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

+ OHDSI Overview / Why OHDSI?
+ OHDSI adoptions
+ Q&A Session
+ OMOP conversion
+ Data Quality
+ Q&A Session
+ How to do research using OMOP and research examples
+ Example Study & Exercise
+ Q&A Session
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
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

What OHDSI is:

- Open Source
- Community
- Data

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

OHDSI Data Partners

Source data

Standardized database (OMOP CDM)

Standardized analytics

Analysis results

Summary statistics

OHDSI Coordinating Center

Network support
Analytics development
Research and education

OHDSI.org
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.
History of OMOP and OHDSI

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

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

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

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

Formation of Korea Chapter
- Led by the coordinating center at the Erasmus University Medical Center in Rotterdam

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

EMA Adoption
Formation of Australia, Japan and Singapore Chapters

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

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
Formation of European Chapter
- FDA Adoption (FDA BEST Launch)
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

Source of Business

Mortality

Adherence

Safety Signals

North America

Southeast Asia

China

Europe

UK

Japan

India

S.Africa

Switzerland

Italy

Israel

Standardized data
Analytics can be remote
Analytics can be behind firewall
IQVIA Research Network - Structure and participants

Customer

- Claims Asset
- EHR Asset

Technology based access control to Network

Coordinating Center

- EHR Asset
- Claims Asset

IQVIA Research Network Partners

- Network Partner
- Network Partner
- Network Partner
- Network Partner

Statistical Programming Service
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

---

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

<table>
<thead>
<tr>
<th>CDM benefits</th>
<th>Standardized tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>• One script fits all</td>
<td>• Community phenotype definitions</td>
</tr>
<tr>
<td>• No switching between dialects</td>
<td>• Comprehensive ecosystem of tools</td>
</tr>
<tr>
<td>• Modular table structure and consistent field names for easy querying</td>
<td>• High parameterization gives flexibility</td>
</tr>
<tr>
<td>• Hierarchical standard vocabularies</td>
<td>• No need to re-code complex analytics</td>
</tr>
</tbody>
</table>
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

Data Network Growth Over Time

- Data Owners
- Databases
- Countries

Patient Lives in OMOP Data Network

Total lives is not unique due to issues with deduplication of US data
NIH Adopts OMOP CDM for National COVID-19 Surveillance

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.
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.

*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.

*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:
• 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?

- Consult with OHDSI Community
- Utilize OMOP Vocabulary
- Create business rules
- Analyze source data

- Quality checks begin at source level
- Test business rules
- Achilles & Data Quality Dashboard
- Iterative process

- Perform Monthly or Quarterly Update Vocabulary
- Data Refresh

- OMOP Conversion
- Data Quality
- Research

- ATHENA
- ATLAS
- SQL
- R and R Shiny
OMOP Conversion
OMOP conversion process flow

Data experts & CDM experts together design the ETL

Medical experts create the code mappings

All are involved in quality control

A technical person implements the ETL

Analysis
- White Rabbit
- Rabbit In a Hat
- Usagi

Quality Control
- Internal Quality Checks
- Achilles
- Data Quality Dashboard

Development
- Jenkins
- Code Repository
OMOP CDM Version 5.3.1 Minimal Viable Product (MVP)

<table>
<thead>
<tr>
<th>Health System Tables</th>
<th>Clinical Data Tables</th>
<th>Derived Tables (Logic Provided)</th>
<th>Health Economic Tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Visit_Occurrence</td>
<td>Drug_Era</td>
<td>Payer_Plan_Period</td>
</tr>
<tr>
<td>Care_Site</td>
<td>Condition_Occurrence</td>
<td>Dose_Era</td>
<td>Cost</td>
</tr>
<tr>
<td>Provider</td>
<td>Drug_Exposure</td>
<td>Condition_Era</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>Procedure_Occurrence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td>Measurement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation_Period</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specimen</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Device_Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fact_Relationship</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Visit_Detail</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Note</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Note_NLP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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

- Google Big Data Query
- Hadoop
- Amazon Web Services
- OMOP CDM
- Microsoft SQL Server
- Oracle
- PostgreSQL
OMOP Agile conversion methodology

What is Agile?
• Project management & software development
• 2 week sprints
• Promotes continuous adaptation

What is Agile?

- Product Backlog Grooming
- Sprint Planning
- Daily Stand-up
- Sprint Demo
- Sprint Retrospective

During Sprint
- Review ETL Specs with Analyst
- Develop & QA ETL Conversion
- Validate
- Execute ETL Conversion

Post Sprint
- Business Validation/Sign-off
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
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
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

<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person_Id</td>
<td></td>
<td>System generated id based on unique source identifier</td>
</tr>
</tbody>
</table>
| 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

- CDM_source
- Vocabulary
- Concept
- Concept_class
- Concept_synonym
- Concept_ancestor
- Source_to_concept_map
- Drug_strength
- Payer_plan_period
- Location
- Location_history
- Care_site
- Provider
- Cohort_definition
- Person
- Survey_conduct
- Observation
- Specimen
- Fact_relationship
- Observation_period
- Visit_occurrence
- Visit_detail
- Condition_occurrence
- Drug_exposure
- Procedure_occurrence
- Device_exposure
- Measurement
- Note
- Note_NLP
- Condition_era
- Drug_era
- Procedure_era
- Results_Schema
- Standardized health economics
- Standardized metadata
- Standardized derived elements
- Standardized vocabularies
- Standardized health system data
Source code mapping to standards

**Concept Code – F17.22**

### Concept Table – Source Concept

<table>
<thead>
<tr>
<th>concept_id</th>
<th>concept_name</th>
<th>domain_id</th>
<th>vocabulary_id</th>
<th>concept_class_id</th>
<th>standard_concept</th>
<th>concept_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>45591117</td>
<td>Nicotine dependence, chewing tobacco</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>5-char nonbill code</td>
<td>NULL</td>
<td>F17.22</td>
</tr>
</tbody>
</table>

### Concept Relationship Table

<table>
<thead>
<tr>
<th>concept_id_1</th>
<th>concept_id_2</th>
<th>relationship_id</th>
<th>valid_start_date</th>
<th>valid_end_date</th>
<th>invalid_reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>45591117</td>
<td>4218741</td>
<td>Maps to</td>
<td>1/1/1970 0:00</td>
<td>12/31/2099 0:00</td>
<td>NULL</td>
</tr>
<tr>
<td>45591117</td>
<td>4209423</td>
<td>Maps to</td>
<td>1/1/1970 0:00</td>
<td>12/31/2099 0:00</td>
<td>NULL</td>
</tr>
</tbody>
</table>

### Concept Table – Standard Concept

<table>
<thead>
<tr>
<th>concept_id</th>
<th>concept_name</th>
<th>domain_id</th>
<th>vocabulary_id</th>
<th>concept_class_id</th>
<th>standard_concept</th>
<th>concept_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>4209423</td>
<td>Nicotine dependence</td>
<td>Condition</td>
<td>SNOMED</td>
<td>Clinical Finding</td>
<td>S</td>
<td>56294008</td>
</tr>
<tr>
<td>4218741</td>
<td>Chews tobacco</td>
<td>Observation</td>
<td>SNOMED</td>
<td>Clinical Finding</td>
<td>S</td>
<td>81703003</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>diagnosis_code (ICD9CM)</th>
<th>diagnosis_description</th>
<th>concept_id (standard)</th>
<th>concept_name (standard)</th>
<th>domain_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>525.5</td>
<td>Partial Edentulism</td>
<td>40481091</td>
<td>Partial edentulism</td>
<td>Condition</td>
</tr>
<tr>
<td>V26.33</td>
<td>Genetic Counseling</td>
<td>4196362</td>
<td>Genetic counseling</td>
<td>Procedure</td>
</tr>
<tr>
<td>V18.2</td>
<td>Family History of Anemia</td>
<td>4167217</td>
<td>Family history of clinical finding</td>
<td>Observation</td>
</tr>
<tr>
<td>790.2</td>
<td>Abnormal Glucose</td>
<td>4149519</td>
<td>Glucose measurement</td>
<td>Measurement</td>
</tr>
</tbody>
</table>
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

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

- Developer
  - RDBMS
  - Hadoop
  - File Server
- VPN
  - Sandbox Server
  - CI – Version Control
  - Dev - Prod Server
  - CI – Version Control
- R Server
  - Remote Studies
- PROD
  - Sandbox
  - Dev
  - QA
  - UAT
- ANALYTICS
  - Native Data Vocabulary Files
  - Final Output
15 Minute Break
Overview of Data Quality

Source Data
• Unique fields and values
• No cleansing

Converted Data
• Data can still be dirty
• Tools for awareness of 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 Dashboard – System Requirements

**DATA QUALITY ASSESSMENT**

**FULL_201905_SOURCE_DATA**

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

<table>
<thead>
<tr>
<th></th>
<th>Verification</th>
<th>Validation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
<td>Fail</td>
<td>Total</td>
</tr>
<tr>
<td>Plausibility</td>
<td>1611</td>
<td>228</td>
<td>1839</td>
</tr>
<tr>
<td>Conformance</td>
<td>590</td>
<td>91</td>
<td>681</td>
</tr>
<tr>
<td>Completeness</td>
<td>329</td>
<td>57</td>
<td>386</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2530</td>
<td>376</td>
<td>2906</td>
</tr>
</tbody>
</table>

**R Installation**

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

**Getting Started**

To install the latest stable version, install from CRAN:

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

To install the latest development version, install from GitHub:

```r
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)
## Exploring Dashboard

### Filtering Options

- **Show** [5] entries

### Select Columns to Show

- **Status**
- **Context**
- **Category**
- **Subcategory**
- **Level**
- **Description**
- **% Records**

### Export to CSV

- [Column visibility]
- [CSV]

### Table

<table>
<thead>
<tr>
<th>Status</th>
<th>Context</th>
<th>Category</th>
<th>Subcategory</th>
<th>Level</th>
<th>Description</th>
<th>% Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>PASS</td>
<td>Validation</td>
<td>Completeness</td>
<td>None</td>
<td>TABLE</td>
<td>The number and percent of persons in the CDM that do not have at least one record in the NOTE table (Threshold=100%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>Validation</td>
<td>Completeness</td>
<td>None</td>
<td>TABLE</td>
<td>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%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>Verification</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>The number and percent of records with a NULL value in the month_of_birth of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>Verification</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>The number and percent of records with a NULL value in the day_of_birth of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>Verification</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>The number and percent of records with a NULL value in the birth_datetime of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 3,312 entries

Over 3,000 Checks
Explaining Results

**Status**
- Pass
- Fail

**Category**
- Completeness
- Conformance
- Plausibility

**Context**
- Verification
- Validation

**Level**
- Concept
- Table
- Field

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

Description

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

Screenshot
SQL

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

- Open-sourced application
- Shiny is an R package
- Interactive web applications
- Enables easy sharing of aggregated results and visualizations
# OMOP data science tool matrix

*When to use what tool?*

<table>
<thead>
<tr>
<th></th>
<th>ATHENA</th>
<th>SQL</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary look-up</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>Database counts</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Population counts</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Characterisations</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Incidence</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Prevalence</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Treatment patterns &amp; pathways</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Patient-level predictions</td>
<td>▲</td>
<td></td>
<td>▲</td>
</tr>
<tr>
<td>Population-level estimation</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data visualisations e.g. sunburst plots</td>
<td>▲</td>
<td></td>
<td>▲</td>
</tr>
</tbody>
</table>
### 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.
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

https://data.ohdsi.org/DoacCohortDiagnostics/
https://ohdsi.github.io/Hades/
### Analytical packages

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

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohort Method</strong></td>
<td>New-user cohort studies using large-scale regression for propensity and outcome models</td>
</tr>
<tr>
<td><strong>Self-Controlled Case Series</strong></td>
<td>Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality.</td>
</tr>
<tr>
<td><strong>Self-Controlled Cohort</strong></td>
<td>A self-controlled cohort design, where time preceding exposure is used as control.</td>
</tr>
<tr>
<td><strong>Patient Level Prediction</strong></td>
<td>Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms.</td>
</tr>
<tr>
<td><strong>Case-control</strong></td>
<td>Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort.</td>
</tr>
<tr>
<td><strong>Case-crossover</strong></td>
<td>Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control).</td>
</tr>
<tr>
<td><strong>Empirical Calibration</strong></td>
<td>Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design.</td>
</tr>
<tr>
<td><strong>Method Evaluation</strong></td>
<td>Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods.</td>
</tr>
<tr>
<td><strong>Evidence Synthesis</strong></td>
<td>Combining study diagnostics and results across multiple sites.</td>
</tr>
<tr>
<td><strong>Database Connector</strong></td>
<td>Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.</td>
</tr>
<tr>
<td><strong>Sql Render</strong></td>
<td>Generate SQL on the fly for the various SQL dialects.</td>
</tr>
<tr>
<td><strong>Cyclops</strong></td>
<td>Highly efficient implementation of regularized logistic, Poisson and Cox regression.</td>
</tr>
<tr>
<td><strong>ParallellLogger</strong></td>
<td>Support for parallel computation with logging to console, disk, or e-mail.</td>
</tr>
<tr>
<td><strong>Feature Extraction</strong></td>
<td>Automatically extract large sets of features for user-specified cohorts using data in the CDM.</td>
</tr>
</tbody>
</table>

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

Standardized outputs for easy interpretation
OMOP research examples
OHDSI Research Program for COVID-19

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

*Virtual Study-a-thon, 26-29 March 2020*
Large-Scale Evidence Generation and Evaluation Across a Network of Databases (LEGEND)

"This study is turning me away from ACE inhibitors as a first line agent for hypertension. There are many other inexpensive options, including thiazide diuretics, and so, until more compelling information becomes available, there is little reason not to change practice."

- Harlan Krumholz, MD, SM

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

Summary

Background

Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines indicating no clear evidence to suggest a preferred class among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers.

Study code: http://www.github.com/ohdsi/LEGEND
Validation through EMA - Consistency between Source and CDM data

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, Karin Hedenmalm, Jim Slattery, Alison Cave, Xavier Kurz, Peter Arlett

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 diseases

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