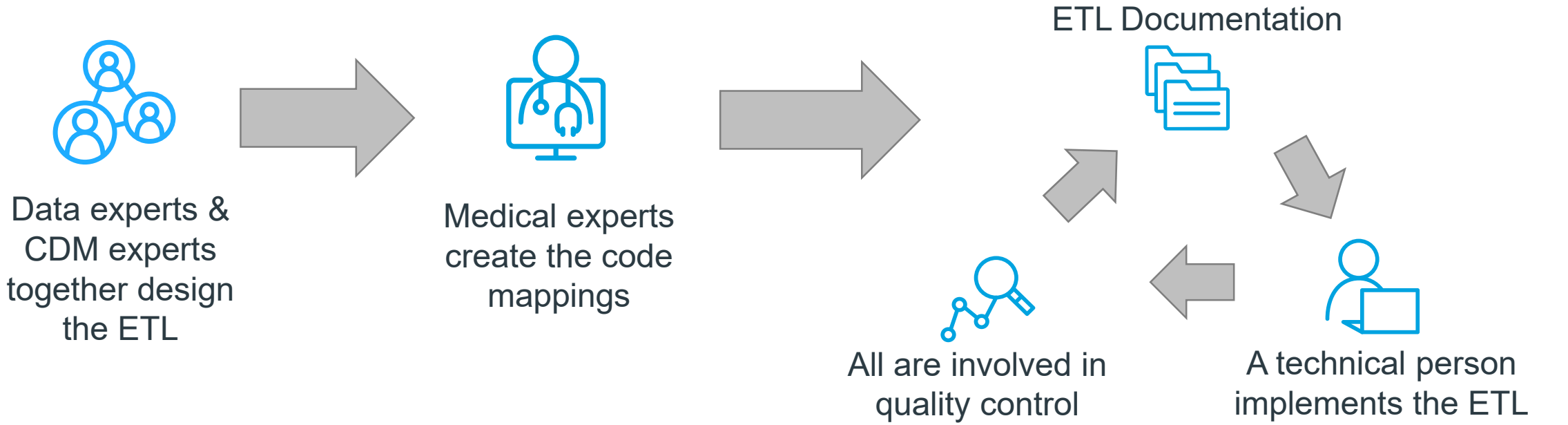
How to get started?




Q&A

OMOP Conversion

OMOP conversion process flow



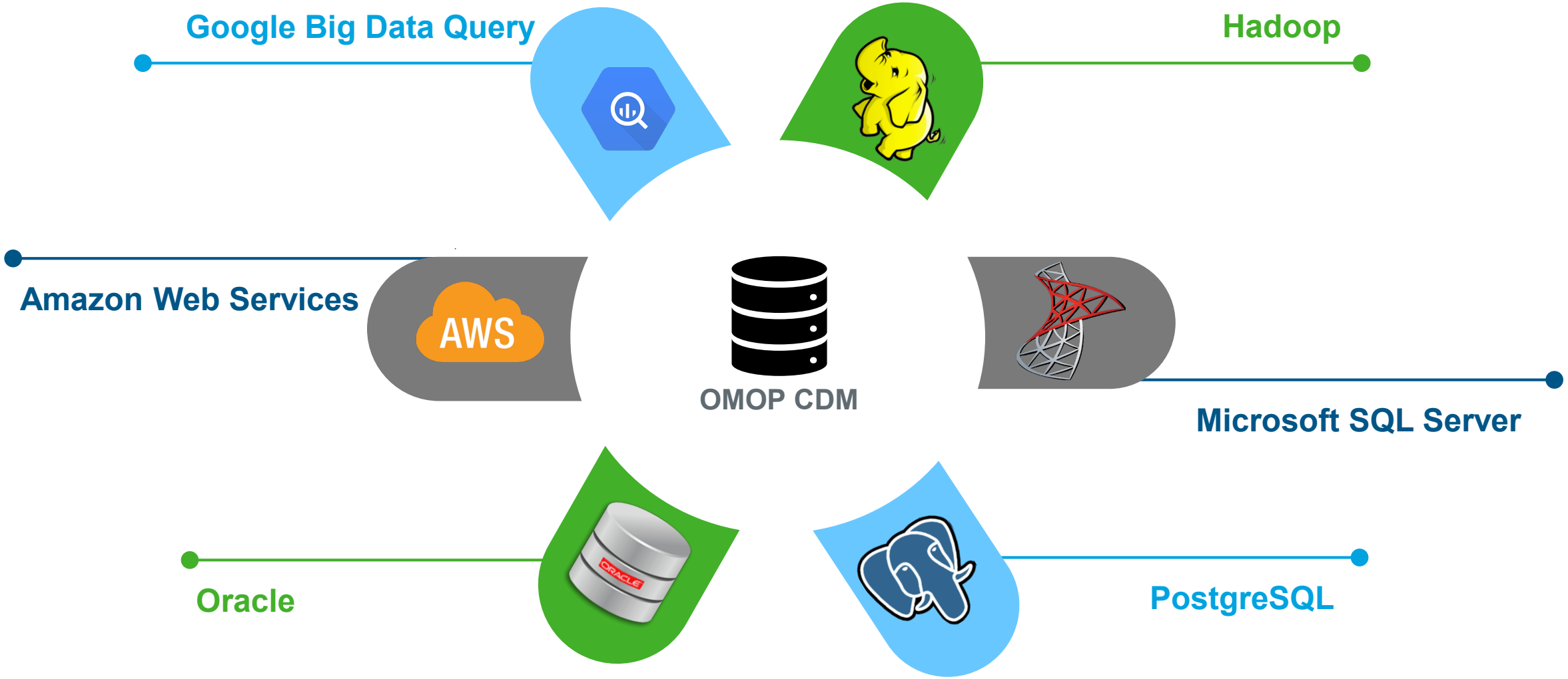
	Analysis			Quality Control			Development	
Tools	White Rabbit	Rabbit In a Hat	Usagi	Internal Quality Checks	Achilles	Data Quality Dashboard	Jenkins	Code Repository

OMOP CDM Version 5.3.1 Minimal Viable Product (MVP)

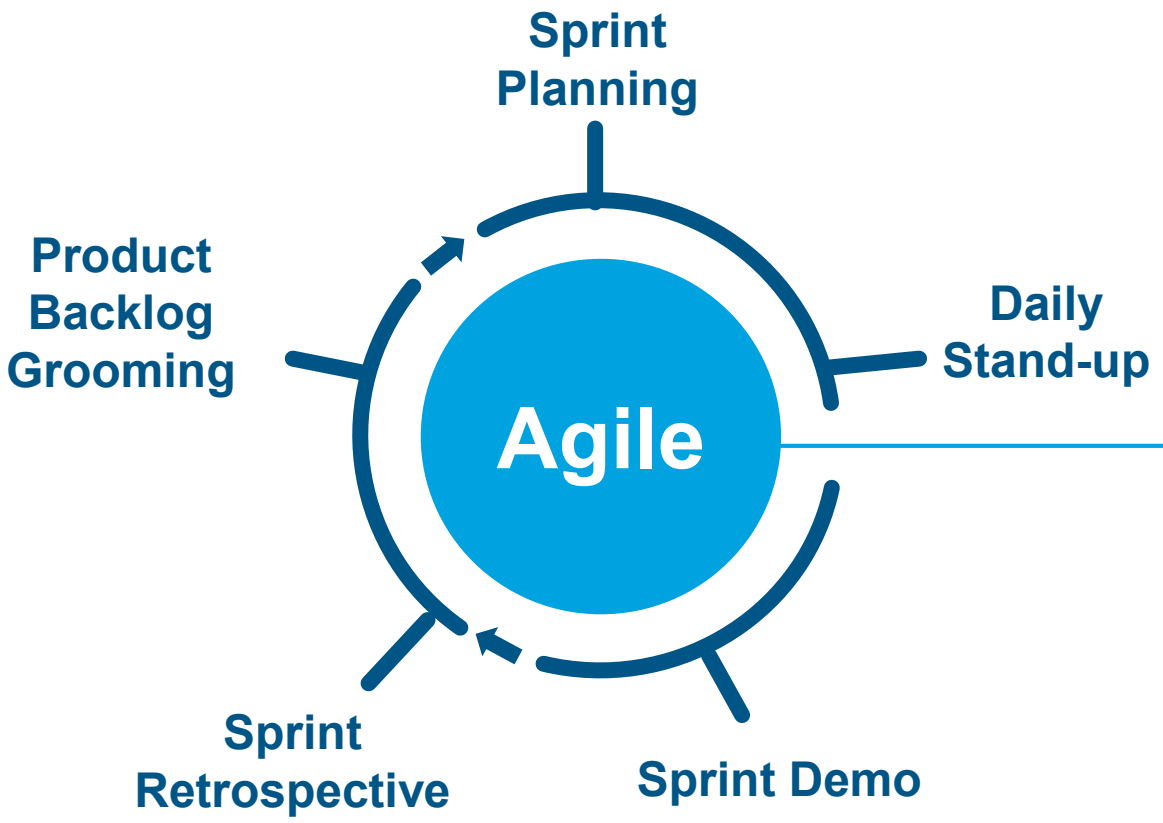
Health System Tables	Clinical Data Tables	Derived Tables (Logic Provided)	Health Economic Tables
• **Location** • **Care_Site** • **Provider** • **Person** • **Death**	• **Visit_Occurrence** • **Condition_Occurrence** • **Drug_Exposure** • **Procedure_Occurrence** • **Measurement** • **Observation** • **Observation_Period** • Specimen • Device_Exposure • Fact_Relationship • Visit_Detail • Note • Note_NLP	• Drug_Era • Dose_Era • Condition_Era	• Payer_Plan_Period • Cost

*12 of 23 tables can get you a functional CDM (Derived tables just need a script to run). Adjustments can be made to accommodate specific use cases

Technology Independent

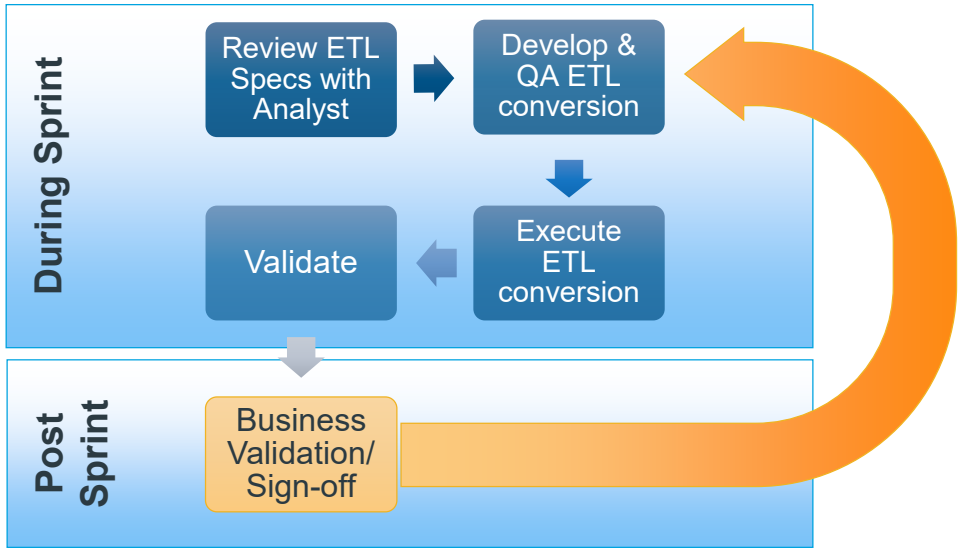


OMOP Agile conversion methodology



What is Agile?

- Project management & software development
- 2 week sprints
- Promotes continuous adaptation



Conversion timeline in sprint

Sprint 0
- Project kick-off
- Prep source data
- Gather documentation
- Analyze source data

Sprint 1
- Create ETL spec for dimension tables
- Identify custom mappings

Sprint 2
- Load dimension tables
- Create ETL spec for fact tables
- Load custom mappings

Sprint 3
- QA dimension tables
- Load fact tables

Sprint 4
- Load fact tables
- QA fact tables

Sprint 5
- QA fact tables
- Load derived tables

Sprint 6
- Obtain sign-off
- Deployment
- Delivery

Sprint 7
- Business Validation

Sprints

Source data profiling

- Used to analyze the structure and content of source data
- Assists with data types, values, frequency, anomalies
- Creates scan report of tables, columns, files
- Starts/continues investigation of source data with data owner
- Used in preparation for creating ETL specification

	A	B	C	D	E	F	G
1	Table	Field	Type	Max length	N rows	N rows checked	Fraction empty
2	beneficiary_summary	desynpuf_id	character varying	16	1031348	100000	0
3	beneficiary_summary	bene_birth_dt	date	10	1031348	100000	0
4	beneficiary_summary	bene_death_dt	date	10	1031348	100000	0.98493
5	beneficiary_summary	bene_sex_ident_cd	character varying	1	1031348	100000	0
6	beneficiary_summary	bene_race_cd	character varying	1	1031348	100000	0
7	beneficiary_summary	bene_esrd_ind	character varying	1	1031348	100000	0
8	beneficiary_summary	sp_state_code	character varying	2	1031348	100000	0
9	beneficiary_summary	bene_county_cd	character varying	3	1031348	100000	0
10	beneficiary_summary	bene_hi_cvrage_tot_	integer	2	1031348	100000	0
11	beneficiary_summary	bene_smi_cvrage_to	integer	2	1031348	100000	0
12	beneficiary_summary	bene_hmo_cvrage_t	integer	2	1031348	100000	0
13	beneficiary_summary	plan_cvrg_mos_num	integer	2	1031348	100000	0
14	beneficiary_summary	sp_alzhdmta	smallint	1	1031348	100000	0
15	beneficiary_summary	sp_chf	smallint	1	1031348	100000	0
16	beneficiary_summary	sp_chrnkidn	smallint	1	1031348	100000	0
17	beneficiary_summary	sp_cncr	smallint	1	1031348	100000	0
18	beneficiary_summary	sp_copd	smallint	1	1031348	100000	0
19	beneficiary_summary	sp_depressn	smallint	1	1031348	100000	0
20	beneficiary_summary	sp_diabetes	smallint	1	1031348	100000	0
21	beneficiary_summary	sp_ischmcht	smallint	1	1031348	100000	0
22	beneficiary_summary	sp_osteoprs	smallint	1	1031348	100000	0
23	beneficiary_summary	sp_ra_oa	smallint	1	1031348	100000	0
24	beneficiary_summary	sp_strketia	smallint	1	1031348	100000	0
25	beneficiary_summary	medreimb_ip	numeric	9	1031348	100000	0
26	beneficiary_summary	benres_ip	numeric	8	1031348	100000	0

Overview | beneficiary_summary | carrier_claims | inpatient_claims | outpatient_claims | prescription_drug_events

Creating ETL specification

1 Analyze Data

- Review the source data table by table, field by field
- Study the data dictionary
- Study any other supporting

2 Work with Data Owners

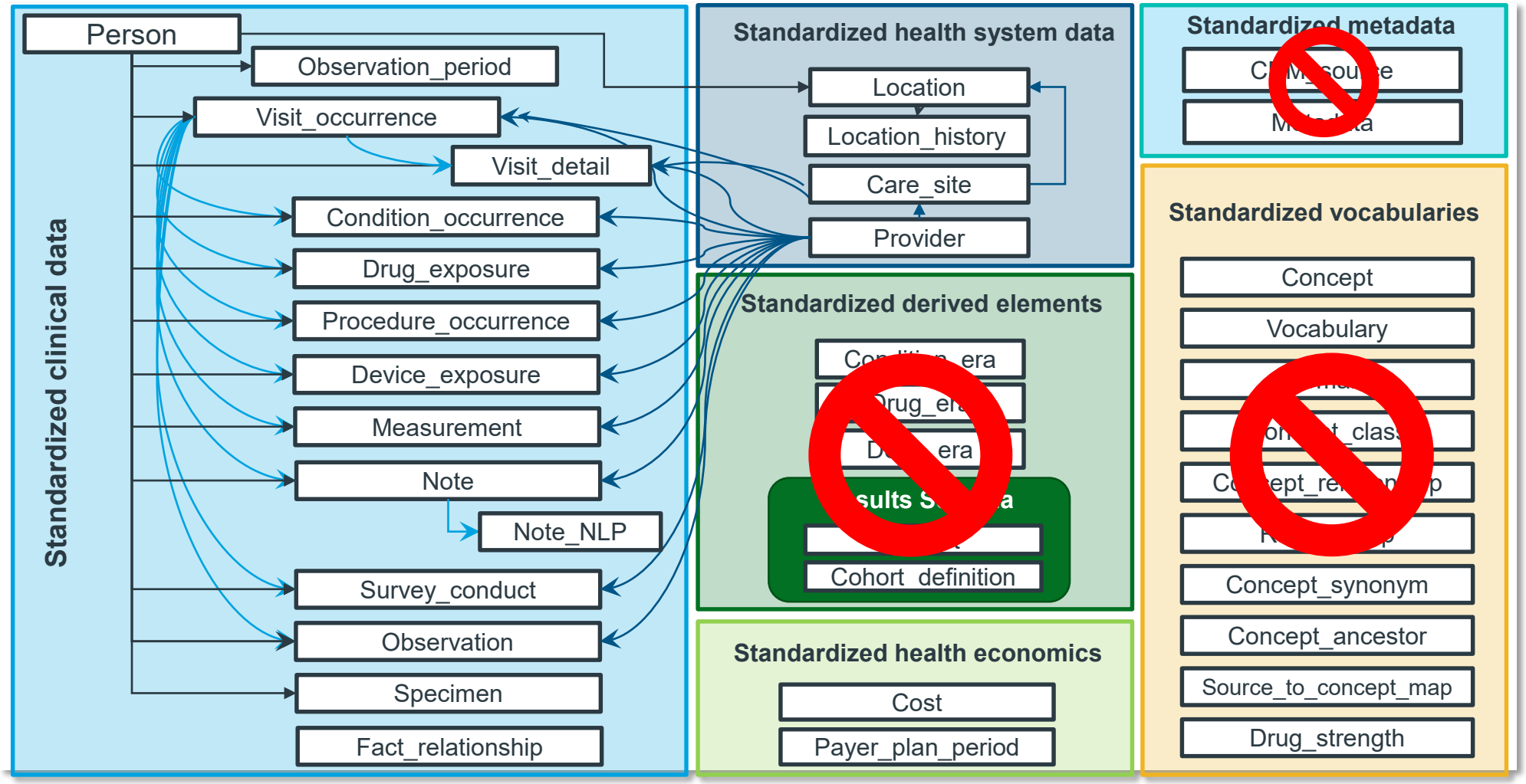
- Confirm your understanding of the data
- Ask questions on things that are not clear

3 Continued Project Review

- Review with team
- Review with data owners

Destination Field	Source Field	Applied Rule
Person_Id		System generated id based on unique source identifier
Gender_concept_id	Bene_sex_ident_cd	If 1 then ‘8507’ If 2 then ‘8532’ All else/unknown = 0
Year_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘YYYY’. Exclude patients with NULL or invalid year of birth
Month_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘MM’.
Day_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘DD’.

CDM sections not covered in ETL spec



Source code mapping to standards

Concept Code – F17.22

Concept Table – Source Concept

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
45591117	Nicotine dependence, chewing tobacco	Condition	ICD10CM	5-char nonbill code	NULL	**F17.22**

Concept Relationship Table

concept_id_1	concept_id_2	relationship_id	valid_start_date	valid_end_date	invalid_reason
45591117	4218741	Maps to	1/1/1970 0:00	12/31/2099 0:00	NULL
45591117	4209423	Maps to	1/1/1970 0:00	12/31/2099 0:00	NULL

Concept Table – Standard Concept

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
4209423	Nicotine dependence	Condition	SNOMED	Clinical Finding	**S**	56294008
4218741	Chews tobacco	Observation	SNOMED	Clinical Finding	**S**	81703003

```
SELECT  *
FROM concept c
LEFT JOIN concept_relationship cr ON c.concept_id = cr.concept_id_1 AND cr.relationship_id = 'Maps to'
LEFT JOIN concept c2 ON cr.concept_id_2 = c2.concept_id
WHERE c.concept_code = ‘F17.22'
```

One source field can go to multiple CDM domains

An example showing source Diagnosis table (diagnosis_code) can be mapped to different domains

diagnosis_code (ICD9CM)	diagnosis_description
525.5	Partial Edentulism
V26.33	Genetic Counseling
V18.2	Family History of Anemia
790.2	Abnormal Glucose

concept_id (standard)	concept_name (standard)	domain_id
40481091	Partial edentulism	**Condition**
4196362	Genetic counseling	**Procedure**
4167217	Family history of clinical finding	**Observation**
4149519	Glucose measurement	**Measurement**

Custom source code mapping

Unmapped Codes

- No existing source code mapping
- No source codes, only text
- Medical coding system doesn't exist in OHDSI

How much mapping is needed?

Usagi

- Free OHDSI tool
- Text based similarity search
- English only

What is done?

- Analyst manually map source codes
- Review with internal stakeholders

Vocabulary Team

- Paid by IQVIA
- Group of medical and technical experts

What is done?

- Send the source codes
- Give us back the mapping
- Review with our internal stakeholders

Difficulties of custom mapping

Requires medical expertise

Non-English descriptions

Time consuming

- No capacity to custom map thousands of codes
- Instead focus on most frequent

Requires updating

- A need to revisit custom mapping
- New codes added
- Old standard concepts become invalid

route_code	route_desc	route_code_vocab	count	% of total
C38288	Oral	NCIT	442,115	68%
C38216	Inhalation	NCIT	81,769	81%
C38304	Topically	NCIT	56,214	89%
C38299	Subcutaneous Injection	NCIT	16,390	92%
C38276	IV Push Slowly	NCIT	7,354	93%
C28161	Intramuscular	NCIT	5,453	94%
C38216	Nebulized inhalation	NCIT	4,386	95%
C38300	Sublingual	NCIT	4,275	95%
C38284	Nares, Both	NCIT	3,926	96%
C38274	Intravenous Push	NCIT	3,695	96%
C38276	Intravenous Infusion	NCIT	3,682	97%
C38299	Subcutaneous Infusion	NCIT	3,564	98%
C38287	Both eyes	NCIT	1,808	99%
C38246	Gastrostomy/PEG Tube	NCIT	979	99%
C38313	Vaginally	NCIT	419	100%

95%

Privacy considerations

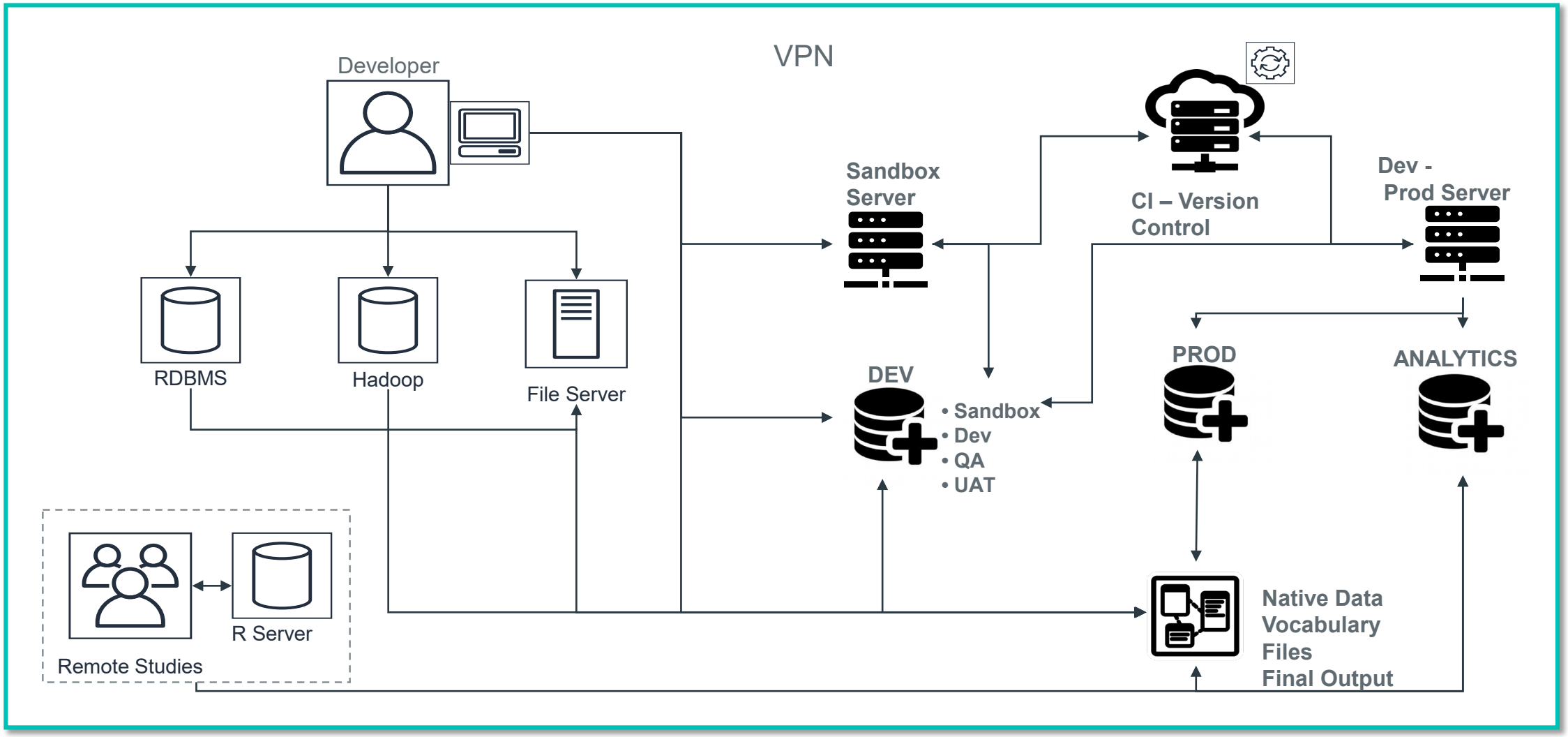
Privacy manipulation can happen at 3 tiers: source data, OMOP data and client delivery

Source data tier + **Data elements are masked at the source level**
Example: Clinical event dates are jittered in source tables

OMOP CDM tier + **Privacy manipulation happened at the OMOP CDM level**
Example: Death dates are not allowed to be loaded into OMOP CDM

Client delivery tier + **Some privacy information are not delivered to clients**
Example: Psychological related clinical conditions are masked during delivery to external clients

ETL environments



15 Minute Break

Data Quality

Overview of Data Quality

Source Data

- Unique fields and values
- No cleansing

Converted Data

- Data can still be dirty
- Tools for awareness of quality

Data Quality

Field Mapping

- Taking from source to converted
- Assumptions can cause incorrect mapping

Concept Mapping

- Understanding CDM
- Understanding vocabualaries

Source Data

1

Status Field

"Entered in Error", "Canceled", "Unauthorized"

2

ICD 9 versus 10

Indicator Field is NULL

V23 (ICD9 – Pregnancy, ICD10 – Motorcyle Accident)

3

Place Holders

"XXYX", "ABC","0000"

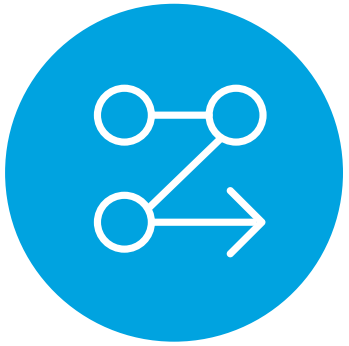
Converted Data

Implausible Values

Example: Body Temperature less than 93 and higher than 113

"Duplicates"

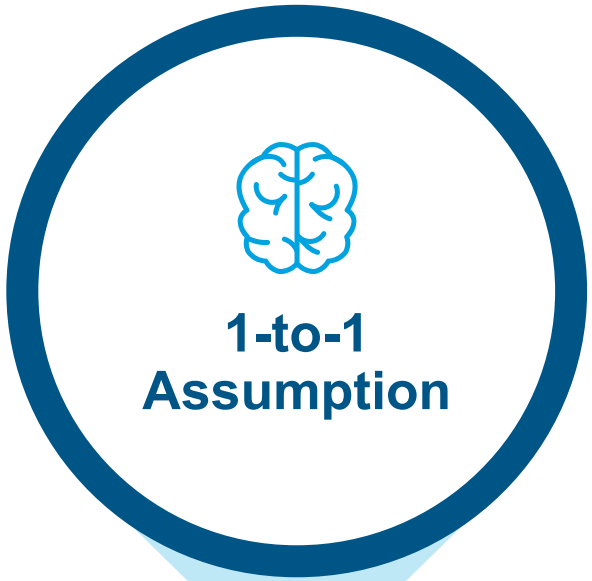
Example: Multiple records on same day, no indication which is erroneous



John Doe's

Example: Fake patients used for testing systems

Field Mapping

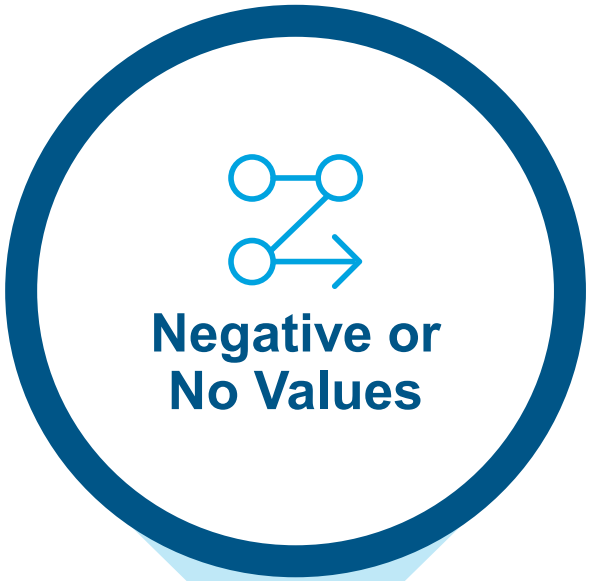


1-to-1 Assumption

- All ICD fields go to Condition
- All CPT4 fields go to Procedure

Dates

- Too far in future
- Too far in past
- Dates before birth/after death



Negative or No Values

- Measurements
- Days Supply
- Procedure Quantity

Concept Mapping

Vocabulary

- Incorrect domains
- Non-standards in Standard fields

Custom Mapping

- Manual process with room for error
- Mapping to '0'

Mapping Upwards

- SNOMED chosen vocabulary
- Correct methodology: ICD9 → SNOMED ←ICD10
- Incorrect methodology: ICD9 → SNOMED → ICD10

Data Quality Dashabord – System Requirements

FULL_201905_SOURCE_DATA

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

FULL_201905_SOURCE_DATA

Results generated at 2019-11-21 06:35:57 in 4 days

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1611	228	1839	88%	274	13	287	95%	1885	241	2126	89%
Conformance	590	91	681	87%	97	7	104	93%	687	98	785	88%
Completeness	329	57	386	85%	13	2	15	87%	342	59	401	85%
Total	2530	376	2906	87%	384	22	406	95%	2914	398	3312	88%

FULL_201905_SOURCE_DATA

OVERVIEW

METADATA

RESULTS

ABOUT

R Installation

```
install.packages("devtools")
devtools::install_github("OHDSI/DataQualityDashboard")
```

Getting Started

To install the latest stable version, install from CRAN:

```
install.packages("DatabaseConnector")
```

To install the latest development version, install from GitHub:

```
install.packages("devtools")
devtools::install_github("ohdsi/DatabaseConnectorJars")
devtools::install_github("ohdsi/DatabaseConnector")
```

To download and use the JDBC drivers for BigQuery, Impala, or Netezza, see these instructions.

Information on Database Connector

GitHub Site https://github.com/OHDSI/DatabaseConnector

Instructions for BigQuery, Impala or Netezza Extra Instructions

Vignette Using DatabaseConnector

Package Manual DatabaseConnector manual

Executing Data Quality Dashboard

- Go to OHDSI GitHub Site (https://github.com/OHDSI/DataQualityDashboard)
- Copy and paste R Scripts
- Edit DatabaseConnector command as needed (manual can be found on links in previous slide)

Executing Data Quality Checks

```
# fill out the connection details ----------------------------------------------------------------------
connectionDetails <- DatabaseConnector::createConnectionDetails(dbms = "",
                                                                user = "",
                                                                password = "",
                                                                server = "",
                                                                port = "",
                                                                extraSettings = "")

cdmDatabaseSchema <- "yourCdmSchema" # the fully qualified database schema name of the CDM
resultsDatabaseSchema <- "yourResultsSchema" # the fully qualified database schema name of the results schema (that
cdmSourceName <- "Your CDM Source" # a human readable name for your CDM source

# determine how many threads (concurrent SQL sessions) to use ---------------------------------------
numThreads <- 1 # on Redshift, 3 seems to work well

# specify if you want to execute the queries or inspect them ----------------------------------------
sqlOnly <- FALSE # set to TRUE if you just want to get the SQL scripts and not actually run the queries

# where should the logs go? -------------------------------------------------------------------------
outputFolder <- "output"

# logging type --------------------------------------------------------------------------------------
verboseMode <- FALSE # set to TRUE if you want to see activity written to the console

# write results to table? ---------------------------------------------------------------------------
writeToTable <- TRUE # set to FALSE if you want to skip writing to a SQL table in the results schema

# if writing to table and using Redshift, bulk loading can be initialized ---------------------------

# Sys.setenv("AWS_ACCESS_KEY_ID" = "",
#            "AWS_SECRET_ACCESS_KEY" = "",
#            "AWS_DEFAULT_REGION" = "",
#            "AWS_BUCKET_NAME" = "",
#            "AWS_OBJECT_KEY" = "",
```

Exploring Dashboard

Filtering Options

Select Columns to Show

Export to CSV

Column visibility | CSV

Show 5 entries

Search:

	STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
⊞	PASS	Validation	Completeness	None	TABLE	The number and percent of persons in the CDM that do not have at least one record in the NOTE table (Threshold=100%).	100.00%
⊞	PASS	Validation	Completeness	None	TABLE	The number and percent of persons in the CDM that do not have at least one record in the PAYER_PLAN_PERIOD table (Threshold=100%).	100.00%
⊞	PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the month_of_birth of the PERSON. (Threshold=100%).	100.00%
⊞	PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the day_of_birth of the PERSON. (Threshold=100%).	100.00%
⊞	PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the birth_datetime of the PERSON. (Threshold=100%).	100.00%

Showing 1 to 5 of 3,312 entries

Previous 1 2 3 4 5 ... 663 Next

Over 3,000 Checks

Explaining Results

Category
- Completeness
- Conformance
- Plausibilty

Status
- Pass
- Fail

Level
- Concept
- Table
- Field

Context
- Verification
- Validation

Subcategory
- Atemporal
- Computational
- Relational
- Temporal
- Value
- None


Q&A

How to do research using OMOP

Tools used for OMOP Research

Athena
Free OHDSI online vocabulary browsing tool

Atlas
Free OHDSI analytic tool to support cohort development

SQL
AWS Redshift environment and other flavors of SQL

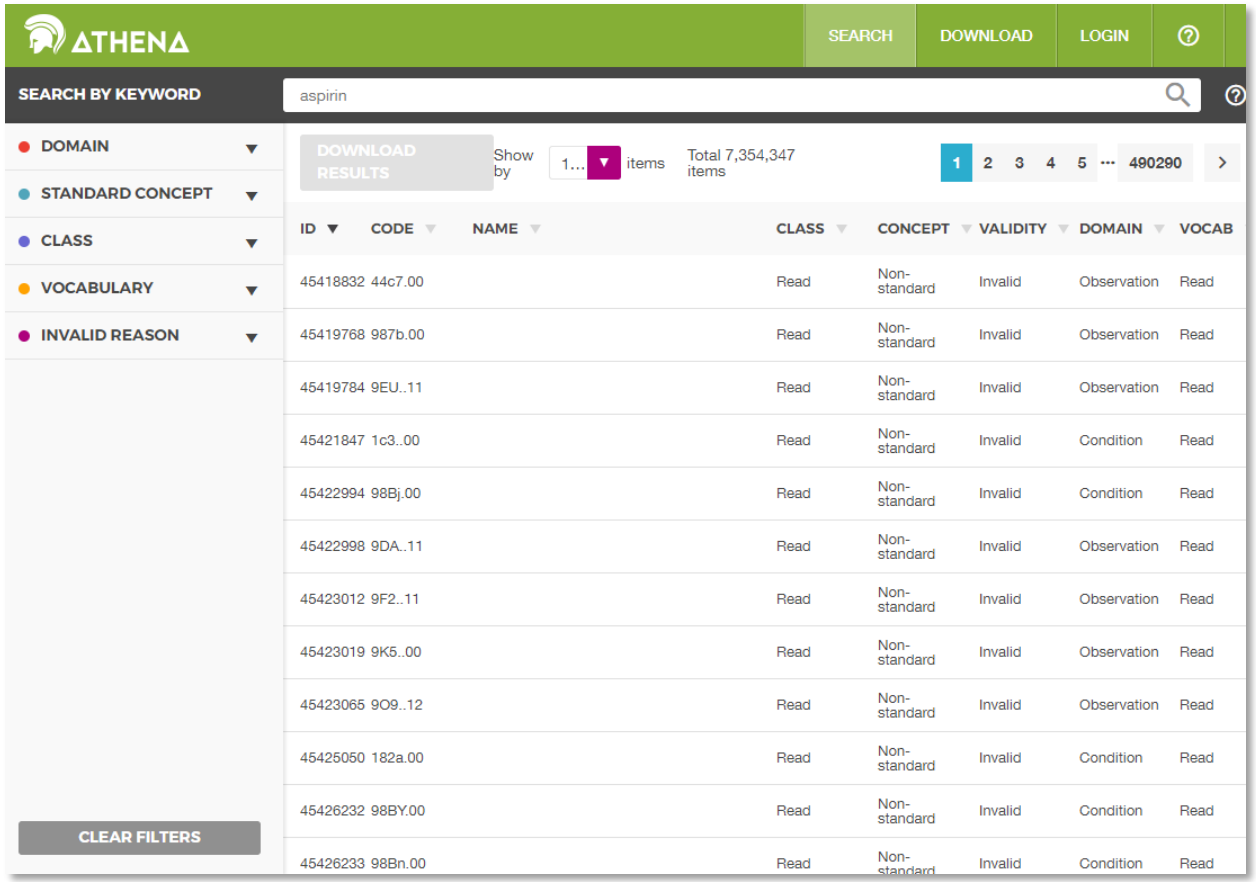
R
Statistical analysis coding program

Athena

Description

- Web-based open-sourced software application
- Developed by the OHDSI community
- Allows faceted search of the vocabularies
- Downloadable vocabulary feature
- User-friendly interface

Screenshot

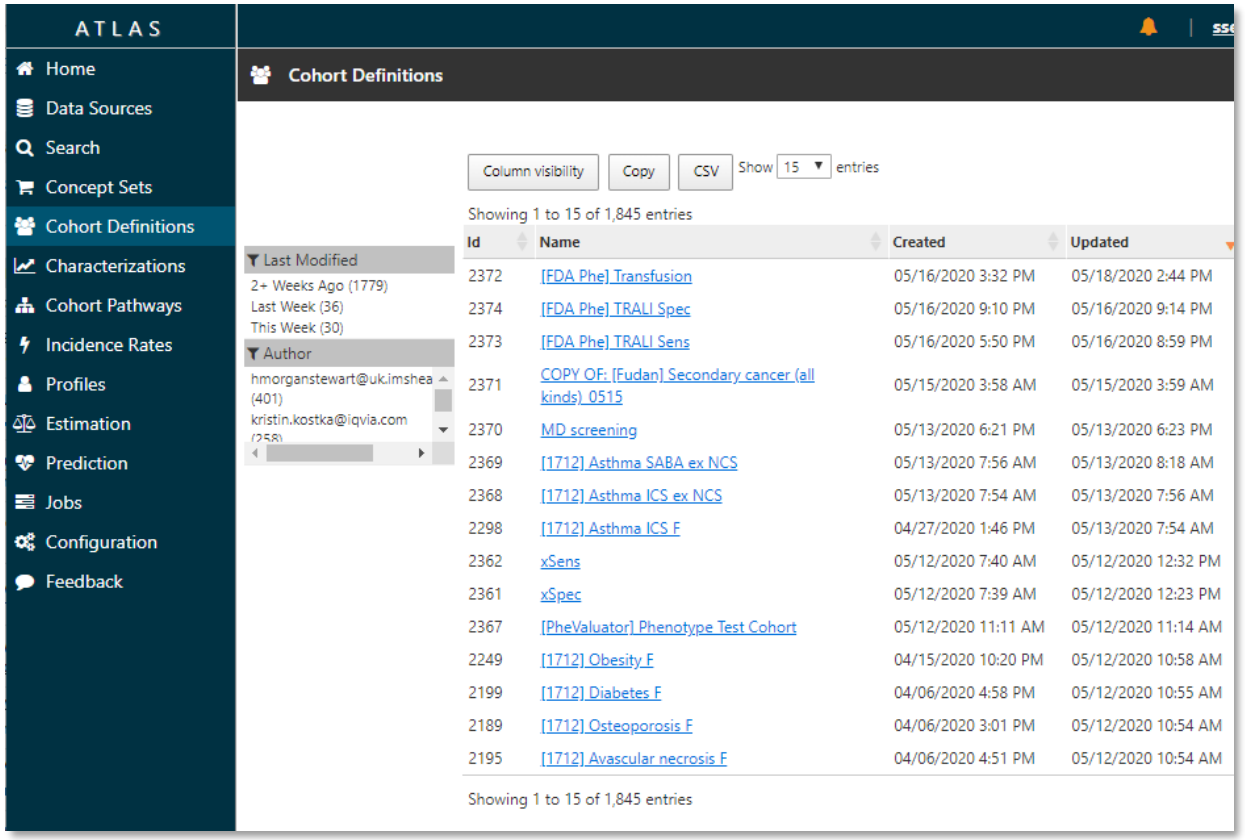


ATLAS

Description

- Web-based open-sourced software application
- Developed by the OHDSI community
- Free and publicly available
- User friendly interface

Screenshot

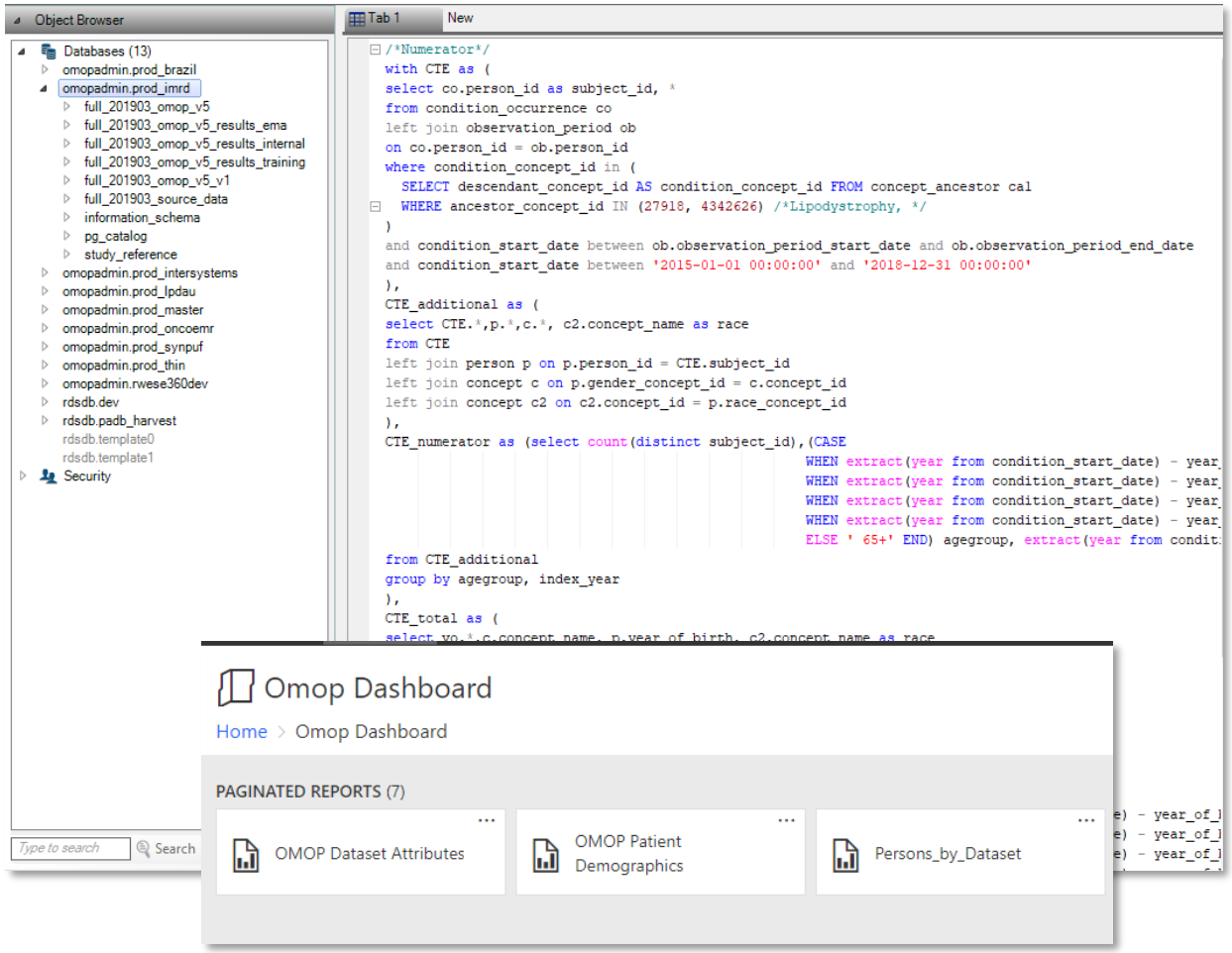


SQL

Description

- Database querying application
- OMOP team uses Redshift by AWS
- In addition, used for OMOP conversions

Screenshot

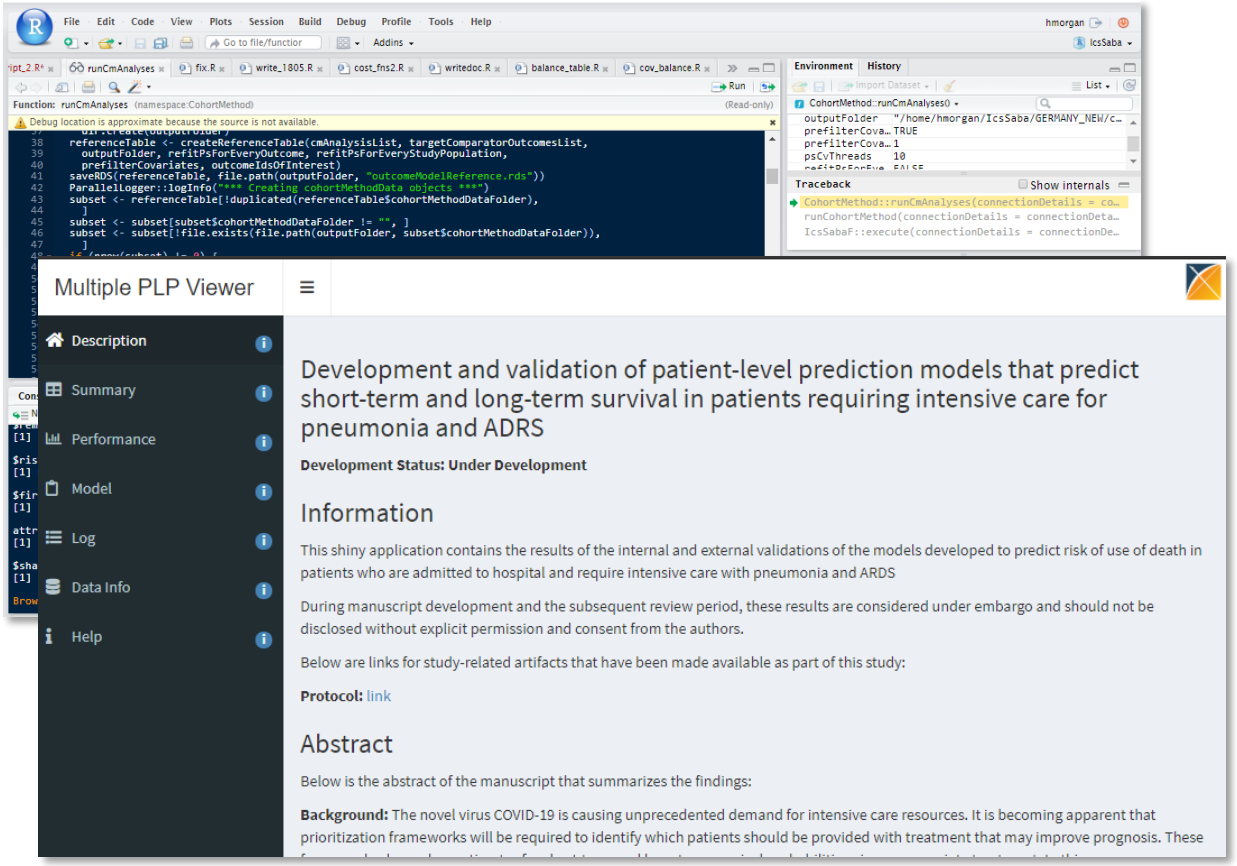


R and R Shiny

Description

- Open-sourced application
- Shiny is an R package
- Interactive web applications
- Enables easy sharing of aggregated results and visualizations

R and R shiny screenshot



OMOP data science tool matrix

When to use what tool?

	ATHENA		SQL	R
Vocabulary look-up	▲	▲		
Database counts	▲	▲	▲	▲
Population counts		▲	▲	▲
Characterisations		▲	▲	▲
Incidence		▲	▲	▲
Prevalence		▲	▲	▲
Treatment patterns & pathways		▲	▲	▲
Patient-level predictions		▲		▲
Population-level estimation		▲		▲
Data visualisations e.g. sunburst plots				▲

When to use what?

Atlas, R, SQL

Atlas

Pros:

- User-friendly
- Pre-defined functions
- Easy to share

Cons:

- Limited functions
- Unable to perform prediction or estimation studies

R

Pros:

- Can manipulate data
- More functions available e.g. build models, loops, etc
- Choice of visualisations

Cons:

- Requires proper set-up
- Requires programming skills
- More validation/reviews required

SQL

Pros:

- ETL Conversions
- Can manipulate data
- Data visualizations e.g. via dashboarding

Cons:

- Requires proper set-up
- Requires programming skills
- More validation/reviews required

Complex cohorts are quick and easy to define

Cohort definitions using ATLAS require no coding and are easily understood by non-technical stakeholders

Cohort Entry Events

Events having any of the following criteria:

+ Add Initial Event

a drug exposure of dabigatran

+ Add attribute... | Delete Criteria

✖ for the first time in the person's history

✖ occurrence start is: On or After 2010-10-19

✖ with age Greater or Equal To 65

with continuous observation of at least 183 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Restrict initial events

Inclusion Criteria

New inclusion criteria

1. Has prior atrial fibrillation of atrial flutter diagnosis
2. Has no prior treatment with comparator drug (warfarin)
3. Has no prior treatment with other anticoagulants (rivaroxaban or apixaban)
4. Not in a skilled nursing facility or nursing home, or receiving hospice care on the index date
5. Not undergoing dialysis or kidney transplant recipient
6. No mitral valve disease, heart valve repair, or replacement in the prior 6 months
7. No deep vein thrombosis or pulmonary embolism in the prior 6 months
8. No joint replacement surgery in the prior 6 months

Has prior atrial fibrillation of atrial flutter diagnosis | Copy | Delete

enter an inclusion rule description

having any of the following criteria:

+ Add criteria to group...

with at least 1 using all occurrences of:

a condition occurrence of Atrial fibrillation

+ Add attribute... | Delete Criteria

where event starts between All days Before and 0 days After index start date *add additional constraint*

☐ restrict to the same visit occurrence

☐ allow events from outside observation period

or with at least 1 using all occurrences of:

a condition occurrence of Atrial flutter

+ Add attribute... | Delete Criteria

where event starts between All days Before and 0 days After index start date *add additional constraint*

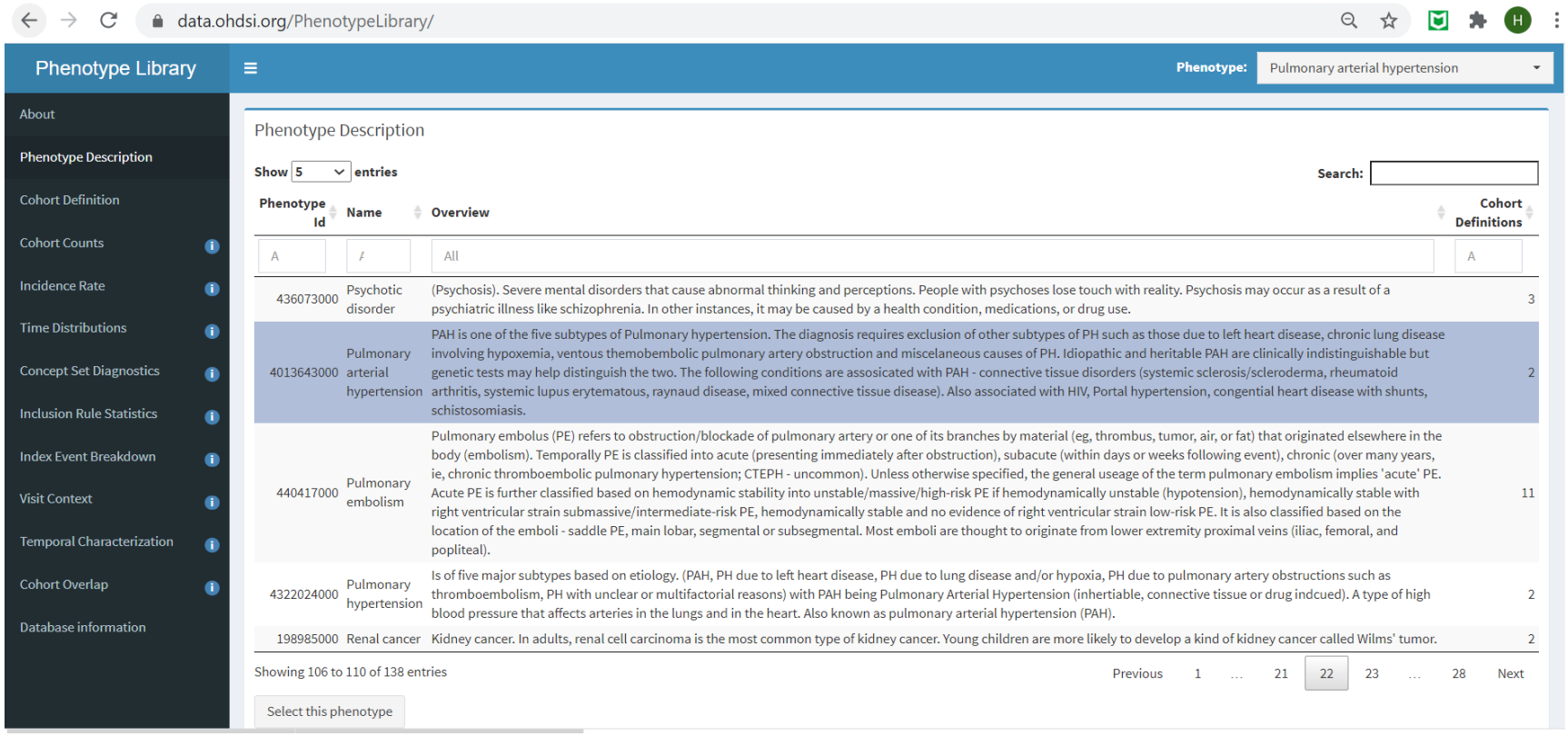
☐ restrict to the same visit occurrence

☐ allow events from outside observation period

Limit qualifying events to: earliest event per person.

The OHDSI phenotype library is growing all the time

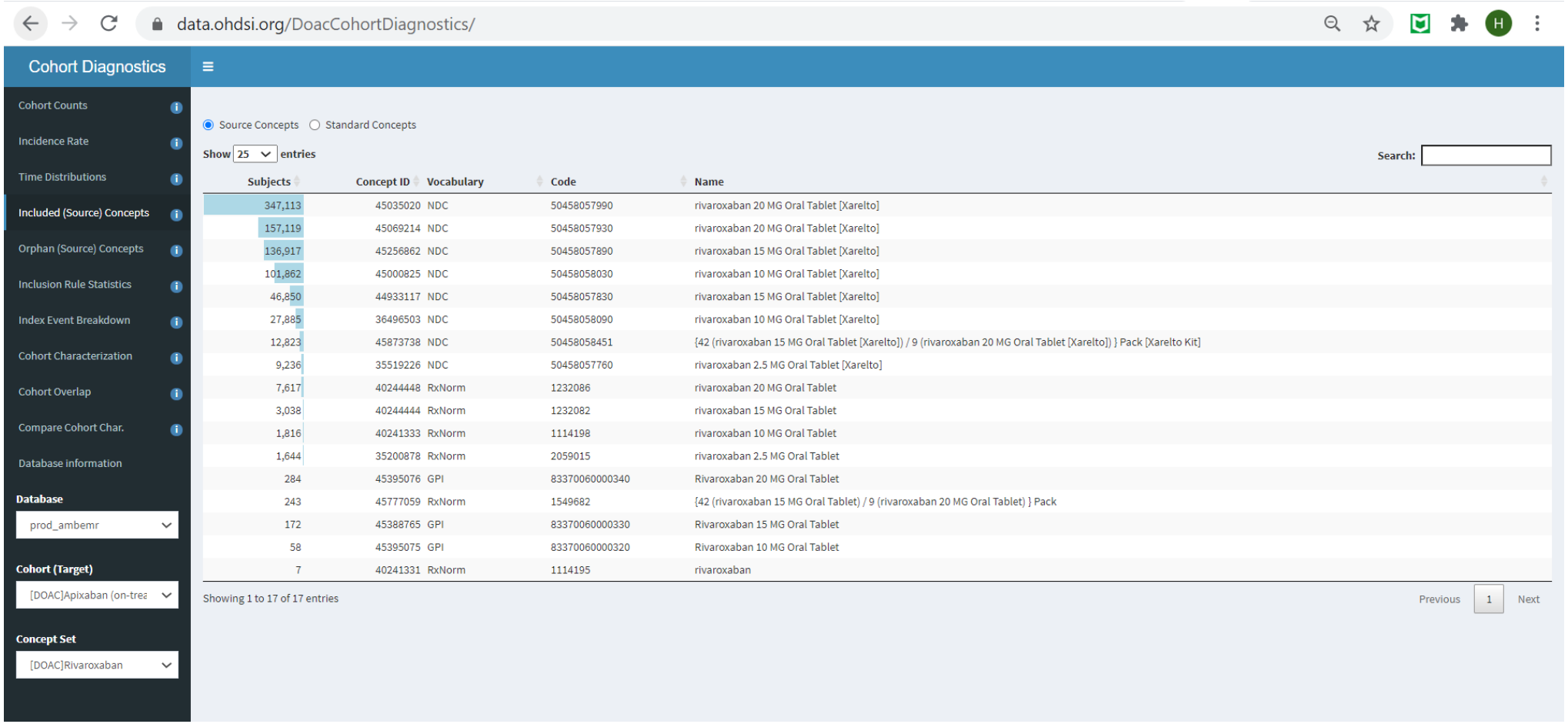
Community phenotypes can be used 'out of the box'



https://data.ohdsi.org/PhenotypeLibrary/

Cohorts can be validated using the Cohort Diagnostics tool

Check for missing codes, prevalence and cohort characteristics



Subjects	Concept ID	Vocabulary	Code	Name
347,113	45035020	NDC	50458057990	rivaroxaban 20 MG Oral Tablet [Xarelto]
157,119	45069214	NDC	50458057930	rivaroxaban 20 MG Oral Tablet [Xarelto]
136,917	45256862	NDC	50458057890	rivaroxaban 15 MG Oral Tablet [Xarelto]
101,862	45000825	NDC	50458058030	rivaroxaban 10 MG Oral Tablet [Xarelto]
46,850	44933117	NDC	50458057830	rivaroxaban 15 MG Oral Tablet [Xarelto]
27,885	36496503	NDC	50458058090	rivaroxaban 10 MG Oral Tablet [Xarelto]
12,823	45873738	NDC	50458058451	{42 (rivaroxaban 15 MG Oral Tablet [Xarelto]) / 9 (rivaroxaban 20 MG Oral Tablet [Xarelto]) } Pack [Xarelto Kit]
9,236	35519226	NDC	50458057760	rivaroxaban 2.5 MG Oral Tablet [Xarelto]
7,617	40244448	RxNorm	1232086	rivaroxaban 20 MG Oral Tablet
3,038	40244444	RxNorm	1232082	rivaroxaban 15 MG Oral Tablet
1,816	40241333	RxNorm	1114198	rivaroxaban 10 MG Oral Tablet
1,644	35200878	RxNorm	2059015	rivaroxaban 2.5 MG Oral Tablet
284	45395076	GPI	83370060000340	Rivaroxaban 20 MG Oral Tablet
243	45777059	RxNorm	1549682	{42 (rivaroxaban 15 MG Oral Tablet) / 9 (rivaroxaban 20 MG Oral Tablet) } Pack
172	45388765	GPI	83370060000330	Rivaroxaban 15 MG Oral Tablet
58	45395075	GPI	83370060000320	Rivaroxaban 10 MG Oral Tablet
7	40241331	RxNorm	1114195	rivaroxaban

https://data.ohdsi.org/DoacCohortDiagnostics/
https://ohdsi.github.io/Hades/

Analytical packages

Highly parameterized tools for characterization, cohort studies (PLE) and prediction studies (PLP)

Prediction and estimation methods

Cohort Method
New-user cohort studies using large-scale regression for propensity and outcome models

Self-Controlled Case Series
Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality.

Self-Controlled Cohort
A self-controlled cohort design, where time preceding exposure is used as control.

Patient Level Prediction
Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms.

Case-control
Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort.

Case-crossover
Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control).

Method characterization

Empirical Calibration
Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design.

Method Evaluation
Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods.

Evidence Synthesis
Combining study diagnostics and results across multiple sites.

Supporting packages

Database Connector
Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.

Sql Render
Generate SQL on the fly for the various SQL dialects.

Cyclops
Highly efficient implementation of regularized logistic, Poisson and Cox regression.

ParallelLogger
Support for parallel computation with logging to console, disk, or e-mail.

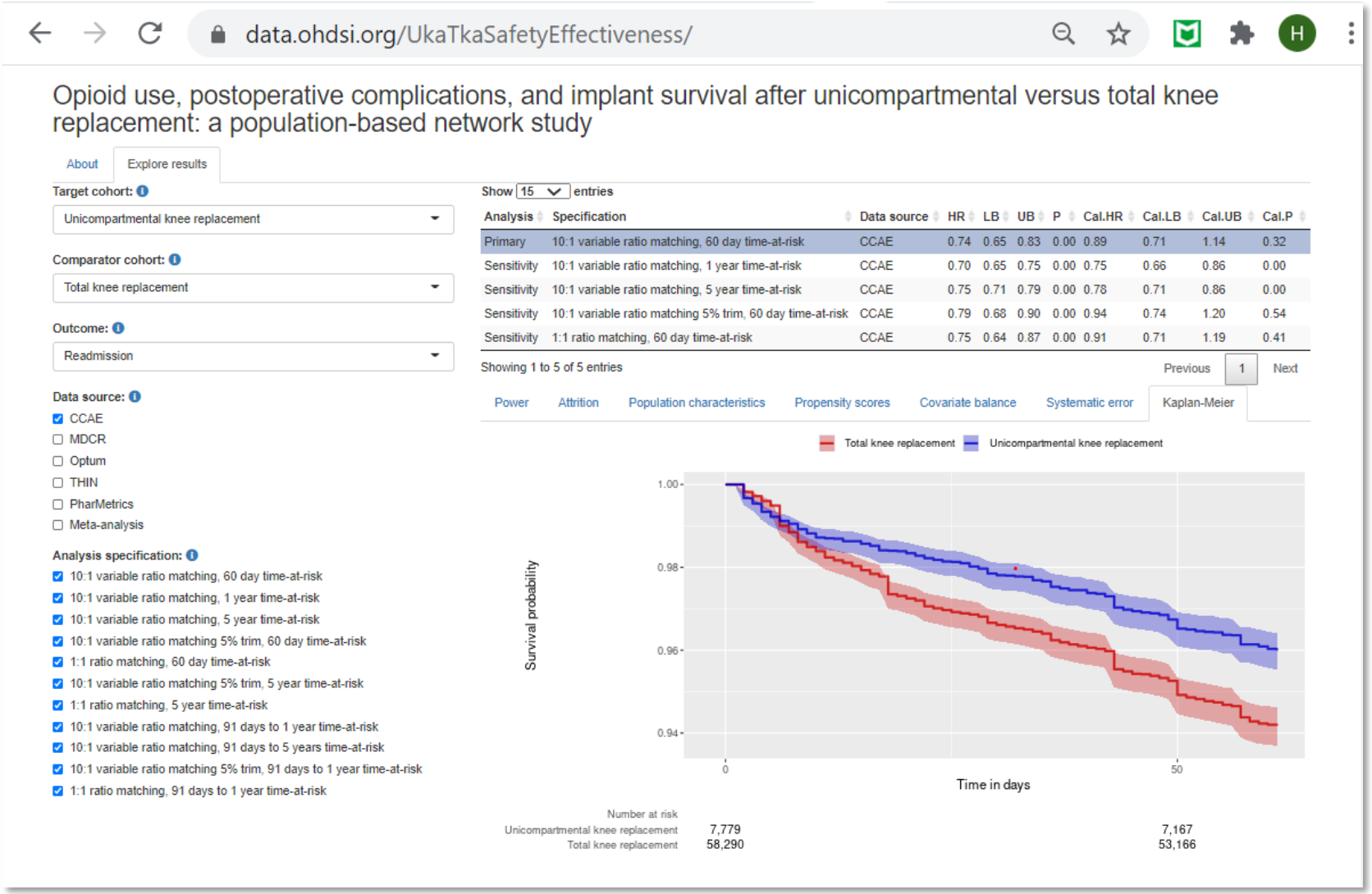
Feature Extraction
Automatically extract large sets of features for user-specified cohorts using data in the CDM.

- Run a complex cohort study or prediction study with minimal coding
- Just define the study in ATLAS to generate an R package
- No need for complicated communication between epi and developers
- Code has already been QC'd and can be used

https://ohdsi.github.io/Hades/packages.html

Standardized outputs for easy interpretation

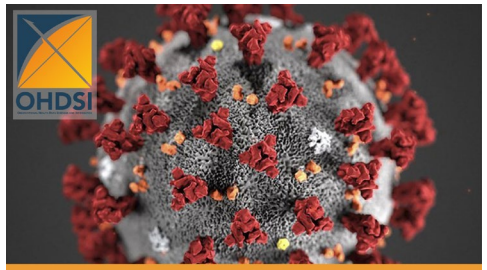
In time, stakeholders know what to expect and results are easy to digest



Analysis	Specification	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
Primary	10:1 variable ratio matching, 60 day time-at-risk	CCAE	0.74	0.65	0.83	0.00	0.89	0.71	1.14	0.32
Sensitivity	10:1 variable ratio matching, 1 year time-at-risk	CCAE	0.70	0.65	0.75	0.00	0.75	0.66	0.86	0.00
Sensitivity	10:1 variable ratio matching, 5 year time-at-risk	CCAE	0.75	0.71	0.79	0.00	0.78	0.71	0.86	0.00
Sensitivity	10:1 variable ratio matching 5% trim, 60 day time-at-risk	CCAE	0.79	0.68	0.90	0.00	0.94	0.74	1.20	0.54
Sensitivity	1:1 ratio matching, 60 day time-at-risk	CCAE	0.75	0.64	0.87	0.00	0.91	0.71	1.19	0.41

OMOP research examples

OHDSI Research Program for COVID-19



OHDSI Kicks Off COVID-19 Research Agenda With 4-Day International Virtual Study-A-Thon

Overview:

- OHDSI's international call to action to generate real-world evidence and inform the COVID-19 pandemic response
- OHDSI community invited to collaborate
- Over 350 participants from 30 countries collaborated on Erasmus MS Teams platform
- 37 databases from 10 countries on 3 continents including 8 databases with COVID-19 patients
- Aims to design and execute a series of observational studies

Research tracks:

- Systematic literature review
- Phenotype development
- ***Characterization studies:*** prognosis and natural history
- ***Population-level effect estimation:*** understanding treatment effectiveness and safety
- ***Patient-level prediction studies:*** prediction of patient outcomes for disease severity and healthcare resource utilization

The OHDSI Research Network

**Virtual Study-a-thon, 26-29 March 2020*

Large-Scale Evidence Generation and Evaluation Across a Network of Databases (LEGEND)

THE LANCET

ARTICLES | ONLINE FIRST

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Prof Marc A Suchard, MD • Martijn J Schuemie, PhD • Prof Harlan M Krumholz, MD • Seng Chan You, MD • RuiJun Chen, MD • Nicole Pratt, PhD • et al. Show all authors

Published: October 24, 2019 • DOI: https://doi.org/10.1016/S0140-6736(19)32317-7 • Check for updates

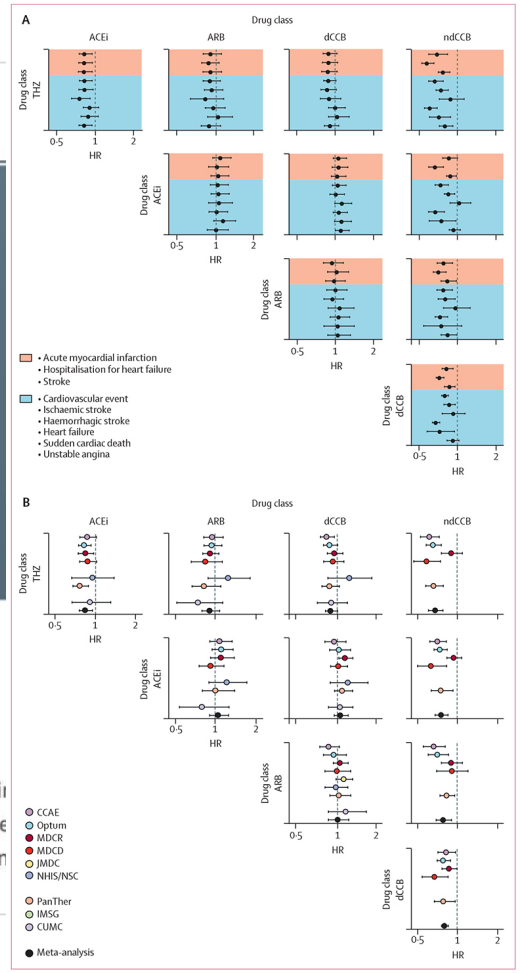
Article Info

Linked Articles

Summary

Background

Uncertainty remains about the optimal monotherapy for hypertension, with current guidelii among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting e blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium chan indications. Randomised trials have not further refined this choice.



Study code: http://www.github.com/ohdsi/LEGEND

Validation through EMA - Consistency between Source and CDM data

Clin Pharmacol Ther. 2020 Apr;107(4):915-925. doi: 10.1002/cpt.1785. Epub 2020 Mar 2.

Can We Rely on Results From IQVIA Medical Research Data UK Converted to the Observational Medical Outcome Partnership Common Data Model?: A Validation Study Based on Prescribing Codeine in Children

Gianmario Candore [1], Karin Hedenmalm [1], Jim Slattery [2], Alison Cave [2], Xavier Kurz [2], Peter Arlett [2]

Affiliations — collapse

Affiliations

1 Business Data Department, European Medicines Agency, Amsterdam, The Netherlands.

2 Pharmacovigilance and Epidemiology Department, European Medicines Agency, Amsterdam, The Netherlands.

PMID: 31956997 DOI: 10.1002/cpt.1785

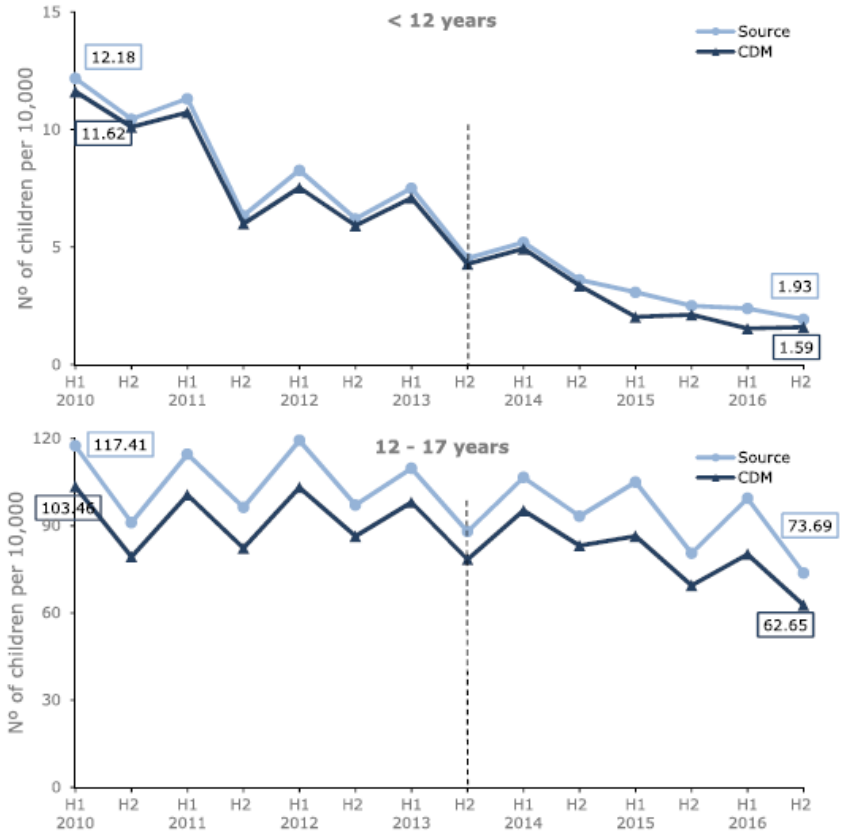


Figure 1-2: Six-monthly prevalence (per 10,000) of codeine prescribing for pain in 0–17 years

Example study & Exercise

Example study – treatments and outcomes of influenza patients during hospital stay

- **Study Topic:**
  - Baseline demographic and clinical characteristics, treatment patterns and outcomes of patients diagnosed with influenza initiating treatment in the US hospital setting: a retrospective cohort study using administrative data.
- **Objectives:**
  - *Primary Objectives*
    - Describe the treatment patterns of hospitalized influenza patients including drugs:
      - (a) antivirals – peramivir, zanamivir, oseltamivir phosphate, baloxavir marboxil (b) antibiotics (c) corticosteroids

      and the following procedures:
      - (a) mechanical ventilation (b) tracheostomy (c) extracorporeal membrane oxygenation (d) oxygen therapy
    - Describe the length of the hospital stay by line of treatment and conditions of interest:
      - (a) diabetes (b) lung disease (c) cancer (d) immunodeficiency (e) heart disease (f) hypertension (g) asthma (h) kidney disease
  - *Secondary Objectives*
    - Describe the baseline demographics and clinical characteristics of hospitalized influenza patients.

Example study – Cohort definitions

- **Study Population**
  - Persons hospitalized during the 2008-2009 influenza season with a diagnosis of influenza 21 days prior or during the hospital stay, with no prior continuous enrollment required and with no influenza hospitalization in the 6 months prior to hospital admission.
- **Inclusion Criteria**
  - Patients with claims for a hospital stay between 1st September 2008 and 1st April 2009 (index date). All hospital stays during the study period are of interest.
  - Patient is ≥ 18 years of age at index date.
  - Patient has at least 1 diagnosis of influenza 21 days prior to index start date (hospital admission) or up to index end date (hospital discharge date).
  - Patient has 0 months of prior continuous enrollment prior to hospital admission.
  - EXCLUDE patients with evidence of hospitalization for influenza in the 6 months prior to index date.

Exercise – Find the OMOP Standard concepts

- influenza
  - *OMOP concept_id = 4266367*
- type 2 diabetes
- lung disease
- cancer
- immunodeficiency
- heart disease
- hypertension
- asthma
- kidney disease

Homework: Find the standard concept(s) for these disease

ATHENA

SEARCH | DOWNLOAD | LOGIN

Influenza

DETAILS

Domain ID	Condition
Concept Class ID	Clinical Finding
Vocabulary ID	SNOMED
Concept ID	4266367
Concept code	6142004
Validity	Valid
Concept	Standard
Synonyms	Grippe Influenza (disorder) Flu
Valid start	31-Jan-2002
Valid end	31-Dec-2099

TERM CONNECTIONS (103) HIERARCHY | RELATED CONCEPTS

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABULARY
Active possibly_equivalent_to inactive (SNOMED)	(Influenza NOS) or (influenza-like illness)	40345755	SNOMED
	(Influenza like illness) or (influenza NOS)	40395532	SNOMED
Active same_as inactive (SNOMED)	Influenza	40316526	SNOMED
Active was_a inactive (SNOMED)	Influenza NOS	3573522	SNOMED
	Influenza NOS	4144103	SNOMED
	Influenza with other manifestations	3531375	SNOMED
	Influenza with other manifestations	4110634	SNOMED
	Influenza with other manifestations NOS	3531376	SNOMED
	Influenza with other manifestations NOS	4110043	SNOMED
	Influenza with other respiratory manifestation	4112663	SNOMED
	Influenza with respiratory manifestations NOS	3536147	SNOMED
	Influenza with respiratory manifestations NOS	4110042	SNOMED
	[X]Influenza with other manifestations, influenza virus identified	44796184	SNOMED
	[X]Influenza with other manifestations, virus not identified	44798590	SNOMED

https://athena.ohdsi.org/search-terms/start


Q&A

Training series plan

+ Session 1 : Course Introduction
  - OMOP CDM and vocabulary overview, OMOP conversion, data quality, examples of previous research and use cases, introducing ATLAS and OHDSI tools

+ Session 2: OMOP CDM/Vocabulary Tutorial
  - Concept, Concept mapping, Hierarchy, Ancestors, and OMOP CDM

+ Session 3: Cohort and Cohort Characterization
  - Concept sets, cohort definition, and cohort characterization

+ Session 4: Treatment Pathways and Incident Rates
  - Treatment pathways, Incident rates, and Characterization using R

Thank you