The Journey from Data to ...
The Journey from Data to ... Data
The Journey from Data to Reliable Evidence

Patient-level data in source system

Patient-level data in Common data model

Reliable evidence
The Journey from Source to Standardized Data

Patient-level data in source system

Patient-level data in Common data model

Reliable evidence
The Journey from Source to Standardized Data

Technical Considerations for Setup
Data Governance
White Rabbit
Rabbit In A Hat
OMOP Vocabulary
Data Quality Dashboard
ACHILLES
Putting it all Together

Patient-level data in source system

Patient-level data in Common data model
The Journey from Source to Standardized Data

- Technical Considerations for Setup
- Data Governance
- White Rabbit
- Rabbit In A Hat
- OMOP Vocabulary
- Data Quality Dashboard
- ACHILLES
- Putting it all Together

Patient-level data in source system

Patient-level data in Common data model
The Journey from Source to Standardized Data

- Vocabulary Mapping & Usagi
- Data Quality
- ACHILLES
- Putting it all Together

Melanie Philofsky  Clair Blacketer  Anthony Molinaro  Frank DeFalco
Vocabulary Mapping and USAGI

Melanie Philofsky

CDM Workshop 2022

2022-03-15
OMOP Vocabulary

• Who, what, when, why & where?
• How does it all work?
• Custom semantic mapping
  • Usagi
  • Source to Concept Map table
  • 2 Billionaires
• Maintenance
OMOP Vocabulary tables

- Where they came from
- Who does all this work
- What they are
  - Quick review of source code to concept_id mapping
- How do I get these
- When & why
Where and who

- Public domain
- Proprietary sources
- Homegrown OMOP concepts

- OHDSI Vocabulary team
- Vocabulary maintenance
  - Source release cycles
  - Mapping to standards
- OHDSI Vocabulary GitHub page
When & why do you need the OMOP Vocabularies

• Now’s a good time to download them
• Can’t properly populate a CDM without them
• Can’t use OHDSI tools without them
Athena.ohdsi.org

Search

aspirin

1. Usage of quotation marks forces an exact-match search
2. In case of a typo, or if there is a similar spelling of the word, the most similar result will be presented

Explore domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drugs</td>
<td>5,250,974</td>
</tr>
<tr>
<td>Conditions</td>
<td>698,822</td>
</tr>
<tr>
<td>Procedures</td>
<td>733,500</td>
</tr>
<tr>
<td>Devices</td>
<td>491,758</td>
</tr>
<tr>
<td>Observations</td>
<td>567,112</td>
</tr>
<tr>
<td>Measurements</td>
<td>368,021</td>
</tr>
</tbody>
</table>
OMOP Standardized Vocabulary Tables

- Concept
- Concept Relationship
- Concept Ancestor
- Source to Concept Map

- Synonym
- Relationship
- Concept Class
- Domain
- Drug Strength
- Vocabulary
Structure of OMOP Vocabulary

All content: concepts in `concept`  

Direct relationships between concepts in `concept_relationship`  

Multi-step hierarchical relationships pre-processed into `concept_ancestor`
Different Categories of Concepts

**Non-standard Concepts**
- **Function**: Unique representation of a source code

**Standard Concepts**
- **Function**: Used for standardized analytics and by OHDSI tools

**Classification Concepts**
- **Function**: Do not directly represent source data
  Used to perform hierarchical queries
The Source for Source Codes

1. May come from international terminology or code system
   - SNOMED, ISBT

2. May come from a country specific terminology or code system
   - Read, BDPM, ICD10CN, CVX

3. May be free text strings
   - Centimeter, Intravenous, Cigarette Smoker

4. May come from a source specific code system
   - EHRs, CRFs, Registries, etc.
## Mapping Source Codes to Concept_IDs

<table>
<thead>
<tr>
<th>Concept_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map to standard or non-standard concept_id</td>
</tr>
<tr>
<td>Do not map to classification concept_id*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Code Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary is known: use to retrieve concept_id</td>
</tr>
<tr>
<td>Vocabulary is unknown, but domain is known: use domain to retrieve concept_id</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Custom/Local Source Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom or local source code: custom map to standard concept_id</td>
</tr>
</tbody>
</table>

*Classification concept_ids are used outside the clinical event tables for research*
Source Code Mapping – Scenario 1

Scenario
• Source code comes from an OMOP supported Vocabulary

Solution
• Use the following condition to perform the mapping:
  Where <source code> = CONCEPT.concept_code and <source vocabulary> = CONCEPT.vocabulary_id

<table>
<thead>
<tr>
<th>Source code</th>
<th>Source vocabulary</th>
<th>Code description</th>
<th>CONCEPT.concept_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>61462000</td>
<td>SNOMED</td>
<td>Malaria</td>
<td>438067</td>
</tr>
<tr>
<td>A663D00</td>
<td>Read</td>
<td>Zika Fever</td>
<td>45489770</td>
</tr>
<tr>
<td>A92.3</td>
<td>ICD10CN</td>
<td>West Nile Virus Infection</td>
<td>1404276</td>
</tr>
</tbody>
</table>
### Source Codes Mapping – Scenario 2

**Scenario**
- Source code is a text string

**Solution**
- Use using the following condition to perform the mapping:
  
  Where `<source string> = CONCEPT.concept_name` and `<source domain> = CONCEPT.domain_id`

<table>
<thead>
<tr>
<th>Source string</th>
<th>Source domain</th>
<th>Source table/field</th>
<th>CONCEPT.concept_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centimeter</td>
<td>Unit</td>
<td>Vital Signs/ unit for height measurement</td>
<td>8582</td>
</tr>
<tr>
<td>Intravenous</td>
<td>Route</td>
<td>Drug/ route for drug administration</td>
<td>4171047</td>
</tr>
<tr>
<td>Male</td>
<td>Gender</td>
<td>Demographics</td>
<td>8507</td>
</tr>
</tbody>
</table>
Source Codes Mapping – Scenario 3

Scenario

• Source data does not map to an OHDSI supported vocabulary

Solution

• Ask OHDSI
• Create custom mapping using one of the following two methods:
  • Source to Concept Map
  • 2 Billionaires
Custom Semantic Mapping

• Usagi
• Source to Concept Map table
• Concept & Concept Relationship tables
• When do you NOT create a custom concept_id?
# Usagi

## Source Code Mapping

<table>
<thead>
<tr>
<th>Status</th>
<th>Code</th>
<th>Source</th>
<th>Source term</th>
<th>Frequency</th>
<th>Match score</th>
<th>Concept ID</th>
<th>Concept name</th>
<th>Domain</th>
<th>Concept class</th>
<th>Vocabulary</th>
<th>Concept code</th>
<th>Standard concept</th>
<th>Parents</th>
<th>Children</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approved</td>
<td>KSH.00</td>
<td>Hypertension</td>
<td>691496</td>
<td>0.81</td>
<td>316866</td>
<td>Hypertension</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>38341003</td>
<td>A</td>
<td>1</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approved</td>
<td>D01.00</td>
<td>Abdominal pain</td>
<td>178784</td>
<td>0.96</td>
<td>197390</td>
<td>Generalized pain</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>102514009</td>
<td>A</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unchecked</td>
<td>S99.00</td>
<td>Skin disease, other</td>
<td>875817</td>
<td>0.75</td>
<td>4317258</td>
<td>Disorder of skin</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>95320006</td>
<td>S</td>
<td>2</td>
<td>193</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Search

- **Query:**
  - Use source term as query
  - Query: 

## Results

<table>
<thead>
<tr>
<th>Score</th>
<th>Term</th>
<th>Concept ID</th>
<th>Concept name</th>
<th>Domain</th>
<th>Concept class</th>
<th>Vocabulary</th>
<th>Concept code</th>
<th>Standard concept</th>
<th>Parents</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>Skin disease</td>
<td>4317258</td>
<td>Disorder of skin</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>95320006</td>
<td>S</td>
<td>2</td>
<td>193</td>
</tr>
<tr>
<td>0.65</td>
<td>Skin Disease, Fungal</td>
<td>1327213</td>
<td>Dermal mycosis</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>14560005</td>
<td>S</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>0.67</td>
<td>AIDS with skin lesions</td>
<td>2224556</td>
<td>Skin disorder associated with AIDS</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>42384009</td>
<td>S</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.66</td>
<td>Chronic skin disease</td>
<td>4134332</td>
<td>Chronic disease of skin</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>128326002</td>
<td>S</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>0.55</td>
<td>Disease, Ophthalmic</td>
<td>3781651</td>
<td>Disorder of ear</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>25960001</td>
<td>S</td>
<td>4</td>
<td>43</td>
</tr>
<tr>
<td>0.55</td>
<td>Disease, Ocular</td>
<td>4163346</td>
<td>Glaucoma, open-angle</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>29291001</td>
<td>S</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0.55</td>
<td>Other peripheral vascular disease</td>
<td>321052</td>
<td>Peripheral vascular disease</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>400947009</td>
<td>S</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>0.52</td>
<td>Disease, Osteoarthropathy</td>
<td>4159512</td>
<td>Lower limb amputation</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>22381000</td>
<td>S</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0.52</td>
<td>Disease, Osteoarthropathy</td>
<td>4159512</td>
<td>Lower limb amputation</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
<td>22381000</td>
<td>S</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

## Comments

- **Comment:**
  - 5/1637 | 8.8% of total frequency

---

**Replace concept** | **Add concept**

- **Approve**
**Source_to_Concept_Map**

Example source code = Pediatric interventional cardiologist

<table>
<thead>
<tr>
<th>Field</th>
<th>Source_code</th>
<th>Source_concept_id</th>
<th>Source_vocabulary_id</th>
<th>Source_code_description</th>
<th>Target_concept_id</th>
<th>Target_vocabulary_id</th>
<th>Valid_start_date</th>
<th>Valid_end_date</th>
<th>Invalid_reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required Field?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Value</td>
<td>Pediatric interventional cardiologist</td>
<td>Unique identifier &gt;2 billion</td>
<td>prov_specialty</td>
<td>Pediatric interventional cardiologist</td>
<td>903276</td>
<td>Medicare Specialty</td>
<td>01/01/1970</td>
<td>12/31/2099</td>
<td></td>
</tr>
</tbody>
</table>
2 Billionaires

• Create a Concept
• Create the Concept Relationship
## Concept

**Example source code = Pediatric interventional cardiologist**

<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>
<th>Valid_start_date</th>
<th>Valid_end_date</th>
<th>Invalid_reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Unique id &gt; 2 billion 21000000</td>
<td>Pediatric interventional cardiologist</td>
<td>Provider</td>
<td>prov_specialty</td>
<td>Physician Specialty</td>
<td>Pediatric interventional cardiologist</td>
<td>01/01/1970</td>
<td>12/31/2099</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
## Concept Relationship

Example source code = Pediatric interventional cardiologist

<table>
<thead>
<tr>
<th>Field</th>
<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>Required Field?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Record #1 value</td>
<td>21000000000</td>
<td>903276</td>
<td>Maps to</td>
<td>01/01/1970</td>
<td>12/31/2099</td>
<td></td>
</tr>
<tr>
<td>Record #2 value</td>
<td>903276</td>
<td>21000000000</td>
<td>Mapped from</td>
<td>01/01/1970</td>
<td>12/31/2099</td>
<td></td>
</tr>
</tbody>
</table>
Comparison of both methods

**STCM**
- Designed for insert statements
- Need to create an additional SQL statement to do a lookup on the STCM during the ETL

**Concept & Concept Relationship**
- Need to modify OHDSI provided tables
  - Concept, Concept Relationship, Vocabulary tables
- Enriches your Vocabulary
- **Enables the display of source concepts in Atlas**
- Great for internal use cases
When do you NOT create a custom concept?

- New source code, already supported vocabulary
- New vocabulary, strong use case
Maintenance

• Every time you update your OHDSI vocabularies, re-run your ETL

• Complete an analysis of your CDM
  – Review top unmapped values
  – Review for duplicates
    • Are source codes now represented by an OHDSI supported concept_id?
  – Update STCM or Concept & Concept Relationship tables
Resources

- Athena - This is a web browser for the most up to date vocabularies
  - http://Athena.ohdsi.org
- USAGI: Download the program, request enhancements, raise issues
  - https://github.com/OHDSI/Usagi
- OHDSI/OMOP Vocabulary GitHub: Log issues/defects & request enhancements
  - https://github.com/OHDSI/Vocabulary-v5.0
- Forums: Ask questions, open discussions, raise ideas
  - https://forums.ohdsi.org
- CDM GitHub page: Log issues/defects & request enhancements
  - https://github.com/OHDSI/CommonDataModel
- CDM wiki page: All conventions,
  - https://ohdsi.github.io/CommonDataModel/faq.html
- Book of OHDSI: Central repository for OHDSI knowledge
  - https://ohdsi.github.io/TheBookOfOhdsi/
Data Quality Dashboard

Clair Blacketer
CDM Workshop 2022
2022-03-15
Data experts and CDM experts together design the ETL.

People with medical knowledge create the code mappings.

All are involved in quality control.

A technical person implements the ETL.

**OHDSI Tools**
- White Rabbit
- Rabbit In a Hat
- Usagi
- White Rabbit
- ACHILLES
- DQD
- Rabbit In a Hat
Investigators should fully understand the quality assurance (QA) and quality control (QC) procedures used by the data holders and how these procedures could have an effect on the integrity of the data and the overall validity of the study. FDA recommends that investigators address the following topics:

- The frequency and type of any data error corrections or changes in data adjudication policies implemented by the data holders during the relevant period of data collection;
- A description of any peer-reviewed publications examining data quality and/or validity, including the relationships of the investigators with the data source(s);
- Any updates and changes in coding practices (e.g., ICD codes) across the study period that are relevant to the outcomes of interest;
- Any changes in key data elements during the study time frame and their potential effect on the study; and
- A report on the extent of missing data over time (i.e., the percentage of data not available for a particular variable of interest) and a discussion on the procedures (e.g., exclusion, imputation) employed to handle this issue. Investigators should also address the implications of the extent of missing data on study findings and the missing data methods used.
In order to include novel data sources as evidence sources for regulatory decision-making, it is critical to understand how much the regulators can rely on the data. Thus, a capability to characterise the quality of data is a strategic objective for regulators. **While pre-defining quality is challenging as need is often driven by the question, it is possible to define some generalised elements for which quality could be defined.**

**What this means for stakeholders:**

A data quality framework will support the trust of patients and healthcare professionals in the decisions reached by regulators when Big Data underpins those decisions. It will aid the choice of data source selected for a study (including those by industry) and it will inform the assessment of the study results and the benefit-risk dossier by regulators.
What is Data Quality?

We need to assess the quality of the conversion AND the quality of the data.
## Data Quality Check Types

<table>
<thead>
<tr>
<th>Check Type</th>
<th>Check Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Completeness</td>
<td>The number and percent of persons in a database that do not have a least one record in the <em>CDM table</em>.</td>
</tr>
<tr>
<td>Is Required</td>
<td>The number and percent of records with a NULL value in a <em>CDM field</em> of a <em>CDM table</em> that is considered not nullable.</td>
</tr>
<tr>
<td>Is Primary Key</td>
<td>The number and percent of records that have a duplicate value in the <em>CDM field</em> of the <em>CDM table</em>.</td>
</tr>
<tr>
<td>Is Foreign Key</td>
<td>The number and percent of records that have a value in a <em>lookup field</em> of a <em>CDM table</em> that does not exist in the <em>lookup table</em>.</td>
</tr>
<tr>
<td>Concept Domain</td>
<td>The number and percent of records that have a concept_id that does not conform to the <em>domain</em>.</td>
</tr>
<tr>
<td>Is Standard Valid Concept</td>
<td>The number and percent of records that do not have a standard, valid concept in the <em>CDM field</em> of a <em>CDM table</em>.</td>
</tr>
</tbody>
</table>
# Data Quality Check Types

<table>
<thead>
<tr>
<th>Check Type</th>
<th>Check Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Concept Completeness</td>
<td>The number and percent of records with a value of 0 in the standard concept field <em>CDM field</em> in the <em>CDM table</em>.</td>
</tr>
<tr>
<td>Plausible Temporal After</td>
<td>The number and percent of records with a value in a <em>CDM field</em> of a <em>CDM table</em> that occurs prior to a <em>plausible date</em>.</td>
</tr>
<tr>
<td>Plausible Value Low</td>
<td>For a given <em>concept_id</em> and <em>unit_concept_id</em> pair, the number and percent of records with a value lower than the <em>plausible low value</em>.</td>
</tr>
<tr>
<td>Plausible Gender</td>
<td>For a given <em>concept_id</em>, the number and percent of records associated with persons with an <em>implausible gender</em>.</td>
</tr>
</tbody>
</table>
# Data Quality Check Types

<table>
<thead>
<tr>
<th></th>
<th>Verification</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plausibility</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Conformance</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Completeness</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

## 20 Check Types
### Data Quality Check Totals

<table>
<thead>
<tr>
<th></th>
<th>Verification</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plausibility</td>
<td>1855</td>
<td>287</td>
</tr>
<tr>
<td>Conformance</td>
<td>563</td>
<td>80</td>
</tr>
<tr>
<td>Completeness</td>
<td>327</td>
<td>12</td>
</tr>
</tbody>
</table>

Total 3,124 Checks
Data Quality Dashboard

Results generated at 2019-09-06 22:20:12 in 7 hours

**RESULTS**

**IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE**

<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>Verification</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>The number and percent of records with a NULL value in the range_high of the MEASUREMENT. (Threshold=100%)</td>
<td>82.14%</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 visit_start_of the MEASUREMENT. (Threshold=100%)</td>
<td>80.90%</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 value_source_value of the MEASUREMENT. (Threshold=100%)</td>
<td>79.99%</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 CM that do not have at least one record in the DEVICE_EXPOSURE table. (Threshold=100%)</td>
<td>76.70%</td>
</tr>
<tr>
<td>FAIL</td>
<td>Verification</td>
<td>Plausibility</td>
<td>Atemporal</td>
<td>CONCEPT</td>
<td>For the combination of CONCEPT_ID 2016649 (Testosterone Free [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8865 (picograms per milliliter), the number and percent of records that have a value less than 5.00g/ml. (Threshold=1%)</td>
<td>72.43%</td>
</tr>
</tbody>
</table>

Showing 126 to 130 of 3,361 entries

https://github.com/OHDSI/DataQualityDashboard
DQD v1.4 released with full support for CDM v5.4

https://github.com/OHDSI/DataQualityDashboard
# Data Quality Dashboard

## Data Quality Assessment

**IBM® MarketScan® Multi-State Medicaid Database**

Results generated at 2020-08-24 15:44:34 in 3 hours

<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>1849</td>
<td>6</td>
<td>1855</td>
</tr>
<tr>
<td>Conformance</td>
<td>550</td>
<td>13</td>
<td>563</td>
</tr>
<tr>
<td>Completeness</td>
<td>322</td>
<td>5</td>
<td>327</td>
</tr>
<tr>
<td>Total</td>
<td>2721</td>
<td>24</td>
<td>2745</td>
</tr>
</tbody>
</table>

[https://data.ohdsi.org/DataQualityDashboardMDCD/](https://data.ohdsi.org/DataQualityDashboardMDCD/)
Increasing trust in real-world evidence through evaluation of observational data quality

Clair Blacketer, Frank J Defalco, Patrick B Ryan, Peter R Rijnbeek

Abstract

Objective: Advances in standardization of observational healthcare data have enabled methodological breakthroughs, rapid global collaboration, and generation of real-world evidence to improve patient outcomes. Standardizations in data structure, such as use of common data models, need to be coupled with standardized approaches for data quality assessment. To ensure confidence in real-world evidence generated from the analysis of real-world data, one must first have confidence in the data itself.

Materials and methods: We describe the implementation of check types across a data quality framework of conformance, completeness, plausibility, with both verification and validation. We illustrate how data quality checks, paired with decision thresholds, can be configured to customize data quality reporting across a range of observational health data sources. We discuss how data quality assessment and management of the standardized and harmonized data sources provide a reliable basis for the generation of evidence.
ACHILLES

Anthony Molinaro
CDM Workshop 2022
2022-03-15
ETL Process

Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

All are involved in quality control

A technical person implements the ETL

OHDSI Tools

White Rabbit
Rabbit In a Hat
Usagi
White Rabbit
ACHILLES
DQD
Rabbit In a Hat
Achilles

Achilles is a data characterization and quality tool available for download here:

https://github.com/OHDSI/Achilles

To visualize the results, a new tool was developed called ARES

https://github.com/OHDSI/Ares
PERSON REPORT

160,465,311
Number of People

Male: 78,667,927 (49%)  Female: 81,997,384 (51%)
Gender Proportions

Population by Age & Sex

Population by Year of Birth
### VISIT OCCURRENCE

<table>
<thead>
<tr>
<th>Concept Id</th>
<th>Concept Name</th>
<th>% People</th>
</tr>
</thead>
<tbody>
<tr>
<td>9202</td>
<td>Outpatient Visit</td>
<td>80.64%</td>
</tr>
<tr>
<td>581458</td>
<td>Pharmacy visit</td>
<td>63.08%</td>
</tr>
<tr>
<td>9203</td>
<td>Emergency Room Visit</td>
<td>23.88%</td>
</tr>
<tr>
<td>32036</td>
<td>Laboratory Visit</td>
<td>16.86%</td>
</tr>
<tr>
<td>9201</td>
<td>Inpatient Visit</td>
<td>13.06%</td>
</tr>
</tbody>
</table>
COVID-19

**Concept ID:** 37311061

**Number of People:** 1,585,889

**% of People:** 1%

**Records per Person:** 2.3

---

**Age at First Diagnosis**

The chart shows the distribution of age at first diagnosis for COVID-19 patients, differentiated by gender (female and male).

**Female**
- Min Value: 0
- P10 Value: 15
- P25 Value: 24
- Median Value: 39
- P75 Value: 52
- P90 Value: 59
- Max Value: 65

**Male**
- Min Value: 0
- P10 Value: 15
- P25 Value: 24
- Median Value: 39
- P75 Value: 52
- P90 Value: 59
- Max Value: 65