CDM Workshop, Part 2

Common Data Model Working Group

2022-03-15

The Journey from Data to …

The Journey from Data to …

Data

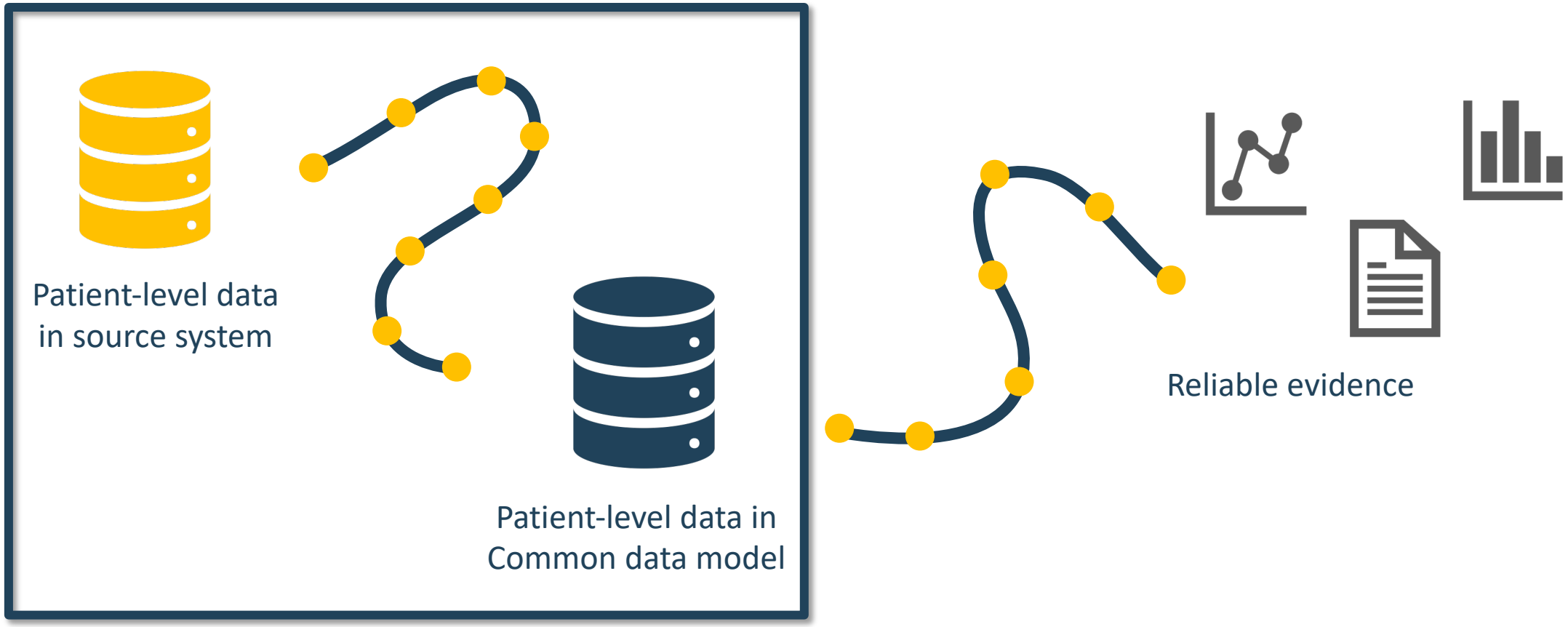
The Journey from Data to Reliable Evidence

Patient-level data in source system

Patient-level data in Common data model

Reliable evidence

The Journey from Source to Standardized Data



The Journey from Source to Standardized Data

The Journey from Source to Standardized Data

The Journey from Source to Standardized Data

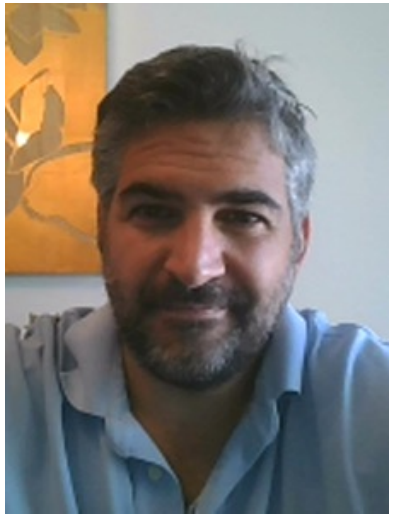
Vocabulary Mapping & Usagi

Melanie Philofsky

Data Quality

Clair Blacketer

ACHILLES



Anthony Molinaro

Putting it all Together

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

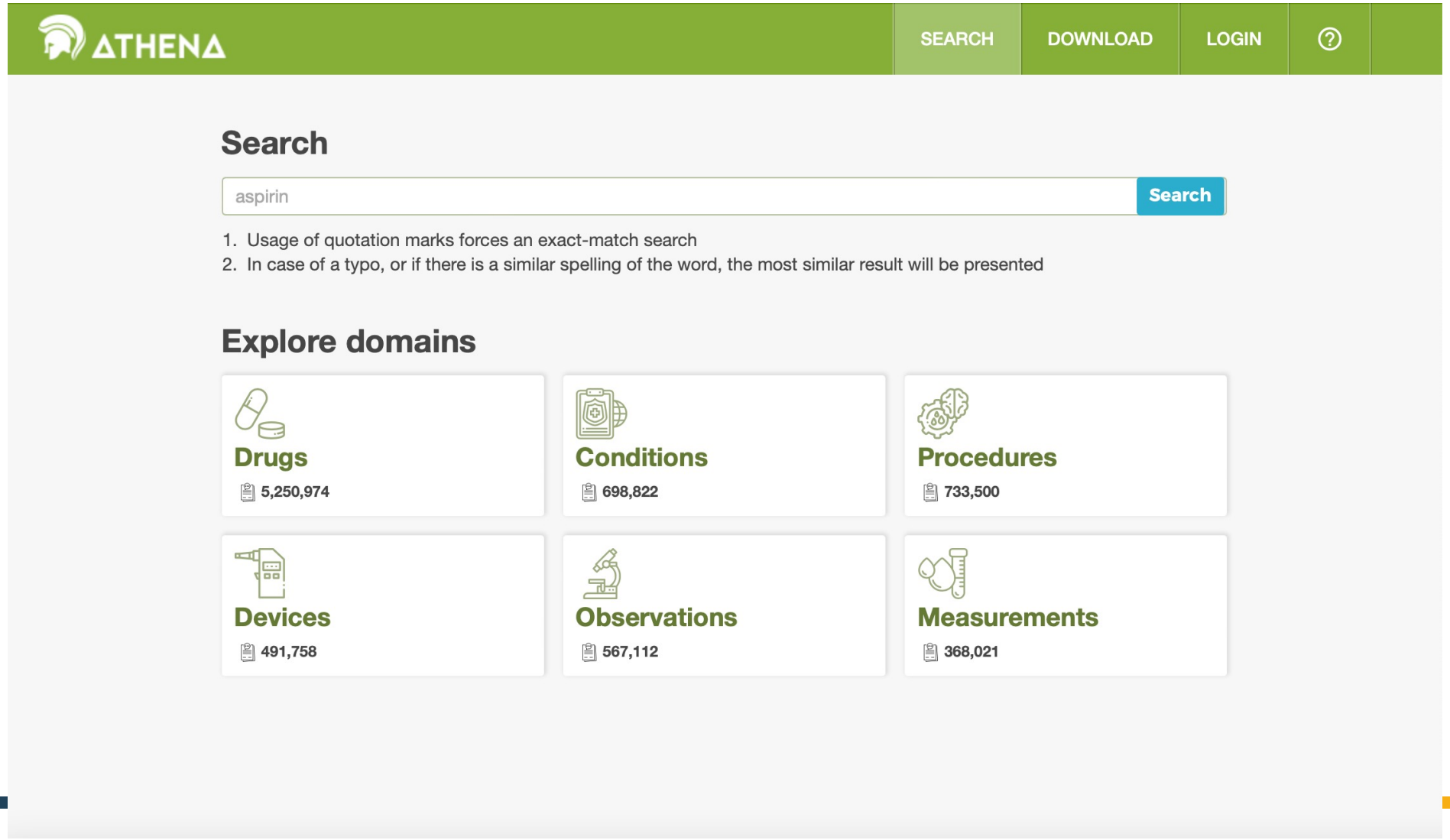
- Public domain
- Proprietary sources
- Homegrown OMOP concepts

- OHDSI Vocabulary team
- Vocabulary maintenance
  - Source release cycles
  - Mapping to standards
- OHDSI Vocabulary GitHub page

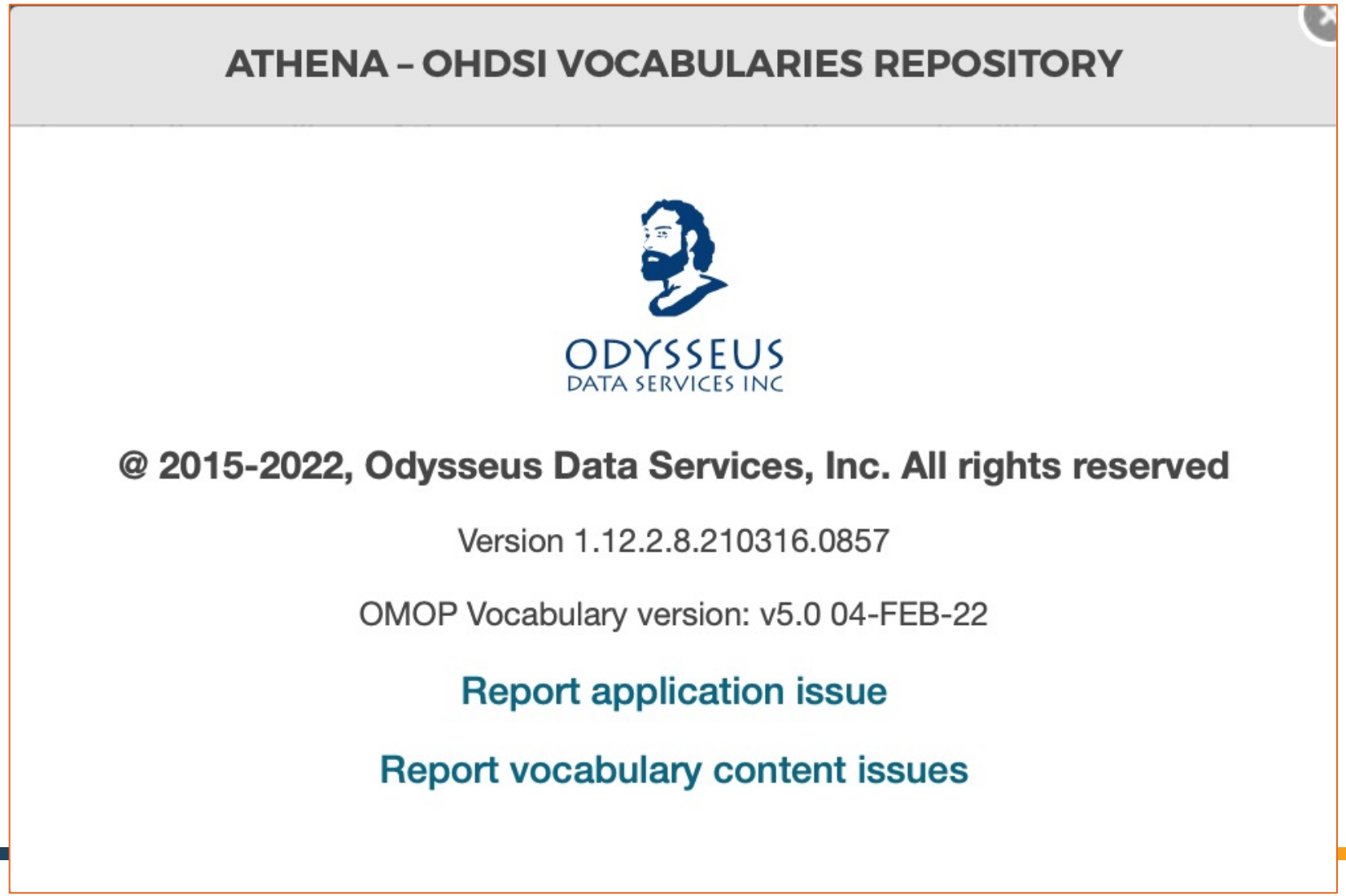
When & why do you need the OMOP Vocabularies

- 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



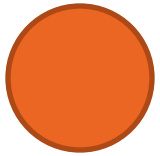
Vocabulary Version



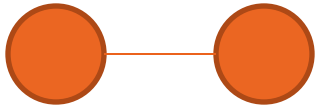
OMOP Standardized Vocabulary Tables

- Concept
- Concept Relationship
- Concept Ancestor
- Source to Concept Map
- Synonym
- Relationship
- Concept Class
- Domain
- Drug Strength
- Vocabulary

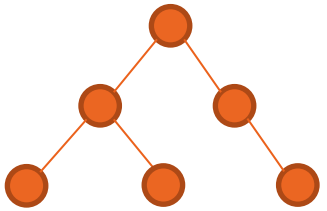
Structure of OMOP Vocabulary



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

Concept_ID

- **Map to** standard or non-standard concept_id
- **Do not map to** classification concept_id*

Source Code Vocabulary

- Vocabulary is **known**: use to retrieve concept_id
- Vocabulary is **unknown**, but ***domain is known***: use domain to retrieve concept_id

Custom/Local Source Code

- Custom or local source code: **custom map** to standard concept_id

*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 using the following condition to perform the mapping:

  Where <source code> = CONCEPT.concept_code and <source vocabulary> = CONCEPT.vocabulary_id

Source code	Source vocabulary	Code description	CONCEPT.concept_id
61462000	SNOMED	Malaria	438067
A663D00	Read	Zika Fever	45489770
A92.3	ICD10CN	West Nile Virus Infection	1404276

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

Source string	Source domain	Source table/field	CONCEPT.concept_id
Centimeter	Unit	Vital Signs/ unit for height measurement	8582
Intravenous	Route	Drug/ route for drug administration	4171047
Male	Gender	Demographics	8507

Source Codes Mapping – Scenario 3

Scenario

- Source data does not map to an OHDSI supported vocabulary

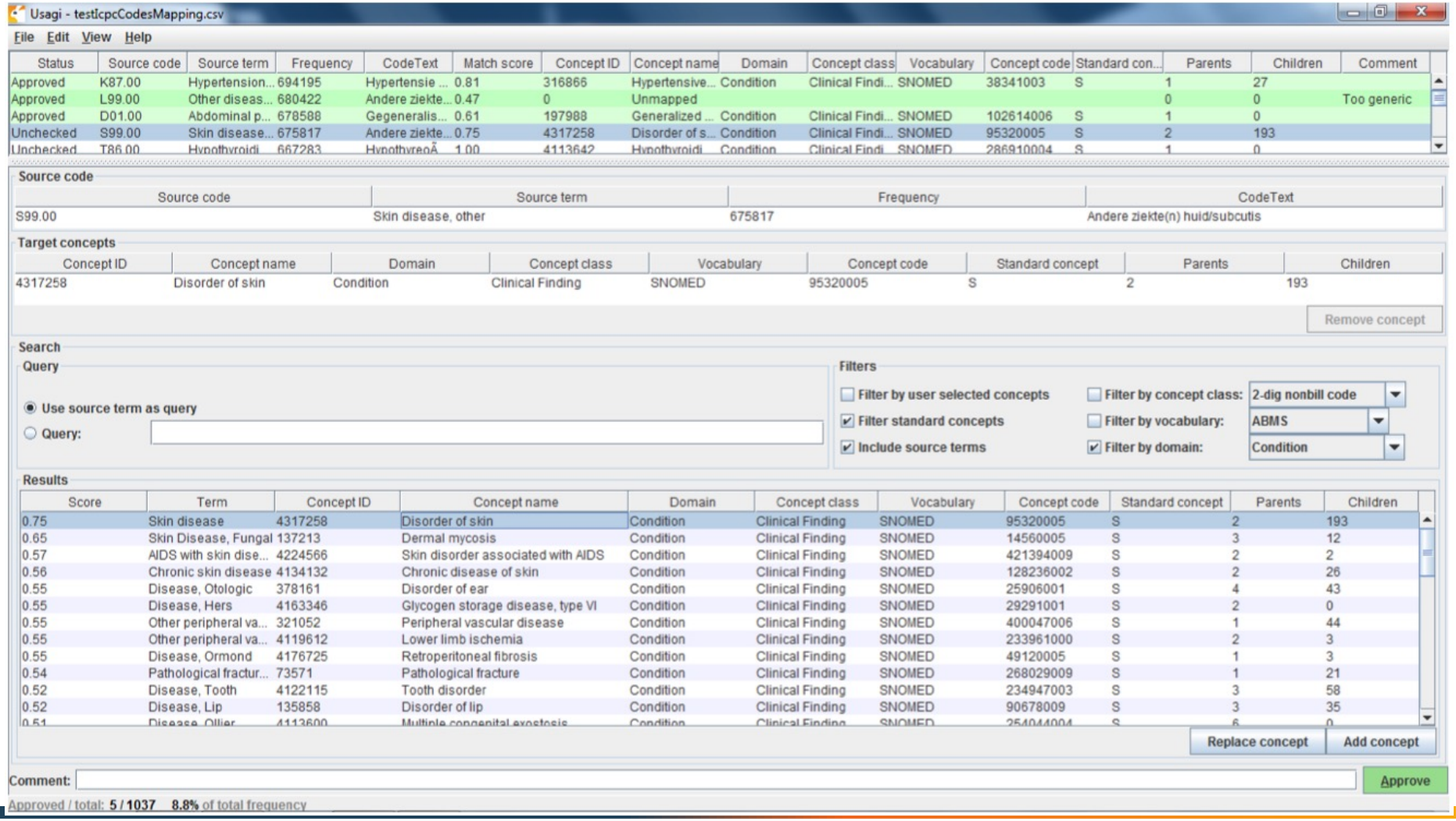
Solution

- 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


Usagi - testIcpcCodesMapping.csv
File Edit View Help
Status | Source code | Source term | Frequency | CodeText | Match score | Concept ID | Concept name | Domain | Concept class | Vocabulary | Concept code | Standard con... | Parents | Children | Comment
Approved | K87.00 | Hypertension... | 694195 | Hypertensie ... | 0.81 | 316866 | Hypertensive... | Condition | Clinical Findi... | SNOMED | 38341003 | S | 1 | 27 |
Approved | L99.00 | Other diseas... | 680422 | Andere ziekte... | 0.47 | 0 | Unmapped | | | | | | 0 | 0 | Too generic
Approved | D01.00 | Abdominal p... | 678588 | Gegeneralis... | 0.61 | 197988 | Generalized ... | Condition | Clinical Findi... | SNOMED | 102614006 | S | 1 | 0 |
Unchecked | S99.00 | Skin disease... | 675817 | Andere ziekte... | 0.75 | 4317258 | Disorder of s... | Condition | Clinical Findi... | SNOMED | 95320005 | S | 2 | 193 |
Source code
Source code | Source term | Frequency | CodeText
S99.00 | Skin disease, other | 675817 | Andere ziekte(n) huid/subcutis
Target concepts
Concept ID | Concept name | Domain | Concept class | Vocabulary | Concept code | Standard concept | Parents | Children
4317258 | Disorder of skin | Condition | Clinical Finding | SNOMED | 95320005 | S | 2 | 193
Remove concept
Search
Query
Use source term as query
Query:
Filters
Filter by user selected concepts
Filter by concept class: 2-dig nonbill code
Filter standard concepts
Filter by vocabulary: ABMS
Include source terms
Filter by domain: Condition
Results
Score | Term | Concept ID | Concept name | Domain | Concept class | Vocabulary | Concept code | Standard concept | Parents | Children
0.75 | Skin disease | 4317258 | Disorder of skin | Condition | Clinical Finding | SNOMED | 95320005 | S | 2 | 193
0.65 | Skin Disease, Fungal | 137213 | Dermal mycosis | Condition | Clinical Finding | SNOMED | 14560005 | S | 3 | 12
0.57 | AIDS with skin dise... | 4224566 | Skin disorder associated with AIDS | Condition | Clinical Finding | SNOMED | 421394009 | S | 2 | 2
0.56 | Chronic skin disease | 4134132 | Chronic disease of skin | Condition | Clinical Finding | SNOMED | 128236002 | S | 2 | 26
0.55 | Disease, Otologic | 378161 | Disorder of ear | Condition | Clinical Finding | SNOMED | 25906001 | S | 4 | 43
0.55 | Disease, Hers | 4163346 | Glycogen storage disease, type VI | Condition | Clinical Finding | SNOMED | 29291001 | S | 2 | 0
0.55 | Other peripheral va... | 321052 | Peripheral vascular disease | Condition | Clinical Finding | SNOMED | 400047006 | S | 1 | 44
0.55 | Other peripheral va... | 4119612 | Lower limb ischemia | Condition | Clinical Finding | SNOMED | 233961000 | S | 2 | 3
0.55 | Disease, Ormond | 4176725 | Retroperitoneal fibrosis | Condition | Clinical Finding | SNOMED | 49120005 | S | 1 | 3
0.54 | Pathological fractur... | 73571 | Pathological fracture | Condition | Clinical Finding | SNOMED | 268029009 | S | 1 | 21
0.52 | Disease, Tooth | 4122115 | Tooth disorder | Condition | Clinical Finding | SNOMED | 234947003 | S | 3 | 58
0.52 | Disease, Lip | 135858 | Disorder of lip | Condition | Clinical Finding | SNOMED | 90678009 | S | 3 | 35
Replace concept
Add concept
Comment:
Approve
Approved / total: 5 / 1037 8.8% of total frequency

Source_to_Concept_Map

Example source code = Pediatric interventional cardiologist

Field	Source_code	Source_concept_id	Source_vocabulary_id	Source_code_description	Target_concept_id	Target_vocabulary_id	Valid_start_date	Valid_end_date	Invalid_reason
Required Field?	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Value	Pediatric interventional cardiologist	Unique identifier >2 billion	prov_specialty	Pediatric interventional cardiologist	903276	Medicare Specialty	01/01/1970	12/31/2099	

2 Billionaires

- Create a Concept
- Create the Concept Relationship

Concept

Example source code = Pediatric interventional cardiologist

Concept_id	Concept_name	Domain_id	Vocabulary_id	Concept_class_id	Standard_concept	Concept_code	Valid_start_date	Valid_end_date	Invalid_reason
Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Unique id > 2 billion 21000000	Pediatric interventional cardiologist	Provider	prov_specialty	Physician Specialty		Pediatric interventional cardiologist	01/01/1970	12/31/2099	

Concept Relationship

Example source code = Pediatric interventional cardiologist

Field	Concept_id_1	Concept_id_2	Relationship_id	Valid_start_date	Valid_end_date	Invalid_reason
Required Field?	Yes	Yes	Yes	Yes	Yes	No
Record #1 value	2100000000	903276	Maps to	01/01/1970	12/31/2099	
Record #2 value	903276	2100000000	Mapped from	01/01/1970	12/31/2099	

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

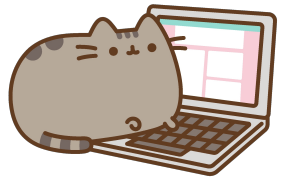
ETL Process

Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

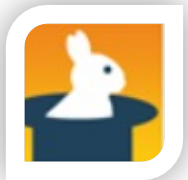
All are involved in quality control



A technical person implements the ETL

OHDSI Tools

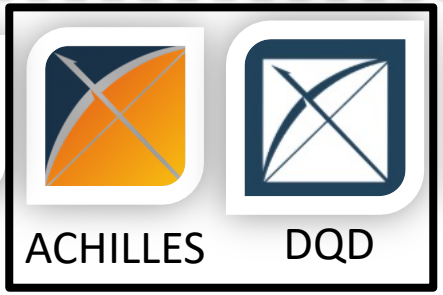
White Rabbit

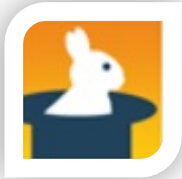


Rabbit In a Hat

Usagi

White Rabbit





Rabbit In a Hat

Data Quality – Thoughts From the FDA

Asses
Use i

Assessing Data Reliability (Data Accrual Data Assurance) and Relevance

The strength of RWE submitted in support of a regu
depends on the clinical study methodology and the r
accrual and data quality control (data assurance)) ar
underlying data. In general, FDA does not endorse o

E. Quality Assurance (QA) and Quality Control (QC)

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.

Data Quality – Thoughts From the EMA

Establish a certification process for data sources '

'Data quality is not a static construct and is context, disease and question dependent and dependent on the healthcare system. Assessments need to be constant and documented every time the data is refreshed'

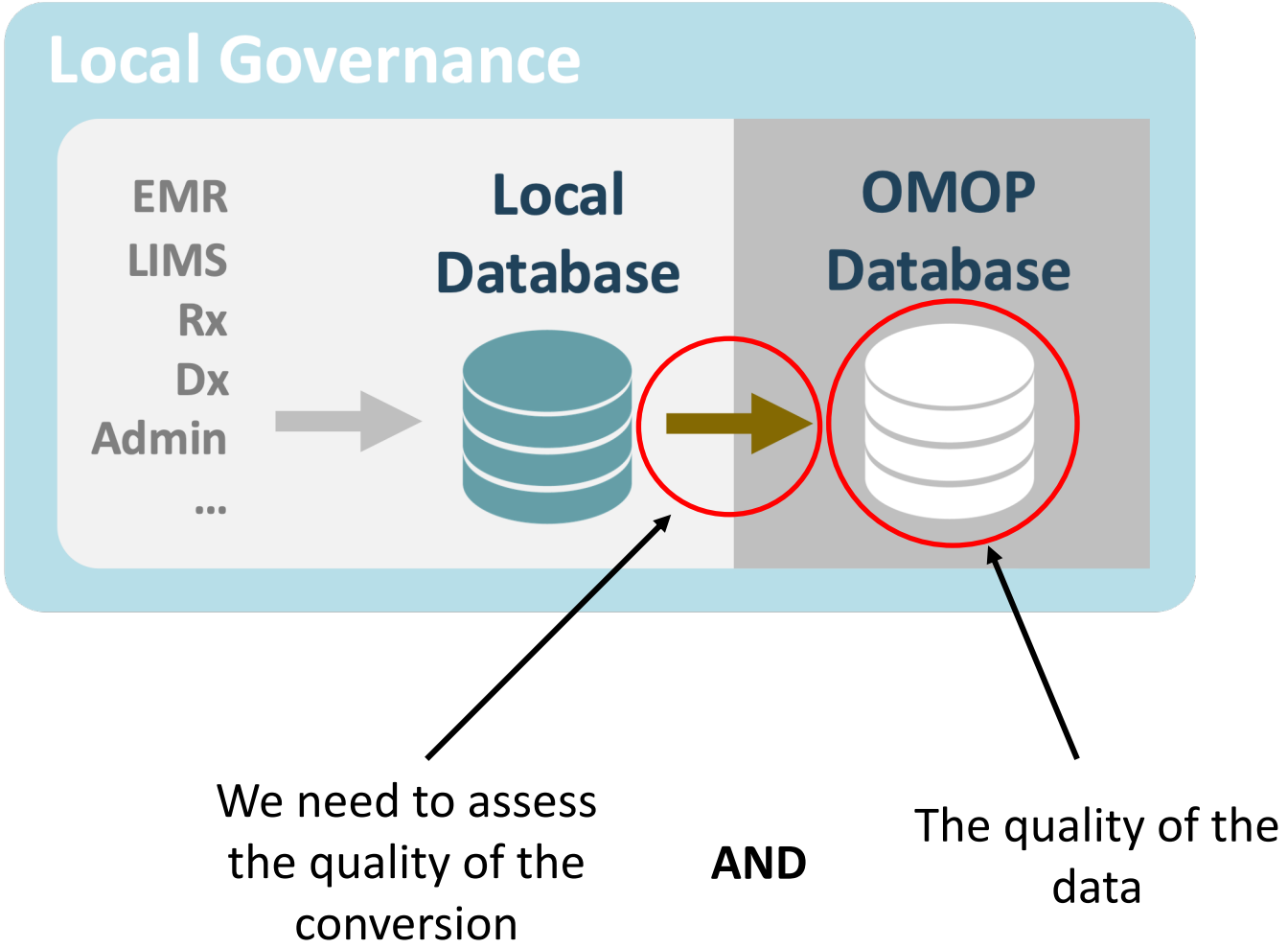
Public consultation comments

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?



Data Quality Check *Types*

Check Type	Check Description
Person Completeness	The number and percent of persons in a database that do not have a least one record in the *CDM table.*
Is Required	The number and percent of records with a NULL value in a *CDM field* of a *CDM table* that is considered not nullable.
Is Primary Key	The number and percent of records that have a duplicate value in the *CDM field* of the *CDM table.*
Is Foreign Key	The number and percent of records that have a value in a *lookup field* of a *CDM table* that does not exist in the *lookup table.*
Concept Domain	The number and percent of records that have a concept_id that does not conform to the *domain.*
Is Standard Valid Concept	The number and percent of records that do not have a standard, valid concept in the *CDM field* of a *CDM table.*

Data Quality Check *Types*

Check Type	Check Description
Standard Concept Completeness	The number and percent of records with a value of 0 in the standard concept field *CDM field* in the *CDM table*.
Plausible Temporal After	The number and percent of records with a value in a *CDM field* of a *CDM table* that occurs prior to a *plausible date.*
Plausible Value Low	For a given *concept_id* and *unit_concept_id* pair, the number and percent of records with a value lower than the *plausible low value.*
Plausible Gender	For a given *concept_id,* the number and percent of records associated with persons with an *implausible gender.*

Data Quality Check *Types*

	Verification	Validation
Plausibility	6	1
Conformance	7	1
Completeness	4	1

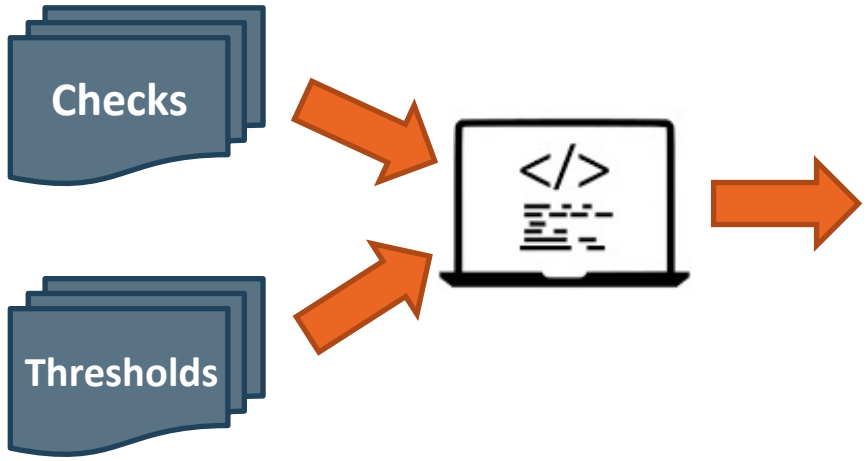
20 Check *Types*

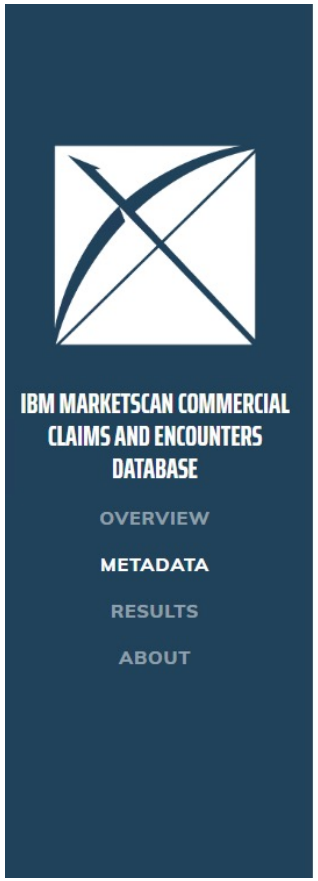
Data Quality Check *Totals*

	Verification	Validation
Plausibility	1855	287
Conformance	563	80
Completeness	327	12

Total 3,124 Checks

Data Quality Dashboard





RESULTS

IBM MARKETSCAN COMMERCIAL CLAIMS AND ENCOUNTERS DATABASE

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

Column visibility | CSV

Show 5 entries | Search:

STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the range_high of the MEASUREMENT. (Threshold=100%).	82.14%
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the visit_detail_id of the MEASUREMENT. (Threshold=100%).	80.90%
PASS	Verification	Completeness	None	FIELD	The number and percent of records with a NULL value in the value_source_value of the MEASUREMENT. (Threshold=100%).	79.89%
PASS	Validation	Completeness	None	TABLE	The number and percent of persons in the CDM that do not have at least one record in the DEVICE_EXPOSURE table (Threshold=100%).	76.70%
FAIL	Verification	Plausibility	Atemporal	CONCEPT	For the combination of CONCEPT_ID 3016049 (Testosterone Free [Mass/volume] in Serum or Plasma) and UNIT_CONCEPT_ID 8845 (picogram per milliliter), the number and percent of records that have a value less than 5.00e+00. (Threshold=1%).	72.43%

Showing 126 to 130 of 3,351 entries

Previous 1 ... 25 26 27 ... 671 Next

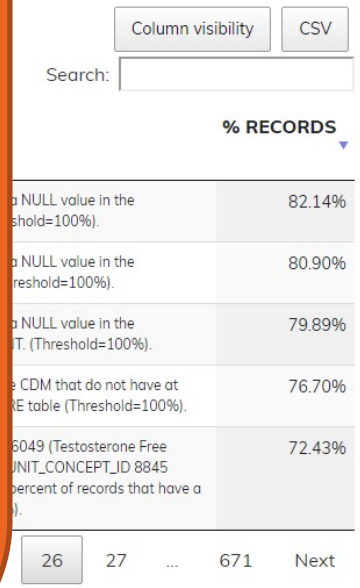
https://github.com/OHDSI/DataQualityDashboard

Data Quality Dashboard

Checks

Thresholds

DQD v1.4 released with full support for CDM v5.4



https://github.com/OHDSI/DataQualityDashboard

Data Quality Dashboard

IBM® MARKETSCAN® MULTI-STATE MEDICAID DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

IBM® MARKETSCAN® MULTI-STATE MEDICAID DATABASE

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

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1849	6	1855	100%	281	6	287	98%	2130	12	2142	99%
Conformance	550	13	563	98%	80	0	80	100%	630	13	643	98%
Completeness	322	5	327	98%	12	0	12	100%	334	5	339	99%
Total	2721	24	2745	99%	373	6	379	98%	3094	30	3124	**99%**

https://data.ohdsi.org/DataQualityDashboardMDCD/

Data Quality Dashboard

J Am Med Inform Assoc. 2021 Sep 18;28(10):2251-2257. doi: 10.1093/jamia/ocab132.

Increasing trust in real-world evidence through evaluation of observational data quality

Clair Blacketer [1 2], Frank J Defalco [1], Patrick B Ryan [1 3], Peter R Rijnbeek [2]

Affiliations + expand

PMID: 34313749 PMCID: PMC8449628 DOI: 10.1093/jamia/ocab132

Free PMC article

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

ACHILLES

Anthony Molinaro

CDM Workshop 2022

2022-03-15

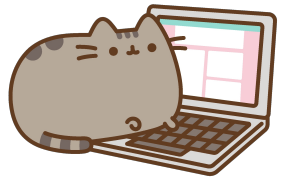
ETL Process

Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

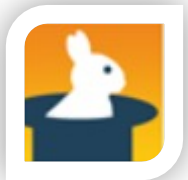
All are involved in quality control



A technical person implements the ETL

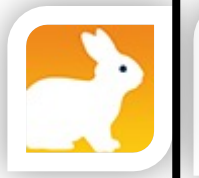
OHDSI Tools

White Rabbit

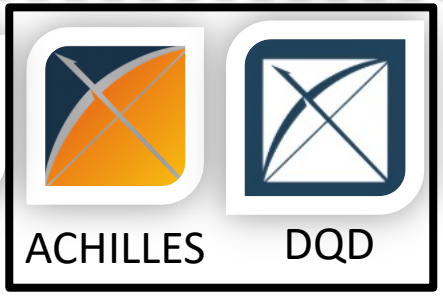


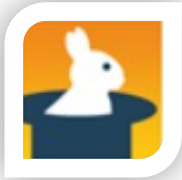
Rabbit In a Hat

Usagi



White Rabbit





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

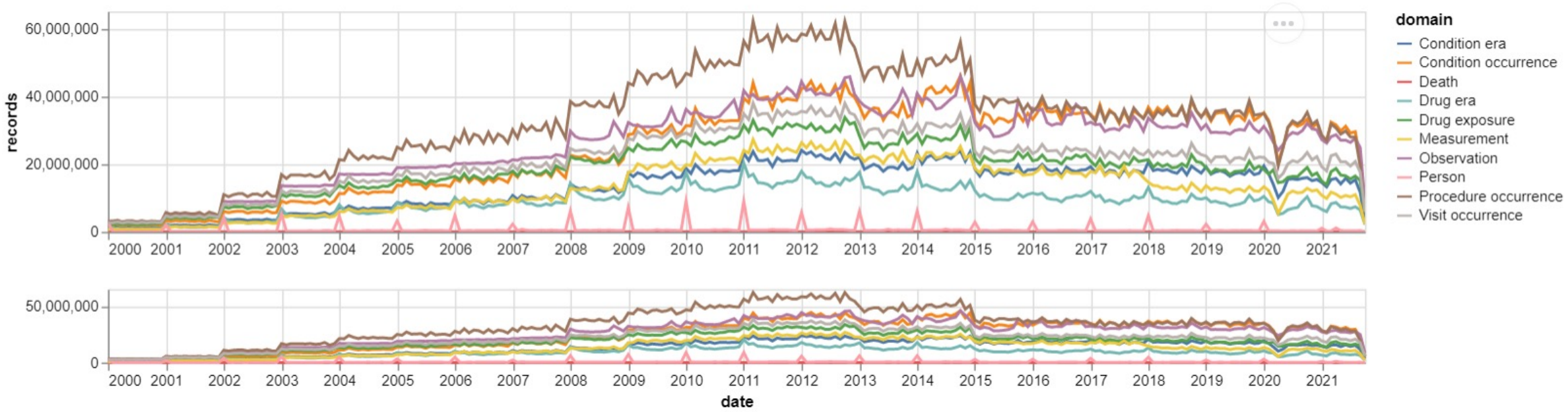
Report Category: Data Source Release

Data Source: IBM CCAE

Data Source Release: 2022-01-22

Report: Data Density

Domain Density



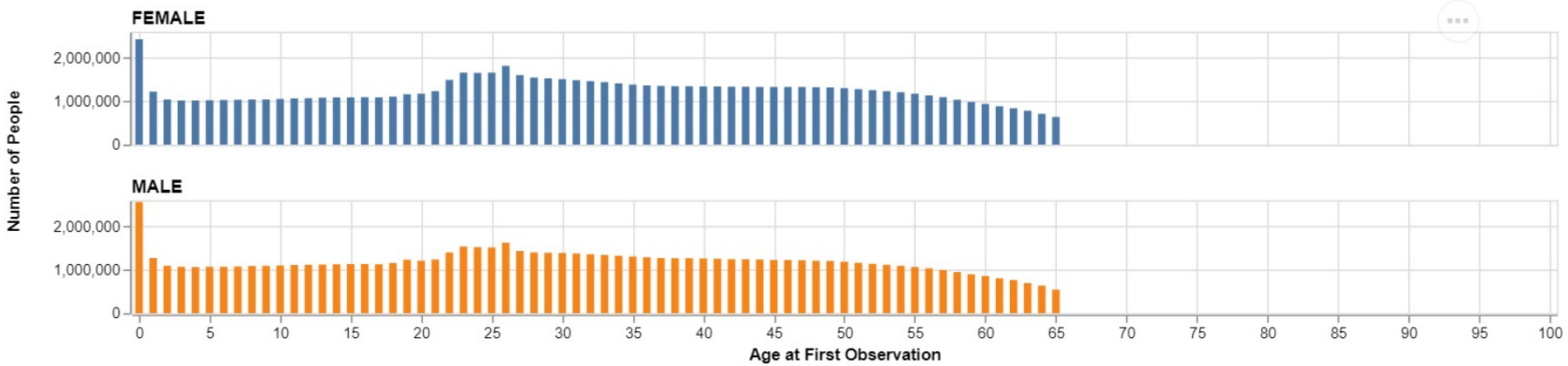
Report Category: Data Source Release | Data Source: IBM CCAE | Data Source Release: 2022-01-22 | Report: Person

PERSON REPORT

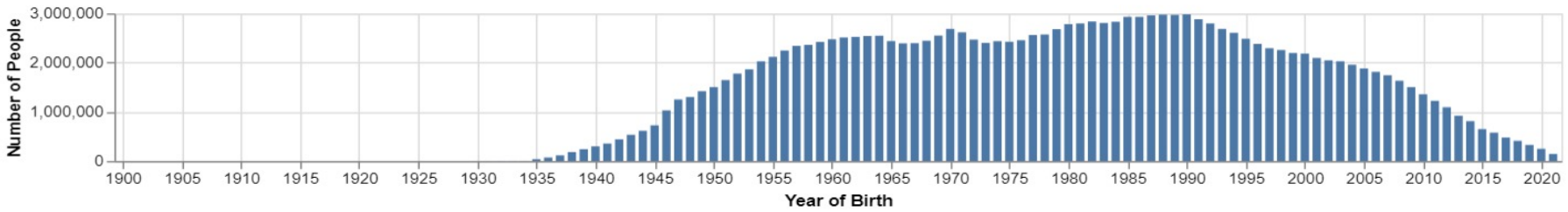
160,465,311
Number of People

Male: 78,467,927 (49%) | Female: 81,997,384 (51%)
Gender Proportions

Population by Age & Sex



Population by Year of Birth



Report Category: Data Source Release

Data Source: IBM CCAE

Data Source Release: 2022-01-22

Report: Visit Occurrence

VISIT OCCURRENCE

Search in Table

CHOOSE COLUMNS TO DISPLAY

Concept Id	Concept Name	↓ % People
9202	Outpatient Visit	**80.64 %**
581458	Pharmacy visit	**63.08 %**
9203	Emergency Room Visit	**23.88 %**
32036	Laboratory Visit	16.86 %
9201	Inpatient Visit	13.06 %

Report Category: Data Source Release

Data Source: IBM CCAE

Data Source Release: 2022-01-22

Report: Conditions

Concept ID: 37311061

COVID-19

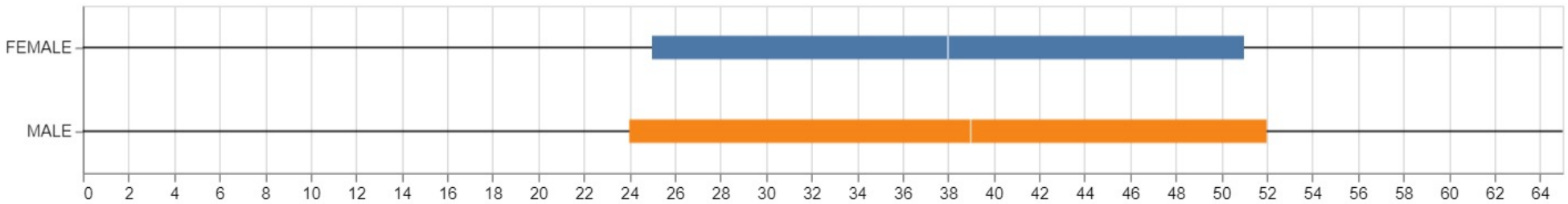
ID 37311061
Concept Identifier

1,585,889
Number of People

1%
% of People

2.3
Records per Person

Age at First Diagnosis




```
CATEGORY: MALE
 MIN_VALUE: 0
 P10_VALUE: 15
 P25_VALUE: 24
 MEDIAN_VALUE: 39
 P75_VALUE: 52
 P90_VALUE: 59
 MAX_VALUE: 65
```