OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

ETL Training – Day 2

Agenda

Aug 13 (Korea Time)	Contents	Speakers
9:00 – 9:45 AM	ETL Specification Review	Jing Li
9:45 – 11:00 AM	Common issues in ETL Conversion	Mui Van Zandt
11:00 – 11:30 PM	OMOP ETL Development (Lecture)	Mui Van Zandt
11:30 – 12:30 PM	Break	
12:30 – 14:00 PM	OMOP ETL Development (Exercise, Review)	Mui Van Zandt
14:00 – 14:15 PM	Break	
14:15 – 16:15 PM	Data Quality Checks (Lecture, Exercise, Review)	Selva Muthu Kumaran Sathappan

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

ETL Specification Review

ETL Specification Review

- Review ETL Specification Exercise

Common issues in ETL Conversion

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Common OMOP CDM issues

CDM table	Frequently encountered issues	Recommended practice
Person	Person without birth year, or birth year in the future	Remove these person from OMOP conversion. OMOP considers them as invalid person.
Condition_occurrence	No standard condition_concept_id mapping available	Populate with standard concept IDs where domain_id='Condition' AND standard_concept='S'. Otherwise, populate this field with 0.
	Condition_source_concept_id has no mapped vocabulary	Populate with mapped concept id corresponding to the source value, otherwise populate with 0.
	Foreign Key provider_id not from Provider table	Check provider_id not null values, if they have the same provider_id in Provider table. If not leave it as null.
Visit_occurrence	Foreign Key provider_id not from Provider table	Check provider_id not null values, if they have the same provider_id in Provider table. If not leave it as null.
	Discharge_to_concept_id has no mapped standard vocabulary	Populate with standard concept IDs where domain_id='Visit' and standard_concept='S'.

Common OMOP CDM issues – Continued

CDM table	Frequently encountered issues	Recommended practice
Drug_exposure	No standard drug_concept_id mapping available	Populate with standard concept IDs where domain_id='Drug' AND standard_concept='S'. Otherwise, populate this field with 0.
	Drug_type_concept_id is not standard and not in Type Concept domain	Populate with standard concept ids which best represents the provenance of the record. Generally, populate with 32810 for Claim or 32817 for HER.
	Drug quantity < 1	Check if quantity < 1 is reasonable from medical perspective. Could leave it < 1 if reasonable.
	Drug_exposure_start_datetime not in correct format	If time value is available, set it between 00:00:00 and 23:59:59. Otherwise, set it to midnight 00:00:00.
	Days_supply <1	Check why days_supply < 1. Default to 1 if not available.
Drug_era	Drug_era_end_date is abnormal	Populate with the end date of the last Drug Exposure.
Device_exposure	Device quantity < 1	Default to 1 if it is 0, null or negative values.

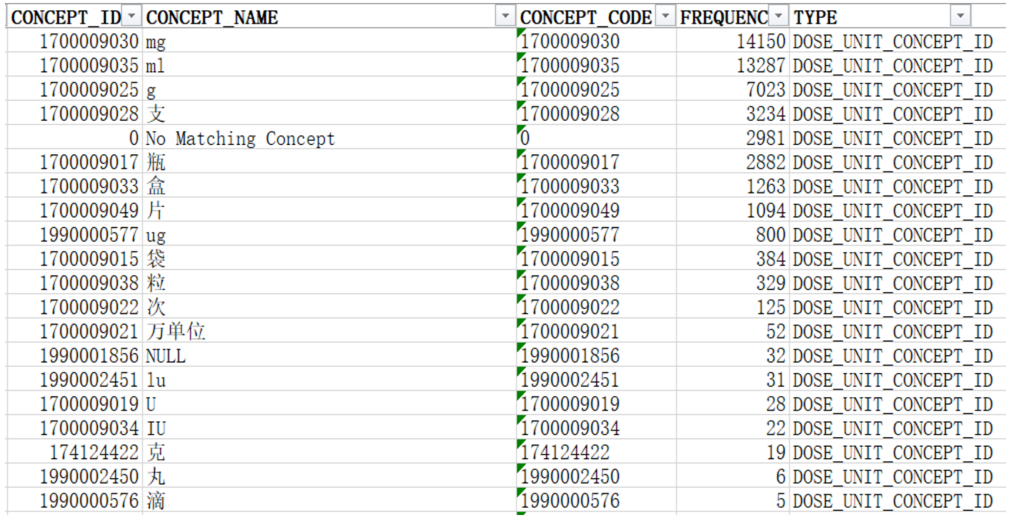
Common OMOP CDM issues – Continued

CDM table	Frequently encountered issues	Recommended practice
Death	Death dates outside observation period	Refresh Observation_period table if Death table data is updated.
	Events after death	Check and fix abnormal event dates. Otherwise remove these person events.
Observation_period	Person without observation period	Remove these person if no events associated with them.
	Events outside observation period	Refresh Observation_period table after event tables updated and abnormal event dates deleted.
Measurement	Value_as_number is negative or null	Check if it's reasonable from medical perspective, or they have value_as_concept_id populated.
	Measurenment_date outside of birth date and death date	Check the records of measurement_date before person's birth date or after death date + 60 days and fix.
Event tables	Records go to wrong CDM tables with different domain	Records should go to corresponding tables with their domain.

1. No standard vocabulary

Issue

- Text fields
- Duplicate and unclear values in source concept names
- Proprietary coding system
- No OMOP standard vocabulary mapping available even though vocabulary is in Athena



CONCEPT_ID	CONCEPT_NAME	CONCEPT_CODE	FREQUENC	TYPE
1700009030	mg	1700009030	14150	DOSE_UNIT_CONCEPT_ID
1700009035	ml	1700009035	13287	DOSE_UNIT_CONCEPT_ID
1700009025	g	1700009025	7023	DOSE_UNIT_CONCEPT_ID
1700009028	支	1700009028	3234	DOSE_UNIT_CONCEPT_ID
0	No Matching Concept	0	2981	DOSE_UNIT_CONCEPT_ID
1700009017	瓶	1700009017	2882	DOSE_UNIT_CONCEPT_ID
1700009033	盒	1700009033	1263	DOSE_UNIT_CONCEPT_ID
1700009049	片	1700009049	1094	DOSE_UNIT_CONCEPT_ID
1990000577	ug	1990000577	800	DOSE_UNIT_CONCEPT_ID
1700009015	袋	1700009015	384	DOSE_UNIT_CONCEPT_ID
1700009038	粒	1700009038	329	DOSE_UNIT_CONCEPT_ID
1700009022	次	1700009022	125	DOSE_UNIT_CONCEPT_ID
1700009021	万单位	1700009021	52	DOSE_UNIT_CONCEPT_ID
1990001856	NULL	1990001856	32	DOSE_UNIT_CONCEPT_ID
1990002451	lu	1990002451	31	DOSE_UNIT_CONCEPT_ID
1700009019	U	1700009019	28	DOSE_UNIT_CONCEPT_ID
1700009034	IU	1700009034	22	DOSE_UNIT_CONCEPT_ID
174124422	克	174124422	19	DOSE_UNIT_CONCEPT_ID
1990002450	丸	1990002450	6	DOSE_UNIT_CONCEPT_ID
1990000576	滴	1990000576	5	DOSE_UNIT_CONCEPT_ID

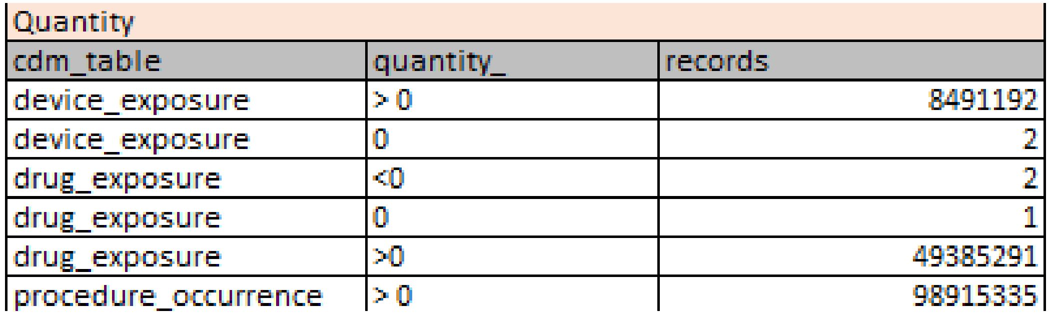
Solution

- Own Mapping Team
  - Mapped translated terms to OMOP standard vocabulary
- OMOP Vocabulary Team
  - Prioritized terms for mapping
  - Verify translated terms
  - Confirm translation with medical team
  - Downloaded latest vocabularies
- If cannot map to a standard vocabulary, use concept_id = 0

2. Abnormal values

Issue

- Negative, 0, decimals, null values of quantity in device_exposure and drug_exposure table
- Negative, null values of value_as_number in measurement table
- Person year_of_birth before 1900 or in the future



Quantity		
cdm_table	quantity_	records
device_exposure	> 0	8491192
device_exposure	0	2
drug_exposure	<0	2
drug_exposure	0	1
drug_exposure	>0	49385291
procedure_occurrence	> 0	98915335

Solution

- Default to 1 for quantity in device_exposure table
- Check source data if it’s reasonable from medical perspective, or they have value_as_concept_id populated
- Check person source birth date
- If valid, leave the values as they were. If not, remove the records as dirty data

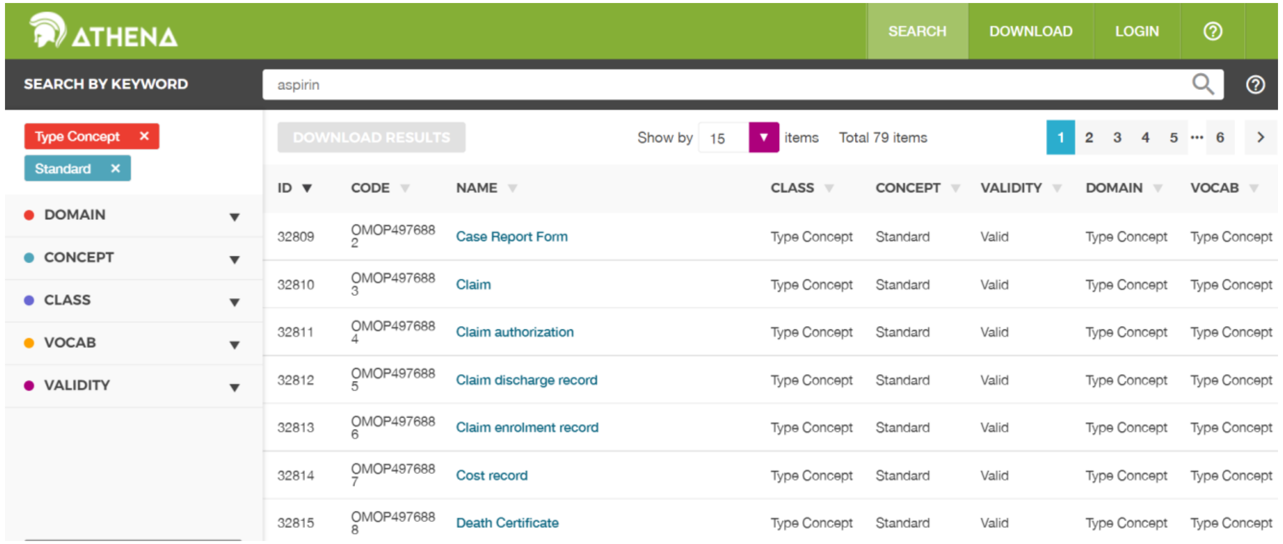
3. Wrong type_concept_ids

Issue

- Wrong provenance of records were assigned to type_concept_ids
- Type_concept_id is not standard

Solution

- Find the best provenance of the record using ATHENA
- Standardize all type_concept_ids in each table
- Guidelines:
  - Use 32810 for Claim
  - Use 32817 for EHR



aspirin

Show by 15 items Total 79 items

ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
32809	OMOP4976882	Case Report Form	Type Concept	Standard	Valid	Type Concept	Type Concept
32810	OMOP4976883	Claim	Type Concept	Standard	Valid	Type Concept	Type Concept
32811	OMOP4976884	Claim authorization	Type Concept	Standard	Valid	Type Concept	Type Concept
32812	OMOP4976885	Claim discharge record	Type Concept	Standard	Valid	Type Concept	Type Concept
32813	OMOP4976886	Claim enrolment record	Type Concept	Standard	Valid	Type Concept	Type Concept
32814	OMOP4976887	Cost record	Type Concept	Standard	Valid	Type Concept	Type Concept
32815	OMOP4976888	Death Certificate	Type Concept	Standard	Valid	Type Concept	Type Concept

4. Missing CDM tables

Issue

- Incomplete OMOP CDM tables
- Potential Missing Tables:
  - Procedure_occurrence
  - Device_exposure
  - Visit_occurrence
  - Observation
  - Payer_plan_period
  - Dose_era
  - Location
  - Cost
  - Death
  - Provider

Solution

- Check source data for related domains
- Provide mapping rules from source data to OMOP CDM, and populate the missing tables

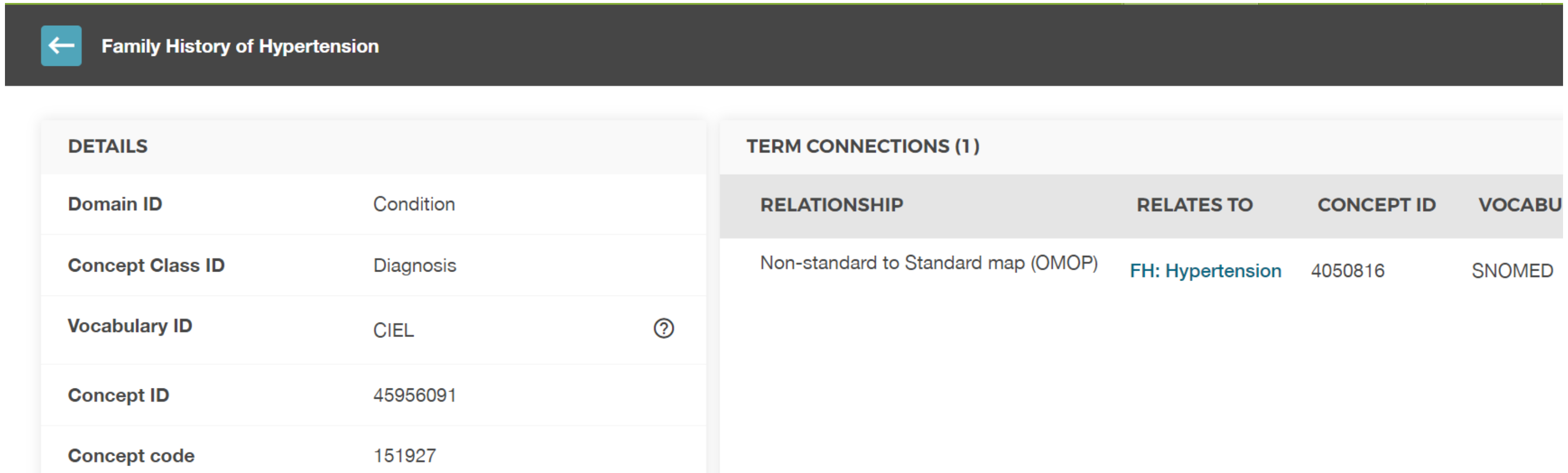
5. Mapping to the wrong domain

Issue

- The source vocabulary domain may differ from its mapped standard vocabulary domain
- Example: for CIEL 151927 – Family History of Hypertension, it maps to concept 4050816 FH: Hypertension, which is not in Condition domain, but Observation domain

Solution

- Use the domain from the mapped OMOP standard vocabulary, not the source vocabulary domain
- For each table, the standard concepts should all be from the corresponding domain



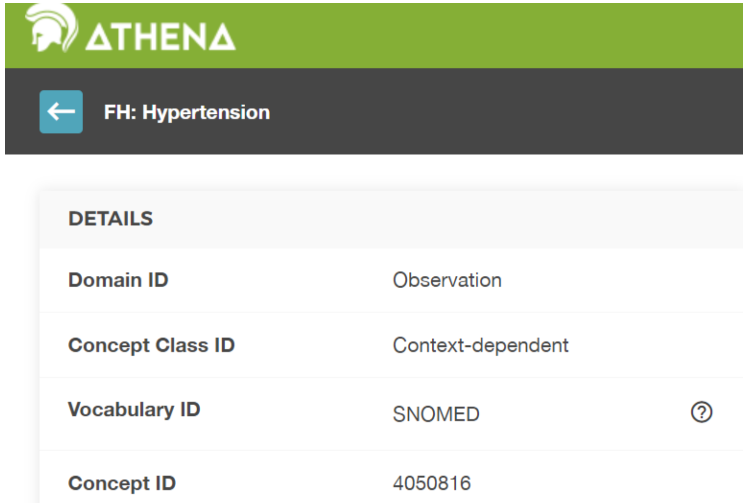
Family History of Hypertension

DETAILS

Domain ID	Condition
Concept Class ID	Diagnosis
Vocabulary ID	CIEL
Concept ID	45956091
Concept code	151927

TERM CONNECTIONS (1)

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABU
Non-standard to Standard map (OMOP)	FH: Hypertension	4050816	SNOMED



FH: Hypertension

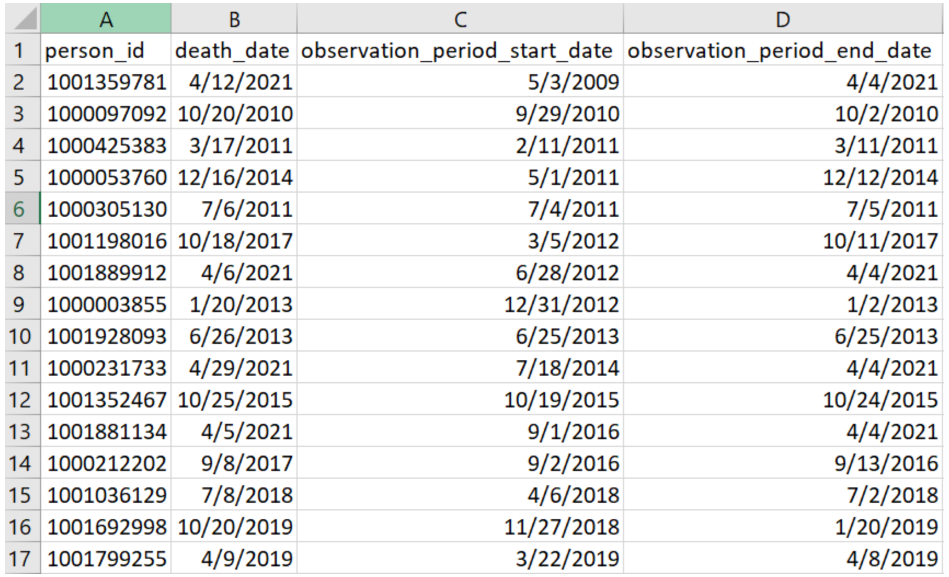
DETAILS

Domain ID	Observation
Concept Class ID	Context-dependent
Vocabulary ID	SNOMED
Concept ID	4050816

6. Incorrect logic for Observation_period

Issue

- Not each person has observation period
- Observation period for patients not cover the whole time period of all events
- Observation period end date less than observation period start date



	A	B	C	D
1	person_id	death_date	observation_period_start_date	observation_period_end_date
2	1001359781	4/12/2021	5/3/2009	4/4/2021
3	1000097092	10/20/2010	9/29/2010	10/2/2010
4	1000425383	3/17/2011	2/11/2011	3/11/2011
5	1000053760	12/16/2014	5/1/2011	12/12/2014
6	1000305130	7/6/2011	7/4/2011	7/5/2011
7	1001198016	10/18/2017	3/5/2012	10/11/2017
8	1001889912	4/6/2021	6/28/2012	4/4/2021
9	1000003855	1/20/2013	12/31/2012	1/2/2013
10	1001928093	6/26/2013	6/25/2013	6/25/2013
11	1000231733	4/29/2021	7/18/2014	4/4/2021
12	1001352467	10/25/2015	10/19/2015	10/24/2015
13	1001881134	4/5/2021	9/1/2016	4/4/2021
14	1000212202	9/8/2017	9/2/2016	9/13/2016
15	1001036129	7/8/2018	4/6/2018	7/2/2018
16	1001692998	10/20/2019	11/27/2018	1/20/2019
17	1001799255	4/9/2019	3/22/2019	4/8/2019

Solution

- Check why person has no observation period, invalid person IDs or person without dates?
- Refresh observation table after event tables updated and abnormal event dates deleted
- Check the logic of generating observation period
- If there is no event associated with the patient, delete the patient records

7. Loss of data

Issue

- Less patients once converted to OMOP
- Not all fields are mapped to OMOP

Solution

- Logic is introduced to ensure patients are valid
  - Test patients
  - Patients without birth year
  - Patients without any transaction
    - Depends on the data and scenario
- Some fields are used to derive the logic of the CDM field
  - For example: ICD Type helps determine if the code is an ICD9 or ICD10 code
- Duplicate records
  - The same diagnosis within the same day of a hospital stay

8. Wrong foreign key identifiers and datetime format

Issue

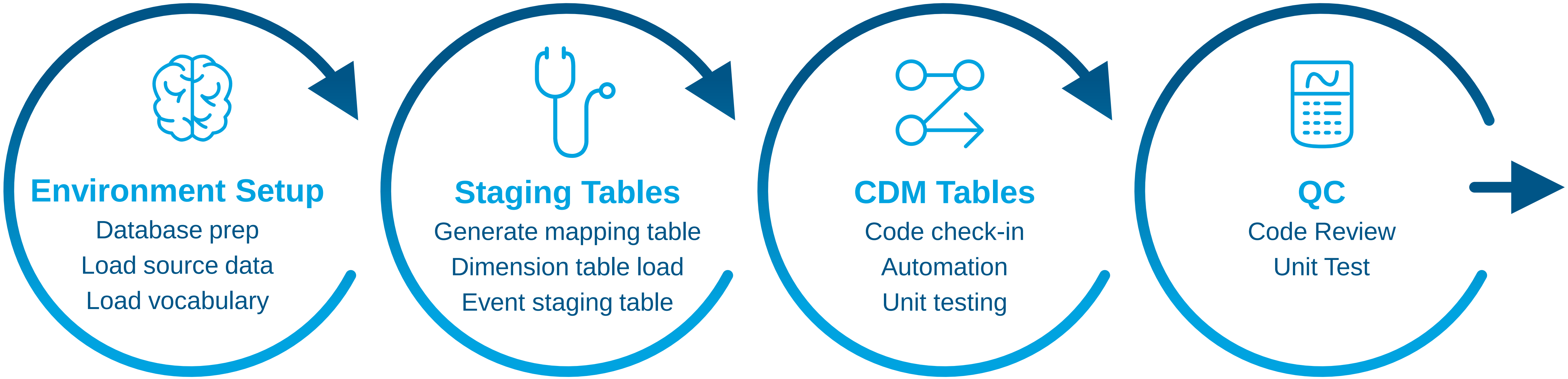
- Foreign key identifiers are invalid
- Event datetime value not in correct format

Solution

- Check why FK identifiers not work, wrong person IDs or ETL logic?
- If time value is available, set it between 00:00:00 and 23:59:59. Otherwise, set it to midnight 00:00:00

OMOP ETL Development

OMOP ETL Workflow



Environment Setup

Database prep

Load source data

Load vocabulary

Staging Tables

Generate mapping table

Dimension table load

Event staging table

CDM Tables

Code check-in

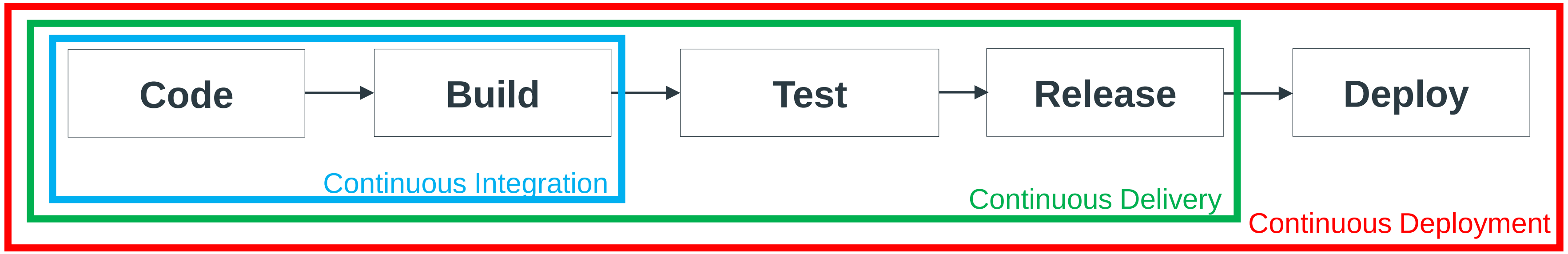
Automation

Unit testing

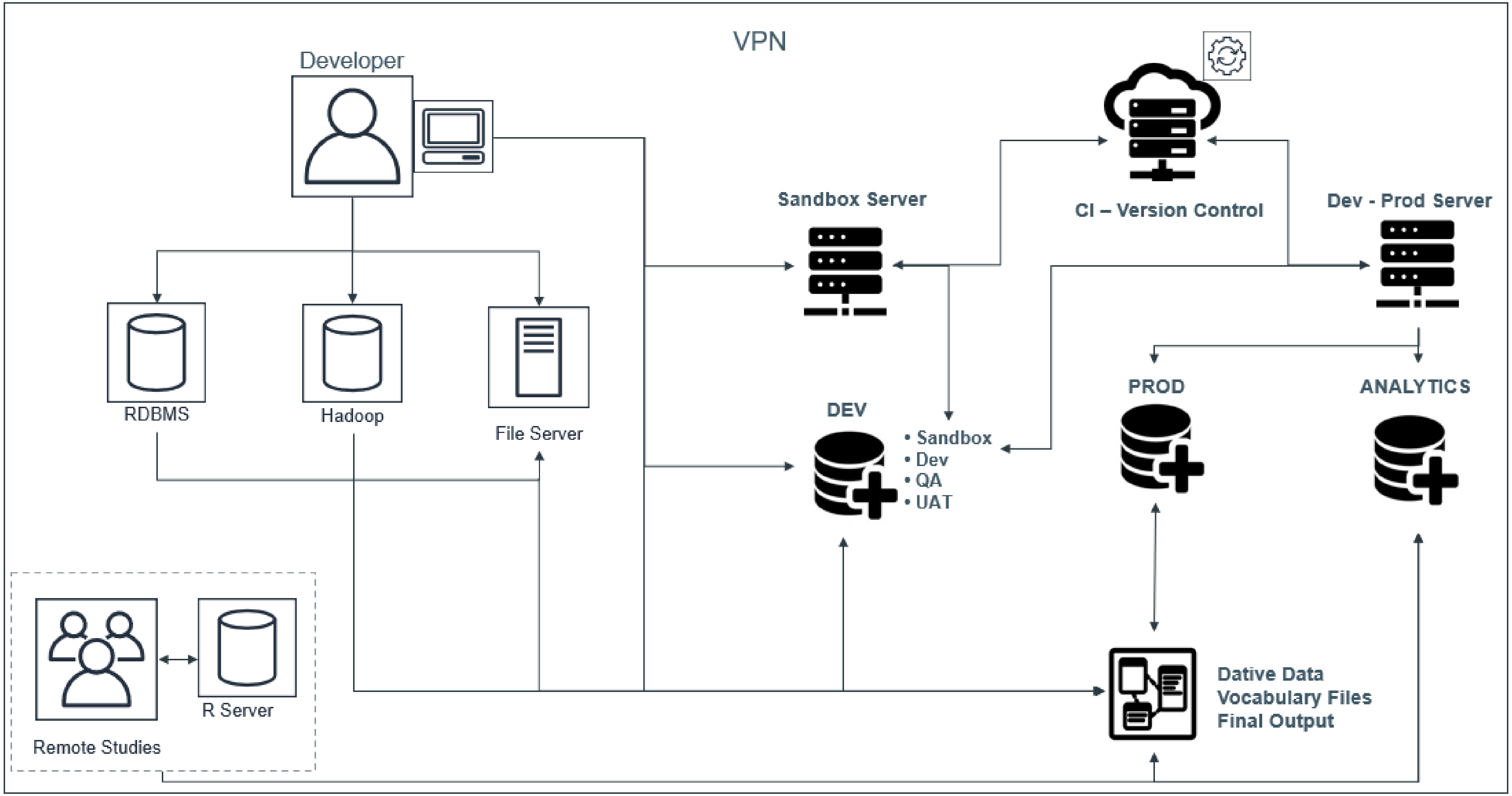
QC

Code Review

Unit Test



ETL Environment



ETL Process

1

Source data to staging table

- DDL
- Load source data
- Load custom mapping
- Load standard vocabulary
- Load staging table

2

Staging table to OMOP CDM tables

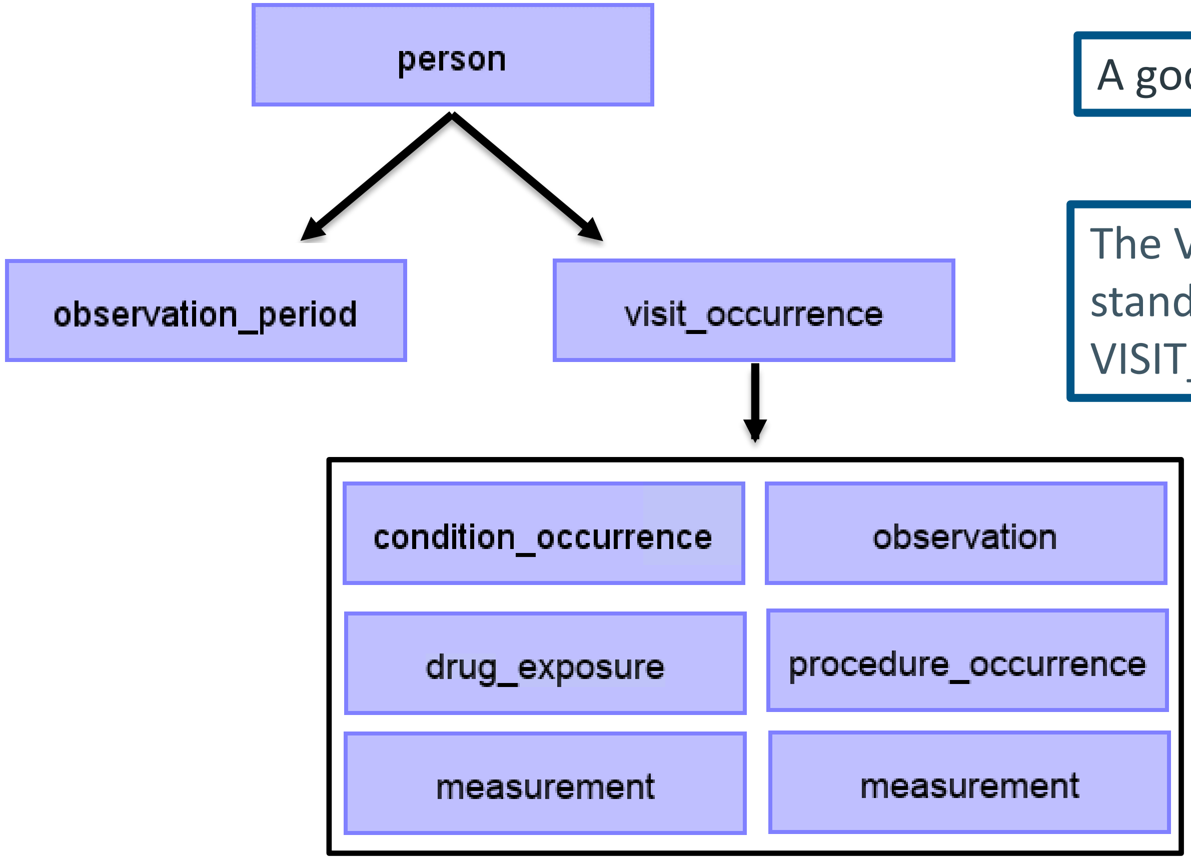
- Generate mapping table
- Dimension table load
- Fact table load

3

QA and Validation

- QA
- Validation
- Demostats
- Achilles

ETL Implementation



A good rule of thumb is to always create the PERSON table first

The VISIT_OCCURRENCE table must be created before the standardized clinical data tables as they all refer to the VISIT_OCCURRENCE_ID

Step 1: Source data to staging table

1. DDL – Create the tables where the source data will land

2. Load the source data into the tables you created

3. Load the custom mapping data into their tables

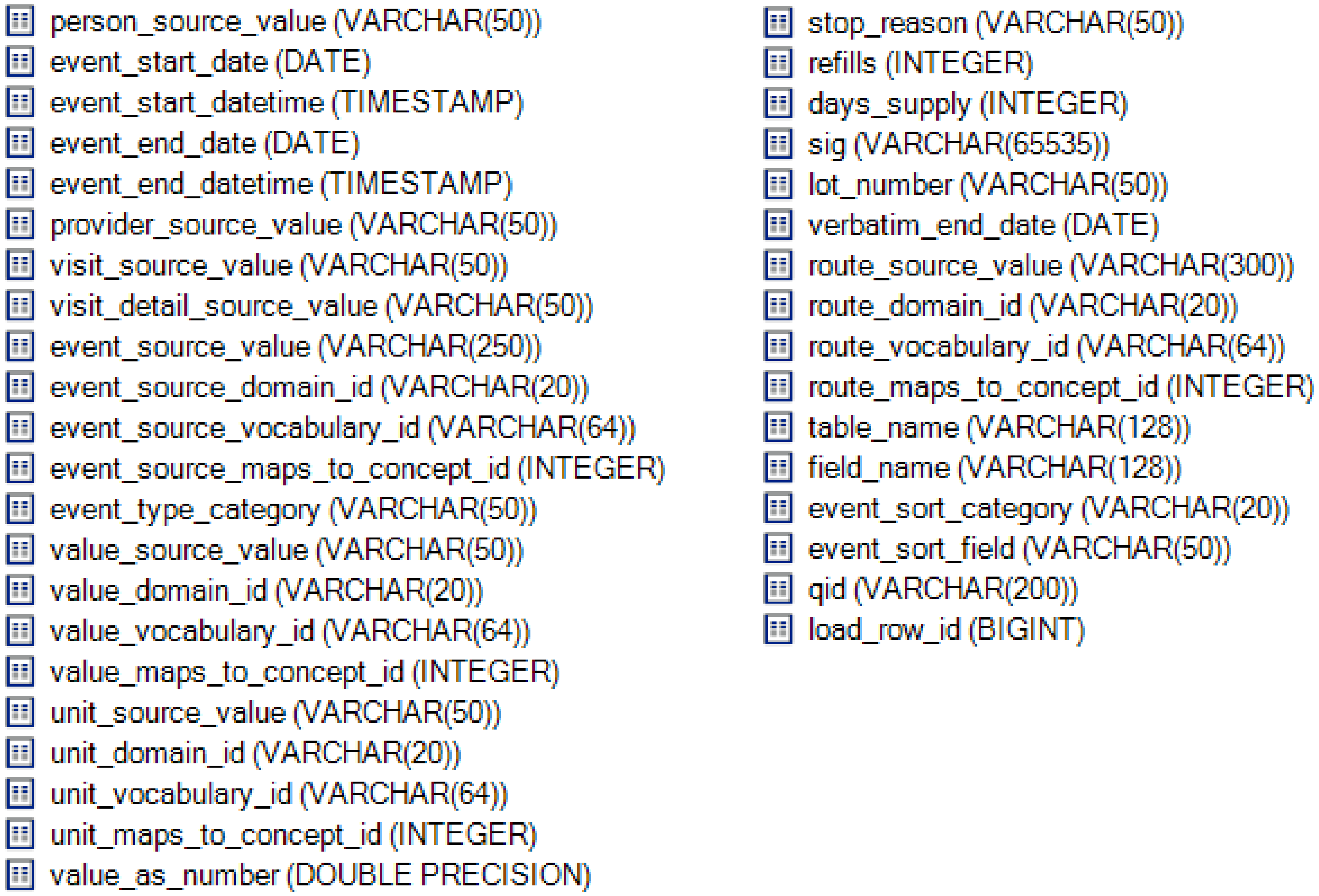
4. Load the vocabulary data from OHDSI into tables

5. Create the code that will move the source data into the staging table

Step 1: Source data to staging table – Staging table

All events associated with patient

- Visit occurrence
- Observation
- Condition occurrence
- Procedure occurrence
- Measurement
- Drug exposure
- Device exposure
- Specimen



person_source_value (VARCHAR(50))
event_start_date (DATE)
event_start_datetime (TIMESTAMP)
event_end_date (DATE)
event_end_datetime (TIMESTAMP)
provider_source_value (VARCHAR(50))
visit_source_value (VARCHAR(50))
visit_detail_source_value (VARCHAR(50))
event_source_value (VARCHAR(250))
event_source_domain_id (VARCHAR(20))
event_source_vocabulary_id (VARCHAR(64))
event_source_maps_to_concept_id (INTEGER)
event_type_category (VARCHAR(50))
value_source_value (VARCHAR(50))
value_domain_id (VARCHAR(20))
value_vocabulary_id (VARCHAR(64))
value_maps_to_concept_id (INTEGER)
unit_source_value (VARCHAR(50))
unit_domain_id (VARCHAR(20))
unit_vocabulary_id (VARCHAR(64))
unit_maps_to_concept_id (INTEGER)
value_as_number (DOUBLE PRECISION)

stop_reason (VARCHAR(50))
refills (INTEGER)
days_supply (INTEGER)
sig (VARCHAR(65535))
lot_number (VARCHAR(50))
verbatim_end_date (DATE)
route_source_value (VARCHAR(300))
route_domain_id (VARCHAR(20))
route_vocabulary_id (VARCHAR(64))
route_maps_to_concept_id (INTEGER)
table_name (VARCHAR(128))
field_name (VARCHAR(128))
event_sort_category (VARCHAR(20))
event_sort_field (VARCHAR(50))
qid (VARCHAR(200))
load_row_id (BIGINT)

Step 2: Staging table to OMOP CDM tables

1 Generate mapping table

- Create the table that will map your source codes to the standard vocabulary

2 Load dimension tables

- Includes person, location, care_site, and provider
- As these are not events, they must be coded separately from the event staging table
- Data from these tables will be used in the fact tables, so this is done first

3 Load fact tables

- Also called event tables, these are where the events go
- The bulk of the data goes into these tables
- Much of the work mapping for these tables will have been done in step 1

Step 3: Code and Unit Testing

QA

- Code review and more in-depth quality assurance
- Code reviews are quick double checks on your code
- More robust QA should be performed at the end as a final check

Validation

- In validation the data itself is checked for consistency and integrity
- Standardized code can be used to check the whole dataset quickly for obvious errors

Achilles

- A program developed by OHDSI to find errors in the data
- Descriptive statistical analysis with reporting and data quality checks

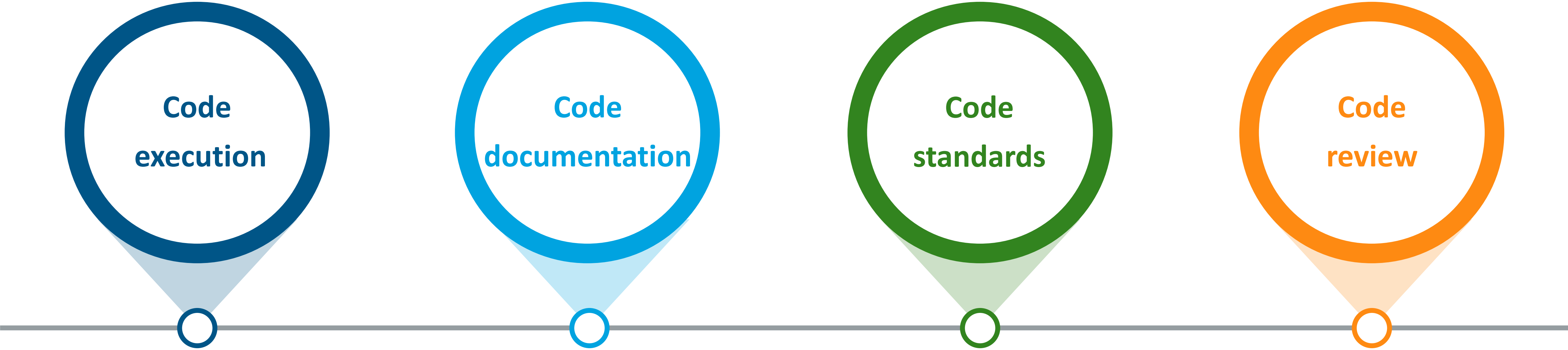
Code Review After ETL Development

- Peer review of new/modified code
- Allows for “another set of eyes”
- Designed to catch bugs/errors
- Enforces standards
- Knowledge Transfer/information sharing
- Reduce rework/troubleshooting in the future

```
/// </summary>
/// <param name="orderedChilIds">A collection of child ids.</param>
/// <param name="movedChildId">The id of the moved child.</param>
public void ChangeChildSortOrder(int[] orderedChilIds, int movedChildId)
{
    if (orderedChilIds == null)
    {
        throw new ArgumentNullException("orderedChildrenIds");
    }

    bool found = false;
    ItemToItem moved = null;
    ItemToItem previous = null;
    ItemToItem next = null;
    foreach (int orderedChildId in orderedChil
    {
        ItemToItem current = ChildItems.FirstO          => c.ChildI
        if (current != null)
        {
            if (current.ChildItem.ItemId == movedChil
            {
                moved = current;
                found = true;
            }
            else
            {
```

Coding Best Practices



Code execution

Does it work?

Does it follow the ETL rules?

Code documentation

Can it be interpreted?

Is a guideline or SOP?

Code standards

Are there coding standards?

Is the code written in the most efficient way?

Code review

Does it conform to internal guidelines?

Does both developer understand the same ETL rules?

Version Control

Version control – Why

Helps manage changes to a software system over time

Allows easy recovery from mistakes or accidents

Necessary when multiple users are making changes to your system

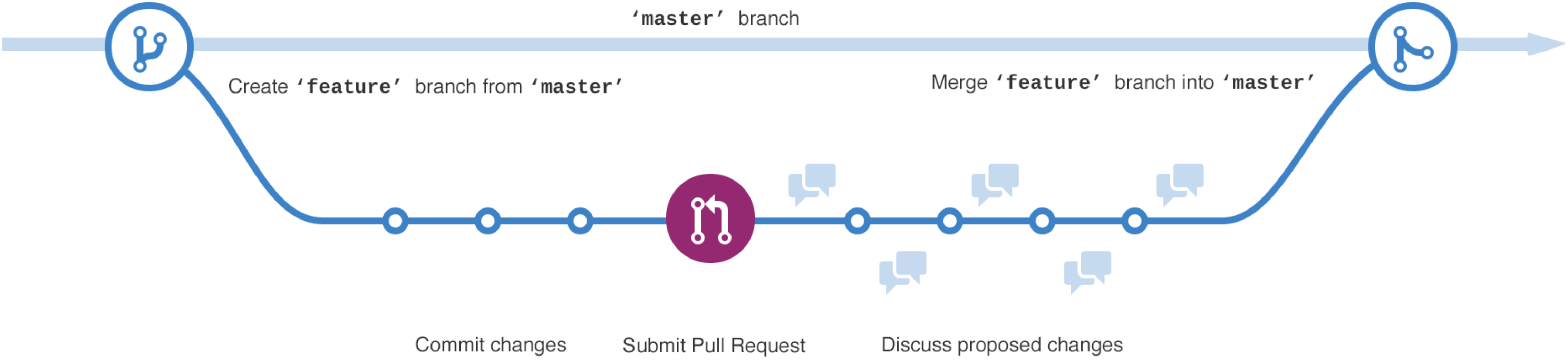
Enables branching and versioning for development, QA and production

Version control – How

Create branch for bug fixing or new feature development

Merge branch to master when it's ready

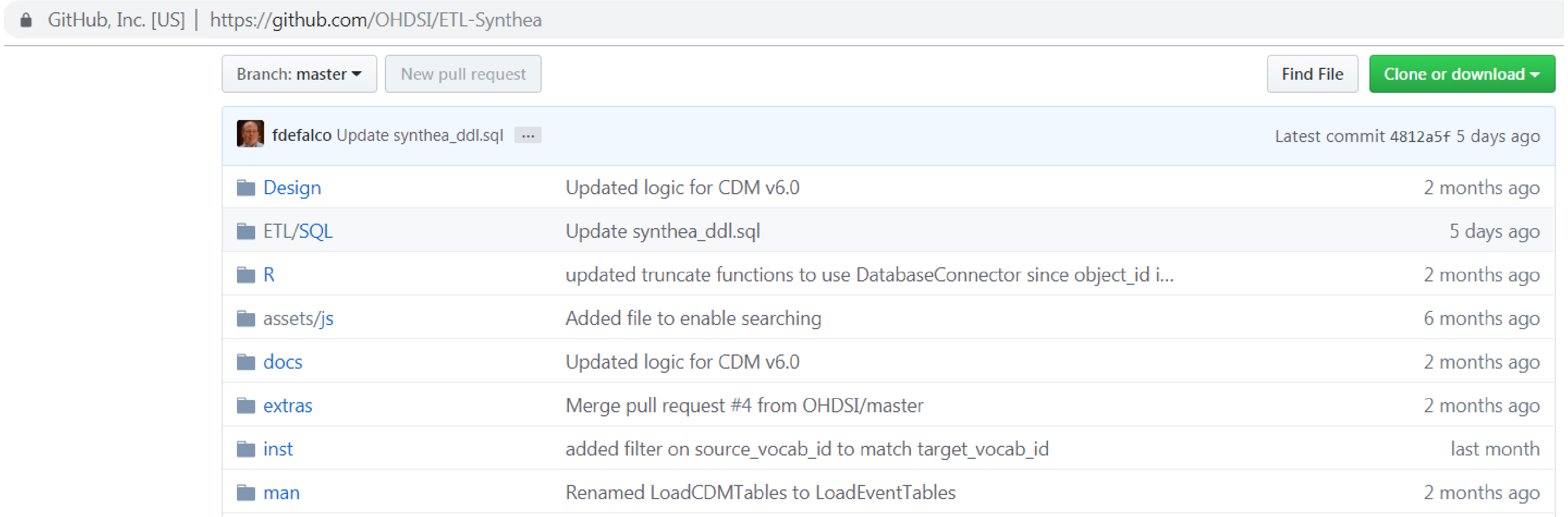
Examples of version control include **Subversion (SVN), Github, GitLab, Bitbucket**



'master' branch
Create 'feature' branch from 'master'
Merge 'feature' branch into 'master'
Commit changes
Submit Pull Request
Discuss proposed changes

Version control – GitHub

- Via web browser
  - https://github.com/



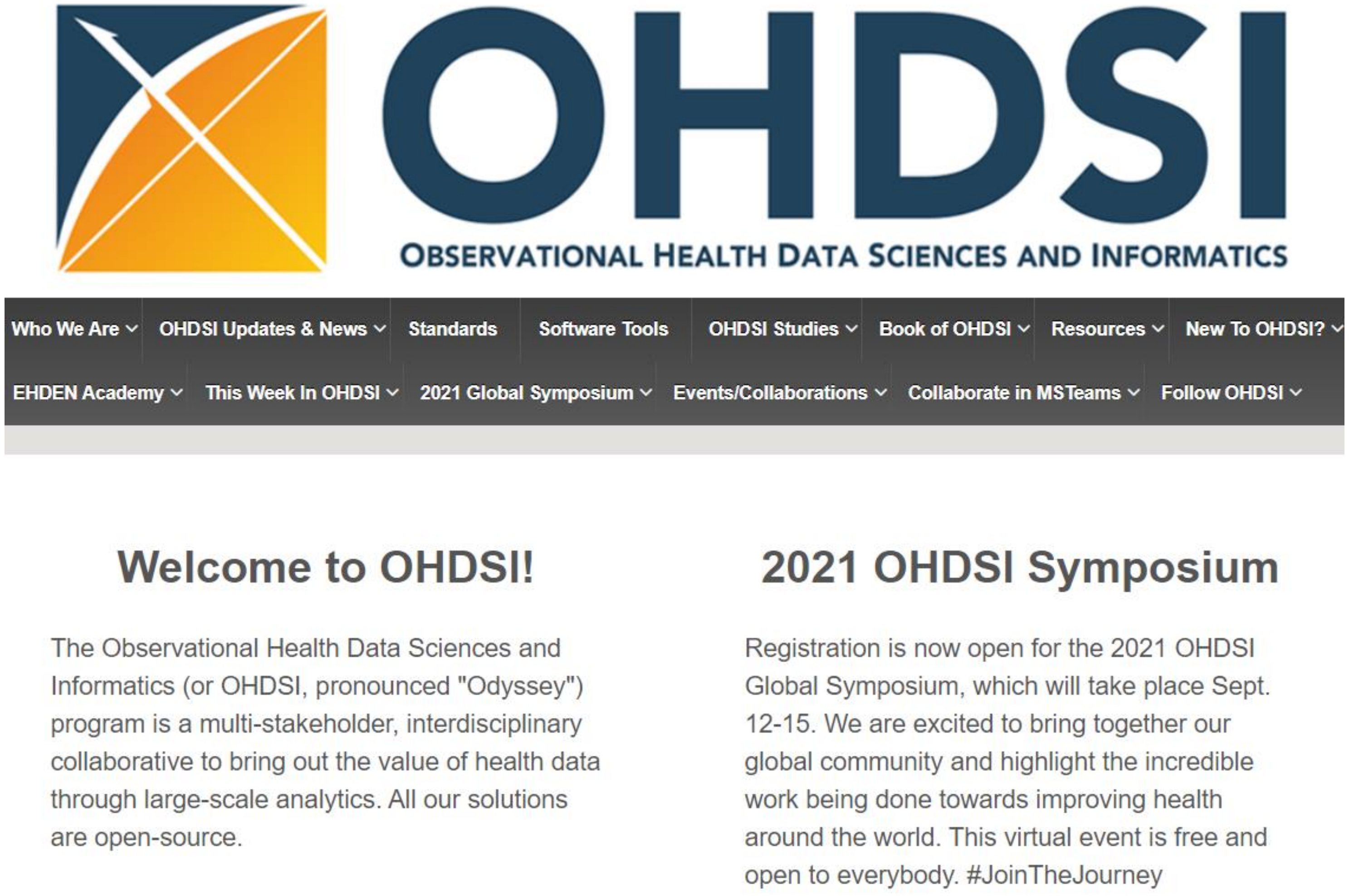
GitHub, Inc. [US] | https://github.com/OHDSI/ETL-Synthea

fdefalco Update synthea_ddl.sql … Latest commit 4812a5f 5 days ago

Design	Updated logic for CDM v6.0	2 months ago
ETL/SQL	Update synthea_ddl.sql	5 days ago
R	updated truncate functions to use DatabaseConnector since object_id i…	2 months ago
assets/js	Added file to enable searching	6 months ago
docs	Updated logic for CDM v6.0	2 months ago
extras	Merge pull request #4 from OHDSI/master	2 months ago
inst	added filter on source_vocab_id to match target_vocab_id	last month
man	Renamed LoadCDMTables to LoadEventTables	2 months ago

OHDSI Technical Resources

OHDSI Site



OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Welcome to OHDSI!

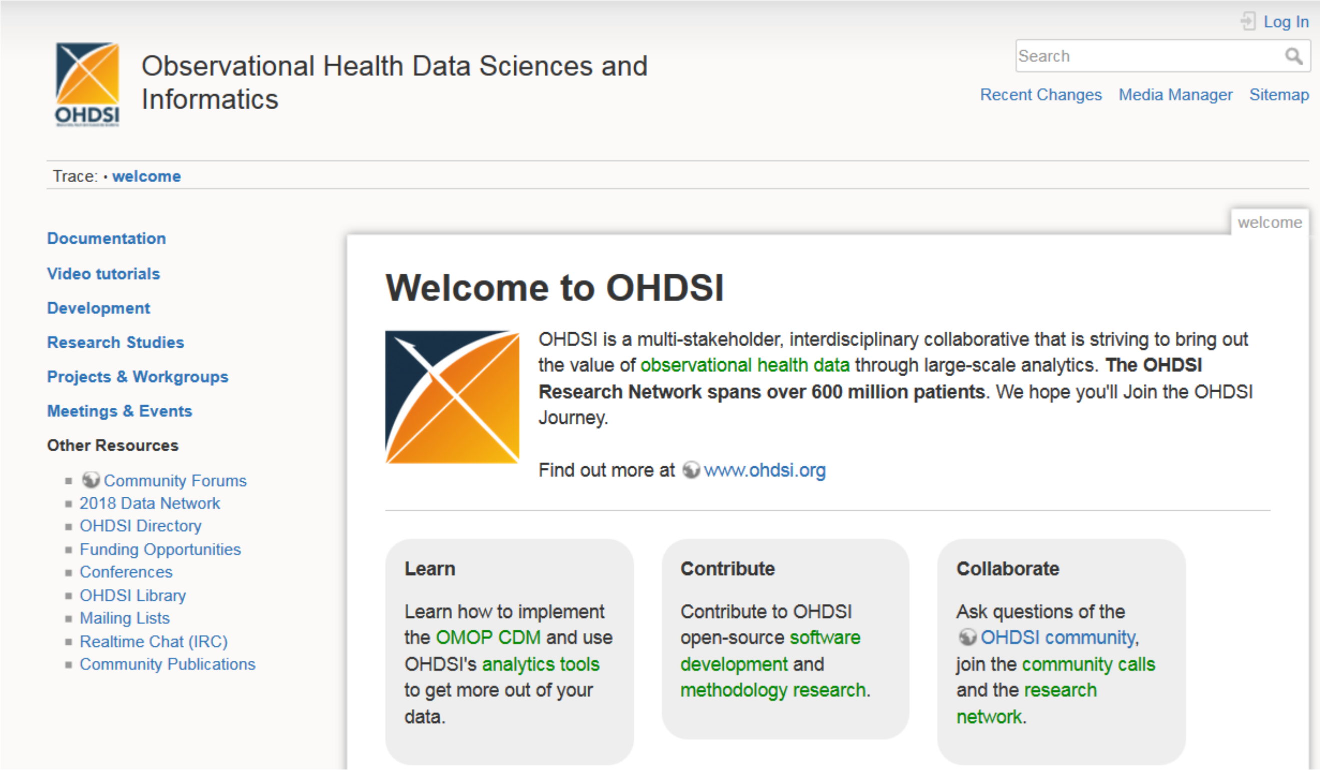
The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source.

2021 OHDSI Symposium

Registration is now open for the 2021 OHDSI Global Symposium, which will take place Sept. 12-15. We are excited to bring together our global community and highlight the incredible work being done towards improving health around the world. This virtual event is free and open to everybody. #JoinTheJourney

Wiki Site

- Helpful documentation and videos
- Notes for meetings and working groups
- Information about studies



Log In

Observational Health Data Sciences and Informatics

Search

Recent Changes Media Manager Sitemap

Trace: • welcome

Documentation

Video tutorials

Development

Research Studies

Projects & Workgroups

Meetings & Events

Other Resources

- Community Forums
- 2018 Data Network
- OHDSI Directory
- Funding Opportunities
- Conferences
- OHDSI Library
- Mailing Lists
- Realtime Chat (IRC)
- Community Publications

welcome

Welcome to OHDSI

OHDSI is a multi-stakeholder, interdisciplinary collaborative that is striving to bring out the value of observational health data through large-scale analytics. **The OHDSI Research Network spans over 600 million patients**. We hope you'll Join the OHDSI Journey.

Find out more at www.ohdsi.org

Learn

Learn how to implement the OMOP CDM and use OHDSI's analytics tools to get more out of your data.

Contribute

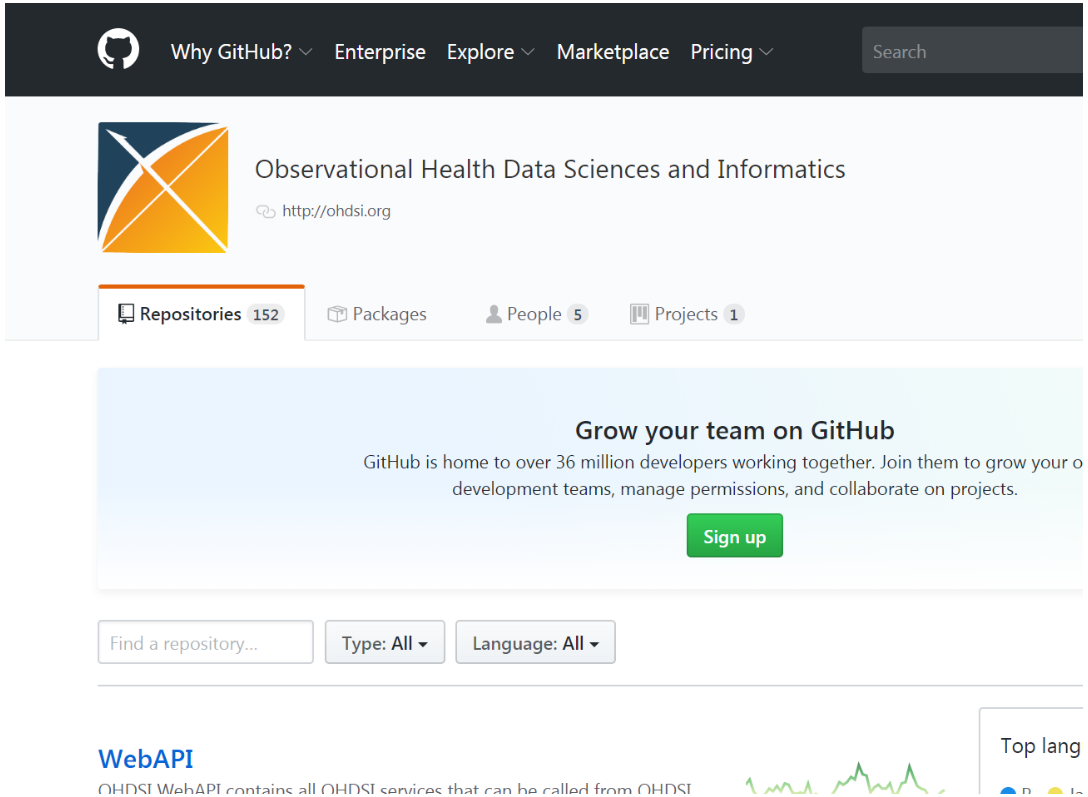
Contribute to OHDSI open-source software development and methodology research.

Collaborate

Ask questions of the OHDSI community, join the community calls and the research network.

GitHub

- Contains code for standard DDL scripts for OMOP database
  - github.com/OHDSI/CommonDataModel
- Wiki site specific to OMOP CDM
  - https://ohdsi.github.io/CommonDataModel/
- Contains all code for available tools such as
  - White Rabbit/Rabbit in a Hat
    - github.com/OHDSI/WhiteRabbit
  - Usagi
    - github.com/OHDSI/Usagi
  - Atlas
    - github.com/OHDSI/Atlas
  - Achillies
    - github.com/OHDSI/Achilles



Observational Health Data Sciences and Informatics

http://ohdsi.org

Grow your team on GitHub

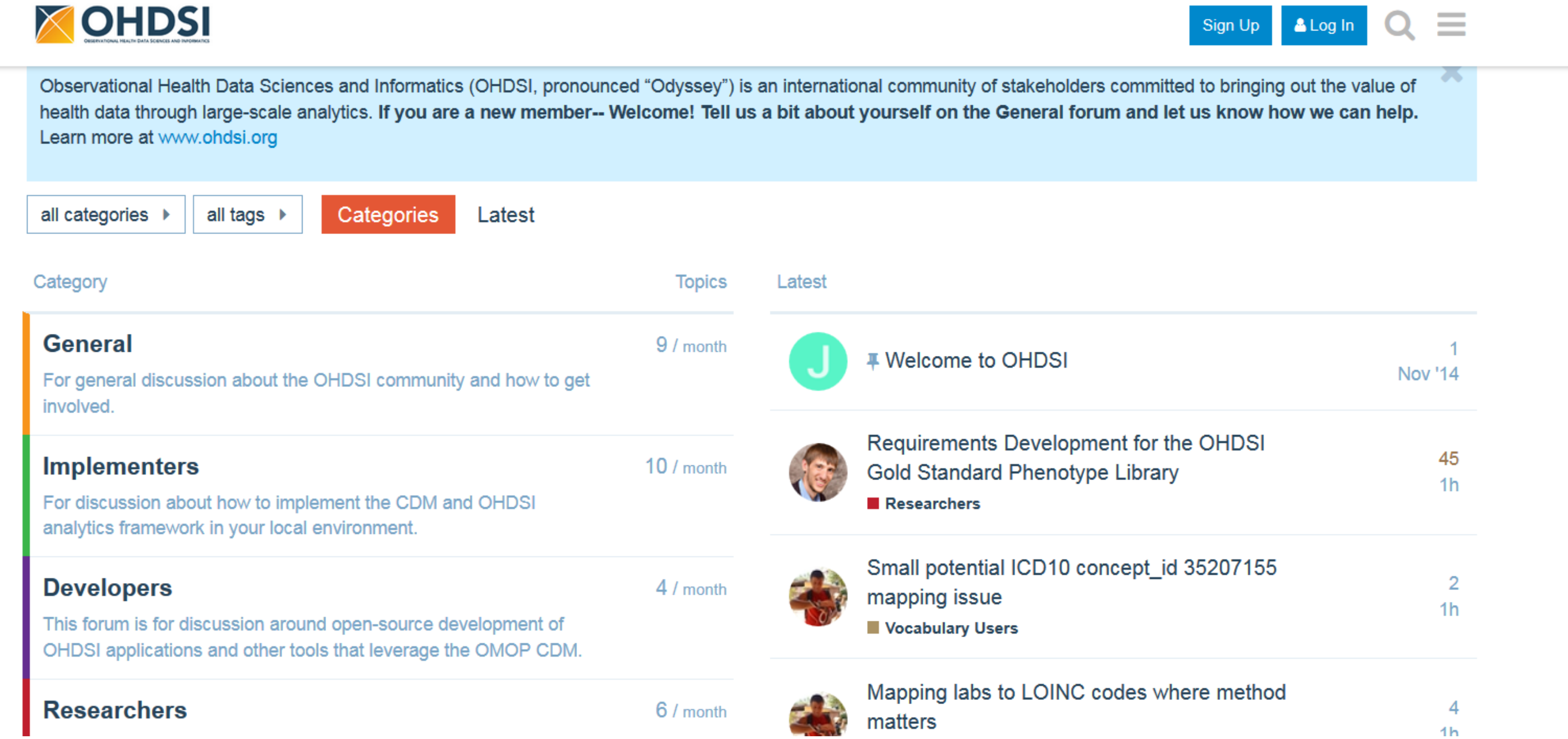
GitHub is home to over 36 million developers working together. Join them to grow your o development teams, manage permissions, and collaborate on projects.

WebAPI

OHDSI WebAPI contains all OHDSI services that can be called from OHDSI

Forums

- Discusses issues and concerns
- Can find answer to questions not in documents
- Helpful to find similar concerns and issues.
- Very active community



OHDSI

Sign Up | Log In

Observational Health Data Sciences and Informatics (OHDSI, pronounced "Odyssey") is an international community of stakeholders committed to bringing out the value of health data through large-scale analytics. **If you are a new member-- Welcome! Tell us a bit about yourself on the General forum and let us know how we can help.** Learn more at www.ohdsi.org

all categories | all tags | Categories | Latest

Category	Topics
General For general discussion about the OHDSI community and how to get involved.	9 / month
Implementers For discussion about how to implement the CDM and OHDSI analytics framework in your local environment.	10 / month
Developers This forum is for discussion around open-source development of OHDSI applications and other tools that leverage the OMOP CDM.	4 / month
Researchers	6 / month

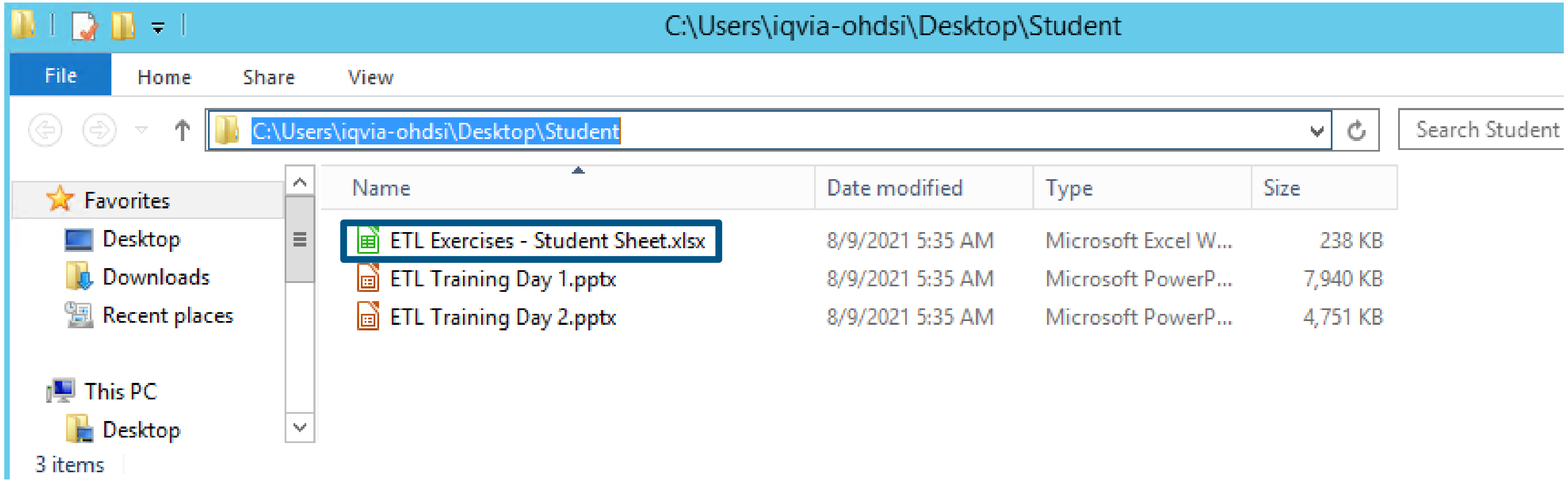
Latest	
Welcome to OHDSI	1 Nov '14
Requirements Development for the OHDSI Gold Standard Phenotype Library Researchers	45 1h
Small potential ICD10 concept_id 35207155 mapping issue Vocabulary Users	2 1h
Mapping labs to LOINC codes where method matters	4 1h

Lunch Break – 1 hour

OMOP ETL Development – Exercise

Exercise

- On the Box, go to folder ‘C:\Users\iqvia-ohdsi\Desktop\Student’
- Open file ‘ETL Exercises - Student Sheet’
- Do exercise in **Day2 ETL Development_1000** tab: take the native data and map to CDM tables
- If time is available, can also do exercises in **Day2 ETL Development_1005** tab and **Day2 ETL Development_1010** tab



C:\Users\iqvia-ohdsi\Desktop\Student

Name	Date modified	Type	Size
ETL Exercises - Student Sheet.xlsx	8/9/2021 5:35 AM	Microsoft Excel W...	238 KB
ETL Training Day 1.pptx	8/9/2021 5:35 AM	Microsoft PowerP...	7,940 KB
ETL Training Day 2.pptx	8/9/2021 5:35 AM	Microsoft PowerP...	4,751 KB

3 items

OHDSI

Observational Health Data Sciences and Informatics

Break – 15min

Data Quality Checks

Data quality checks

Data Quality Dashboard

Free tool developed by OHDSI with over 3,000 quality checks. Designed with FDA and EMA in mind

Data quality checks

Achilles

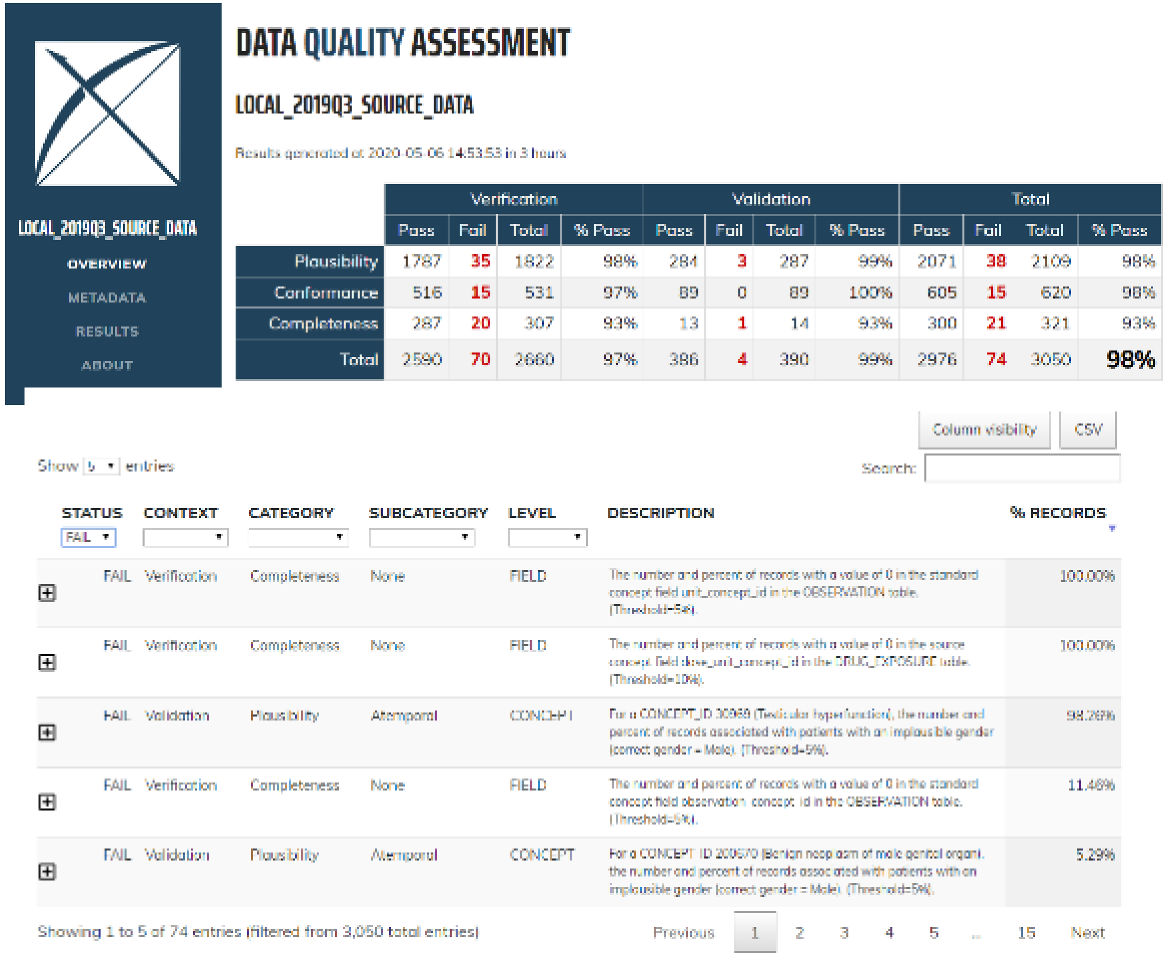
Pre-generated high-level analytics available in a user-friendly webpage

Data Quality Dashboard (DQD)

Description

- Developed in 2019 by OHDSI
  > IQVIA part of core development team
- Follows the Kahn Framework
  > https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/
- 3000+ checks on plausibility, conformance, completeness
- Executed with each data refresh

Deliverable



DATA QUALITY ASSESSMENT

LOCAL_2019Q3_SOURCE_DATA

Results generated at 2020-05-06 14:53:53 in 3 hours

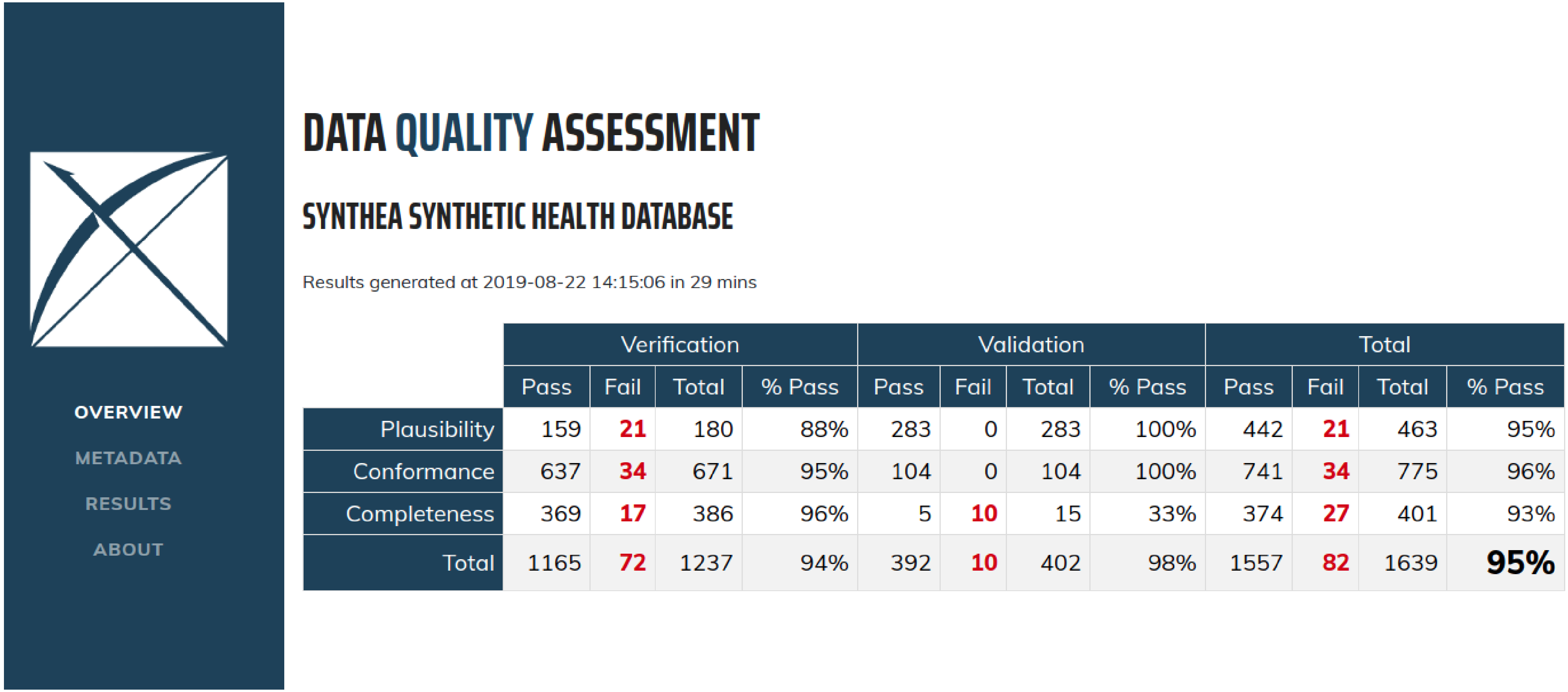
	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1787	35	1822	98%	284	3	287	99%	2071	38	2109	98%
Conformance	516	15	531	97%	89	0	89	100%	605	15	620	98%
Completeness	287	20	307	93%	13	1	14	93%	300	21	321	93%
Total	2590	70	2660	97%	386	4	390	99%	2976	74	3050	98%

	STATUS FAIL	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field unit_concept_id in the OBSERVATION table. (Threshold=5%).	100.00%
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the source concept field dose_unit_concept_id in the DRUG_EXPOSURE table. (Threshold=10%).	100.00%
⊞	FAIL	Validation	Plausibility	Atemporal	CONCEPT	For a CONCEPT_ID 30969 (Testicular hyperfunction), the number and percent of records associated with patients with an implausible gender (correct gender = Male). (Threshold=5%).	98.26%
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field observation_concept_id in the OBSERVATION table. (Threshold=5%).	11.46%
⊞	FAIL	Validation	Plausibility	Atemporal	CONCEPT	For a CONCEPT_ID 200670 (Benign neoplasm of male genital organ), the number and percent of records associated with patients with an implausible gender (correct gender = Male). (Threshold=5%).	5.29%

Showing 1 to 5 of 74 entries (filtered from 3,050 total entries)

Data Quality Dashboard (DQD)

- Runs a prespecified set of data quality checks and thresholds on the CDM



DATA QUALITY ASSESSMENT

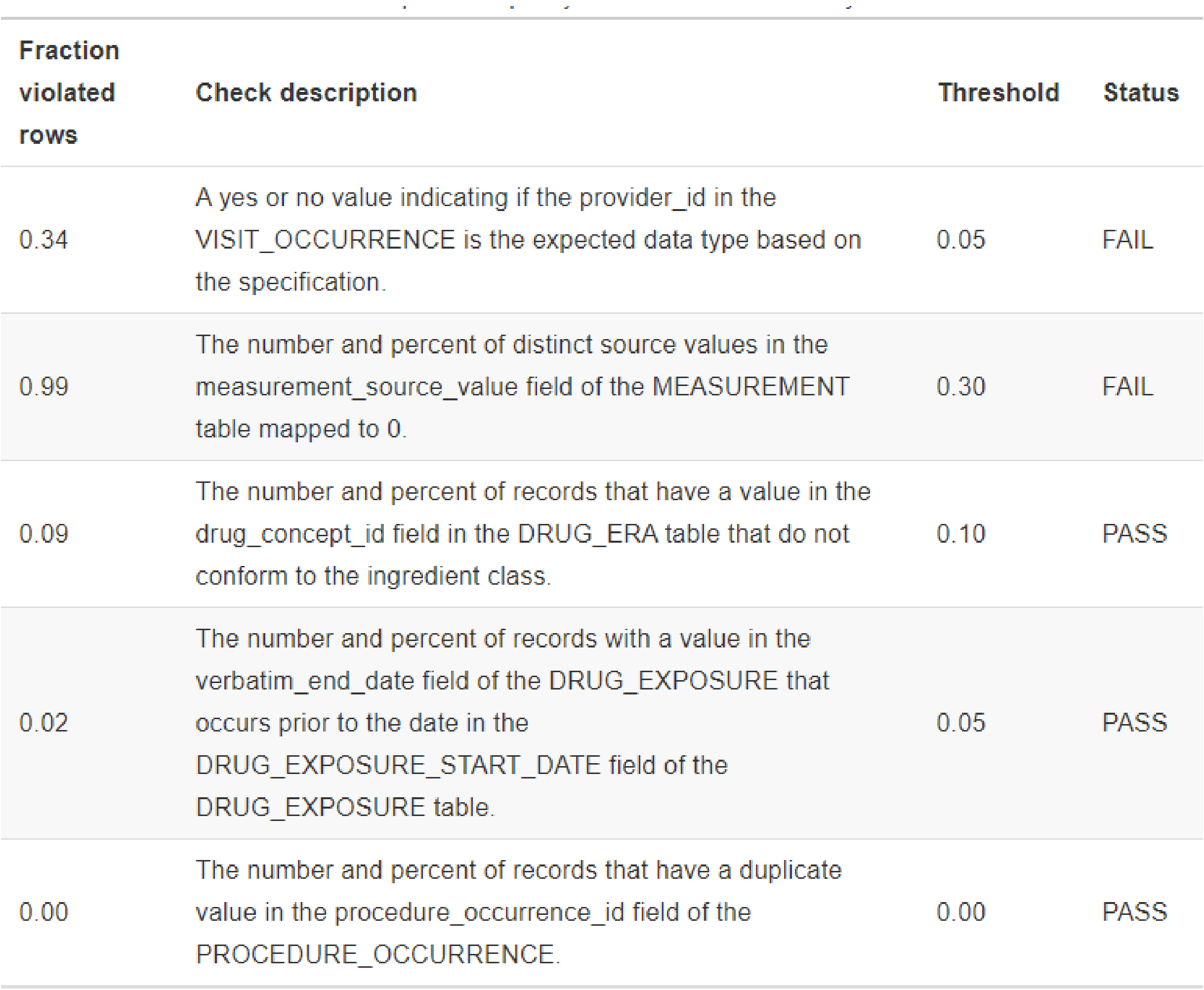
SYNTHEA SYNTHETIC HEALTH DATABASE

Results generated at 2019-08-22 14:15:06 in 29 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	**21**	180	88%	283	0	283	100%	442	**21**	463	95%
Conformance	637	**34**	671	95%	104	0	104	100%	741	**34**	775	96%
Completeness	369	**17**	386	96%	5	**10**	15	33%	374	**27**	401	93%
Total	1165	**72**	1237	94%	392	**10**	402	98%	1557	**82**	1639	**95%**

Data Quality Dashboard (DQD)

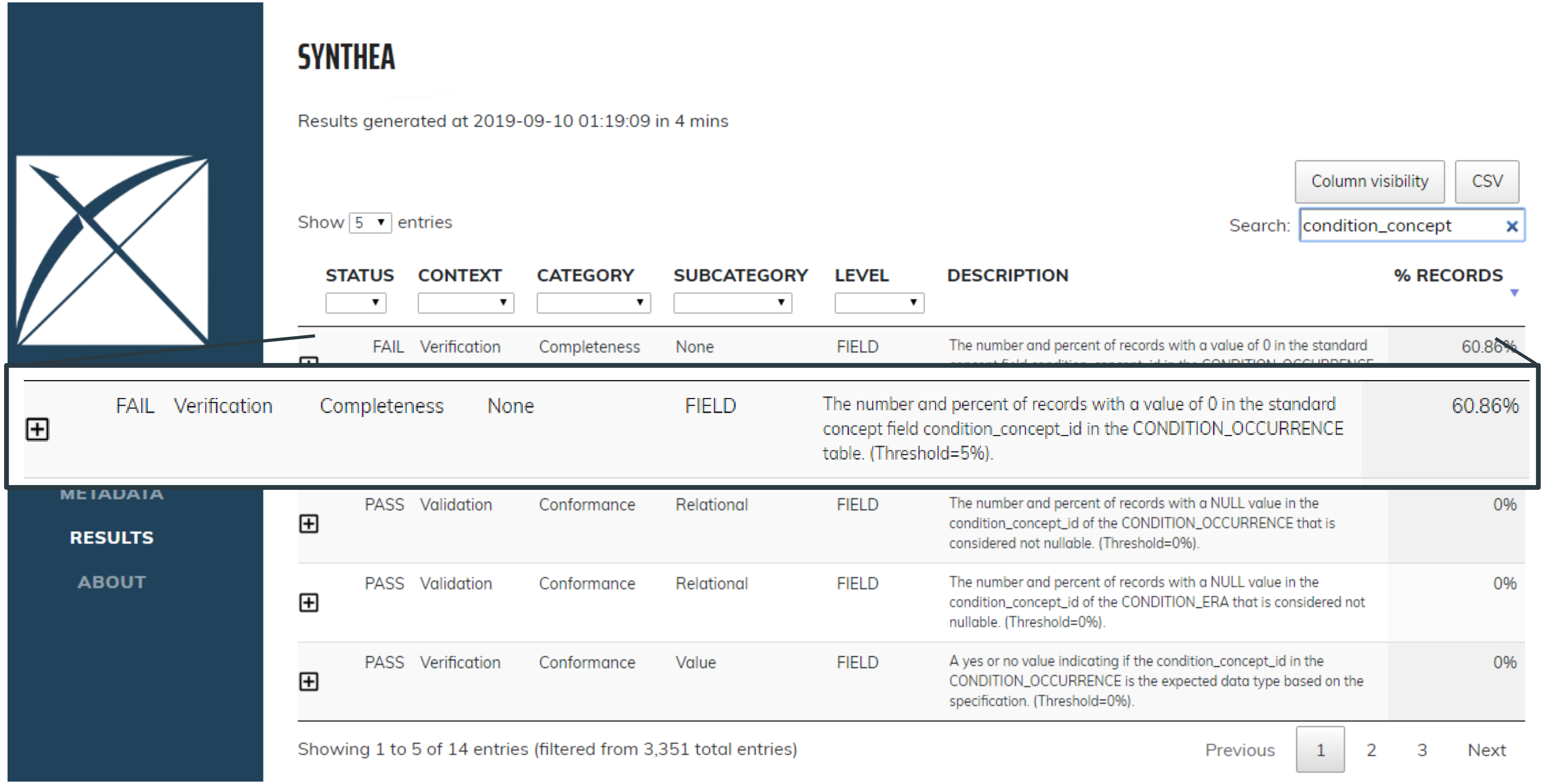
- DQD Example Rules



Fraction violated rows	Check description	Threshold	Status
0.34	A yes or no value indicating if the provider_id in the VISIT_OCCURRENCE is the expected data type based on the specification.	0.05	FAIL
0.99	The number and percent of distinct source values in the measurement_source_value field of the MEASUREMENT table mapped to 0.	0.30	FAIL
0.09	The number and percent of records that have a value in the drug_concept_id field in the DRUG_ERA table that do not conform to the ingredient class.	0.10	PASS
0.02	The number and percent of records with a value in the verbatim_end_date field of the DRUG_EXPOSURE that occurs prior to the date in the DRUG_EXPOSURE_START_DATE field of the DRUG_EXPOSURE table.	0.05	PASS
0.00	The number and percent of records that have a duplicate value in the procedure_occurrence_id field of the PROCEDURE_OCCURRENCE.	0.00	PASS

DQD – Issues in our data?

- Did DQD notice anything?



SYNTHEA

Results generated at 2019-09-10 01:19:09 in 4 mins

Show 5 entries

Search: condition_concept

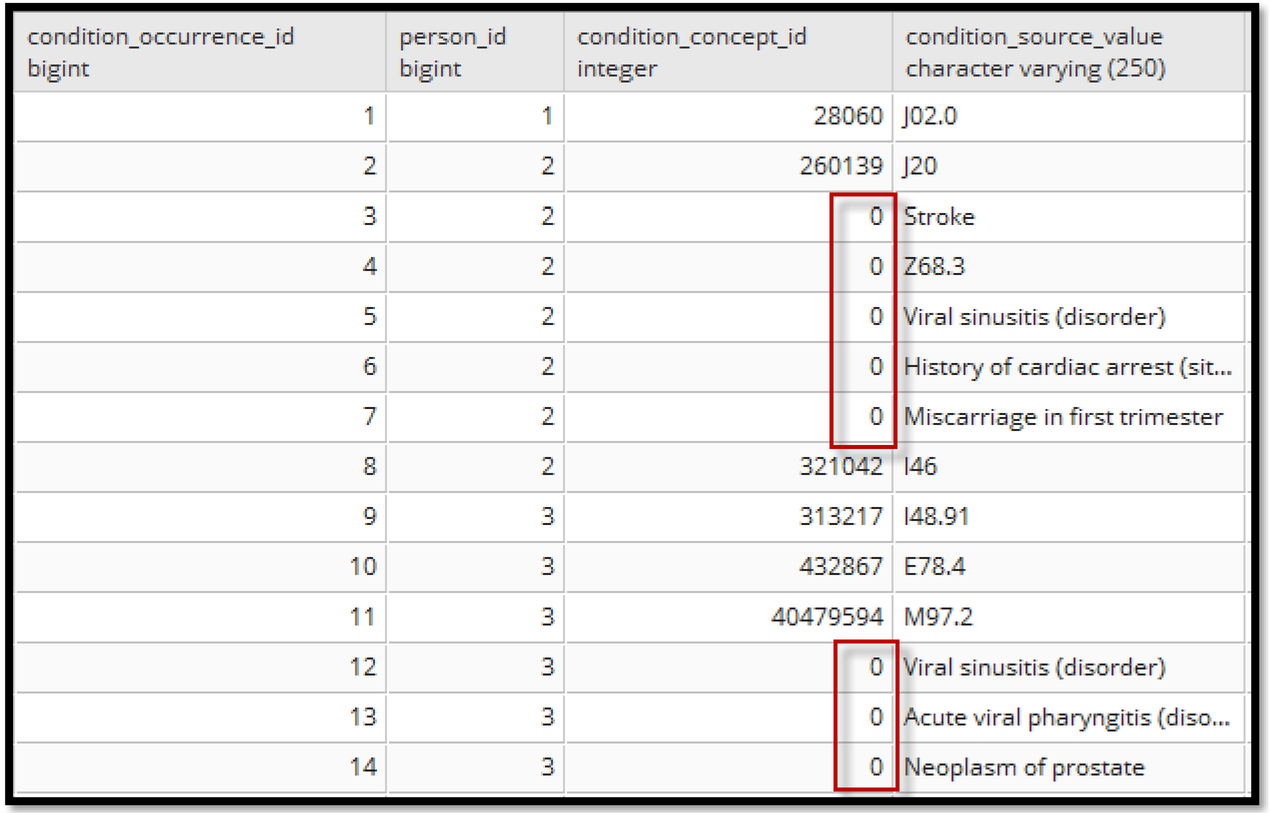
	STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field condition_concept_id in the CONDITION_OCCURRENCE table. (Threshold=5%).	60.86%
⊞	PASS	Validation	Conformance	Relational	FIELD	The number and percent of records with a NULL value in the condition_concept_id of the CONDITION_OCCURRENCE that is considered not nullable. (Threshold=0%).	0%
⊞	PASS	Validation	Conformance	Relational	FIELD	The number and percent of records with a NULL value in the condition_concept_id of the CONDITION_ERA that is considered not nullable. (Threshold=0%).	0%
⊞	PASS	Verification	Conformance	Value	FIELD	A yes or no value indicating if the condition_concept_id in the CONDITION_OCCURRENCE is the expected data type based on the specification. (Threshold=0%).	0%

Showing 1 to 5 of 14 entries (filtered from 3,351 total entries)

Previous 1 2 3 Next

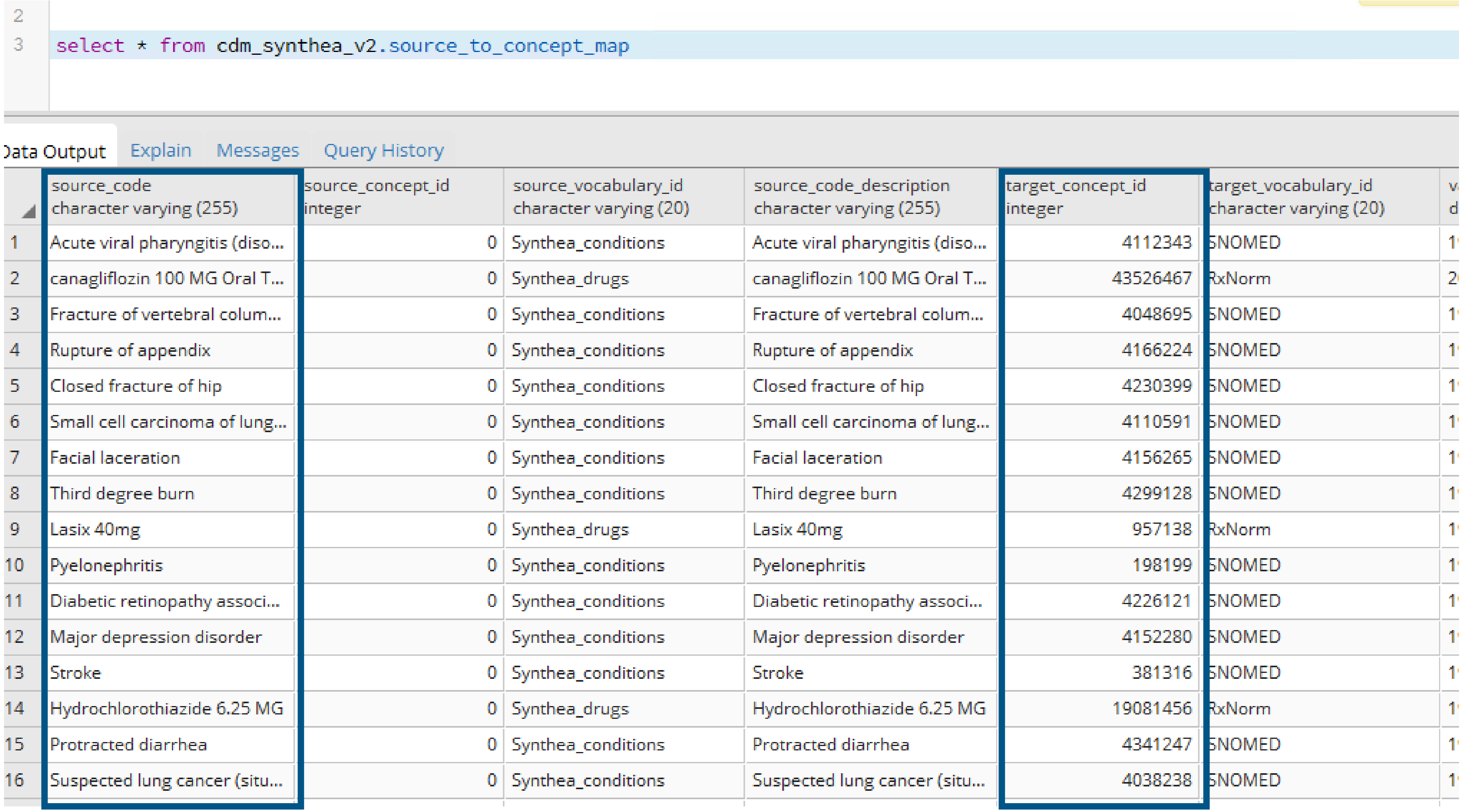
DQD – Maybe we have a bug?

- In the CONDITION_OCCURRENCE, 61% rows are mapped to 0



condition_occurrence_id bigint	person_id bigint	condition_concept_id integer	condition_source_value character varying (250)
1	1	28060	J02.0
2	2	260139	J20
3	2	0	Stroke
4	2	0	Z68.3
5	2	0	Viral sinusitis (disorder)
6	2	0	History of cardiac arrest (sit...
7	2	0	Miscarriage in first trimester
8	2	321042	I46
9	3	313217	I48.91
10	3	432867	E78.4
11	3	40479594	M97.2
12	3	0	Viral sinusitis (disorder)
13	3	0	Acute viral pharyngitis (diso...
14	3	0	Neoplasm of prostate

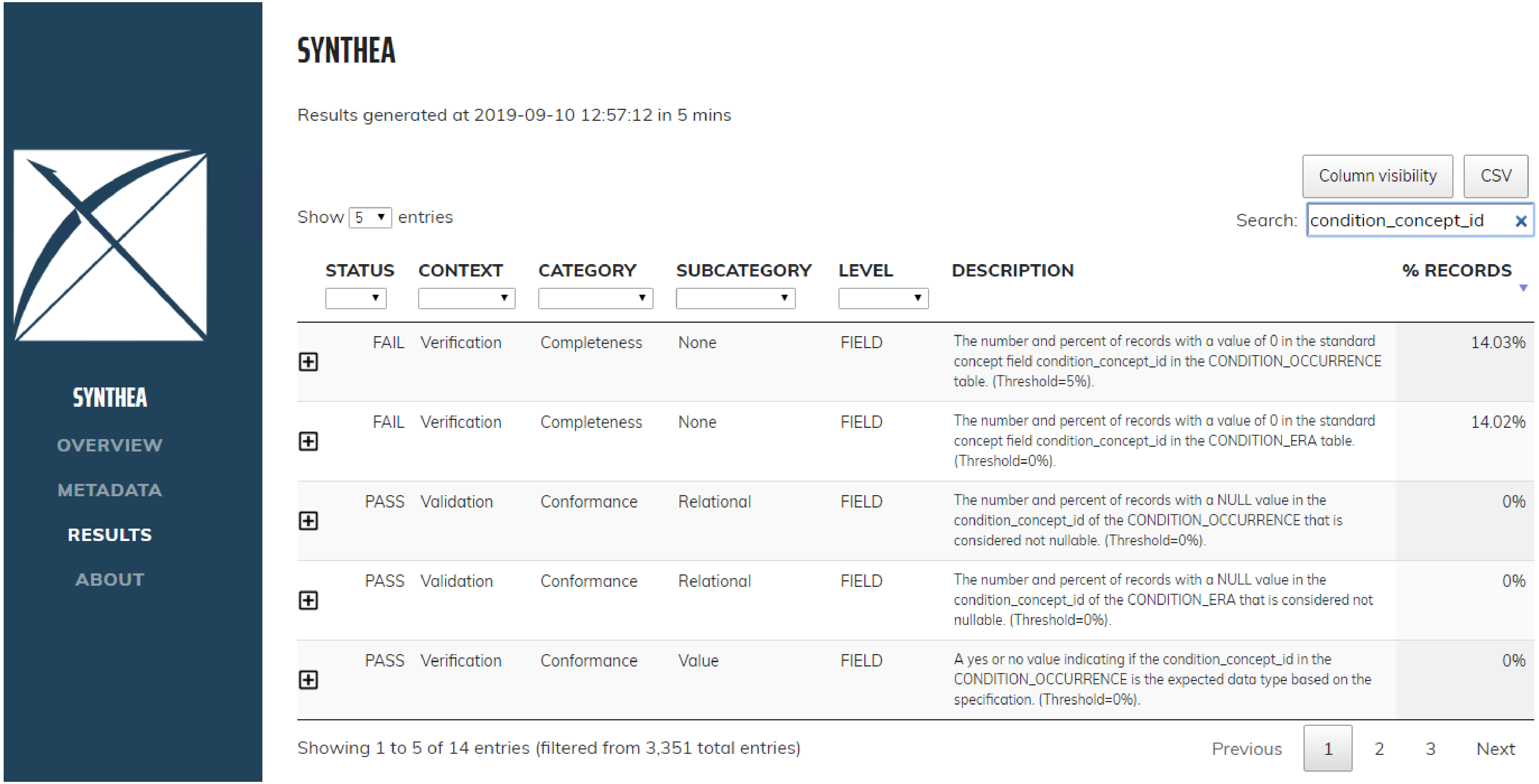
DQD – Vocabulary to fix the problem



```
select * from cdm_synthea_v2.source_to_concept_map
```

	source_code	source_concept_id	source_vocabulary_id	source_code_description	target_concept_id	target_vocabulary_id	v
	character varying (255)	integer	character varying (20)	character varying (255)	integer	character varying (20)	d
1	Acute viral pharyngitis (diso...	0	Synthea_conditions	Acute viral pharyngitis (diso...	4112343	SNOMED	1
2	canagliflozin 100 MG Oral T...	0	Synthea_drugs	canagliflozin 100 MG Oral T...	43526467	RxNorm	2
3	Fracture of vertebral colum...	0	Synthea_conditions	Fracture of vertebral colum...	4048695	SNOMED	1
4	Rupture of appendix	0	Synthea_conditions	Rupture of appendix	4166224	SNOMED	1
5	Closed fracture of hip	0	Synthea_conditions	Closed fracture of hip	4230399	SNOMED	1
6	Small cell carcinoma of lung...	0	Synthea_conditions	Small cell carcinoma of lung...	4110591	SNOMED	1
7	Facial laceration	0	Synthea_conditions	Facial laceration	4156265	SNOMED	1
8	Third degree burn	0	Synthea_conditions	Third degree burn	4299128	SNOMED	1
9	Lasix 40mg	0	Synthea_drugs	Lasix 40mg	957138	RxNorm	1
10	Pyelonephritis	0	Synthea_conditions	Pyelonephritis	198199	SNOMED	1
11	Diabetic retinopathy associ...	0	Synthea_conditions	Diabetic retinopathy associ...	4226121	SNOMED	1
12	Major depression disorder	0	Synthea_conditions	Major depression disorder	4152280	SNOMED	1
13	Stroke	0	Synthea_conditions	Stroke	381316	SNOMED	1
14	Hydrochlorothiazide 6.25 MG	0	Synthea_drugs	Hydrochlorothiazide 6.25 MG	19081456	RxNorm	1
15	Protracted diarrhea	0	Synthea_conditions	Protracted diarrhea	4341247	SNOMED	1
16	Suspected lung cancer (situ...	0	Synthea_conditions	Suspected lung cancer (situ...	4038238	SNOMED	1

DQD – Re-run the DQD



SYNTHEA

OVERVIEW

METADATA

RESULTS

ABOUT

SYNTHEA

Results generated at 2019-09-10 12:57:12 in 5 mins

	STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field condition_concept_id in the CONDITION_OCCURRENCE table. (Threshold=5%).	14.03%
⊞	FAIL	Verification	Completeness	None	FIELD	The number and percent of records with a value of 0 in the standard concept field condition_concept_id in the CONDITION_ERA table. (Threshold=0%).	14.02%
⊞	PASS	Validation	Conformance	Relational	FIELD	The number and percent of records with a NULL value in the condition_concept_id of the CONDITION_OCCURRENCE that is considered not nullable. (Threshold=0%).	0%
⊞	PASS	Validation	Conformance	Relational	FIELD	The number and percent of records with a NULL value in the condition_concept_id of the CONDITION_ERA that is considered not nullable. (Threshold=0%).	0%
⊞	PASS	Verification	Conformance	Value	FIELD	A yes or no value indicating if the condition_concept_id in the CONDITION_OCCURRENCE is the expected data type based on the specification. (Threshold=0%).	0%

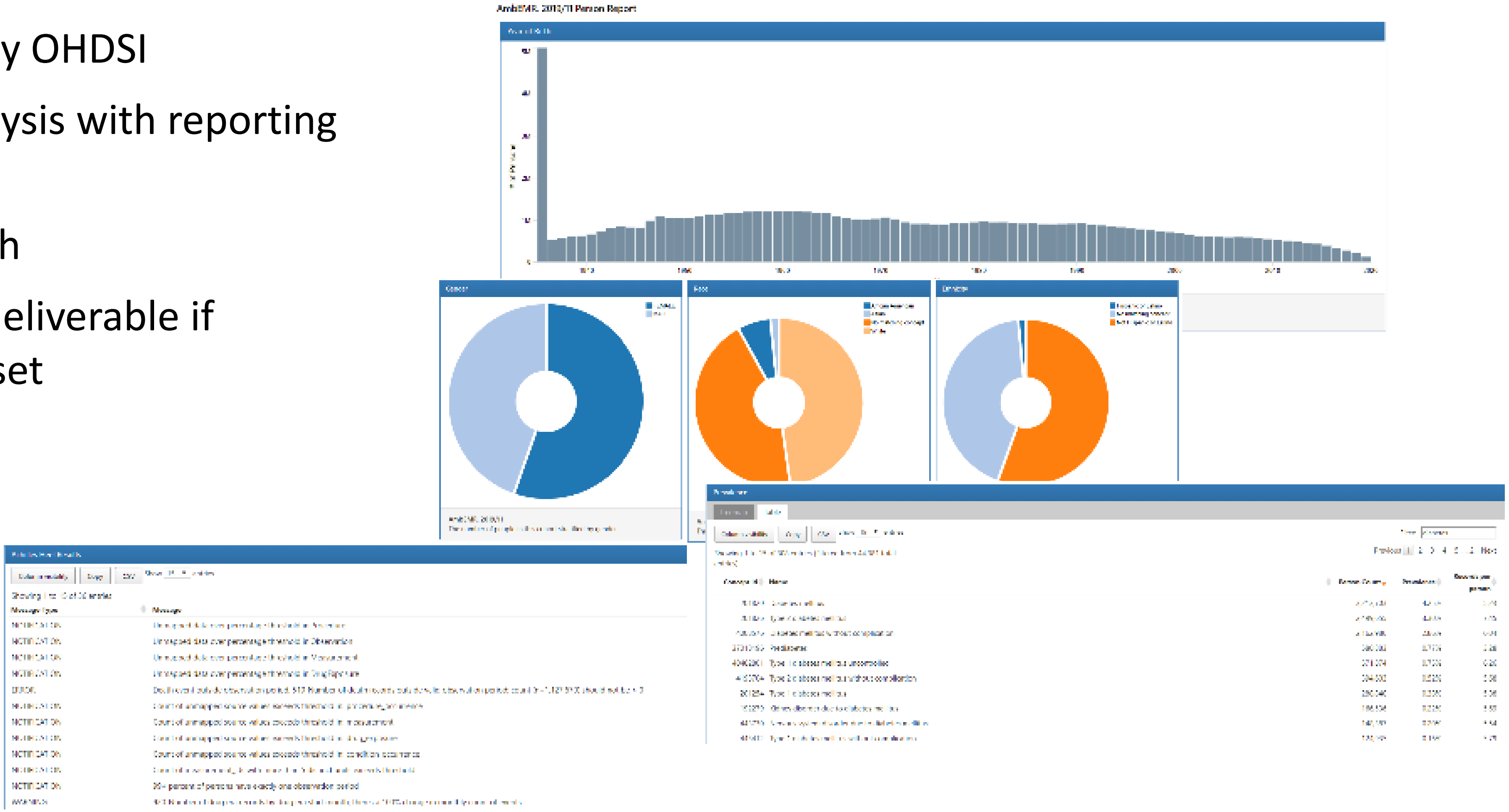
Showing 1 to 5 of 14 entries (filtered from 3,351 total entries)

Achilles

Description

- Created and maintained by OHDSI
- Descriptive statistical analysis with reporting and data quality checks
- Executed with each refresh
- Sent to clients as part of deliverable if purchased OMOP data asset

Deliverable



AmbEMR 2013/11 Person Report

Message Type | Message

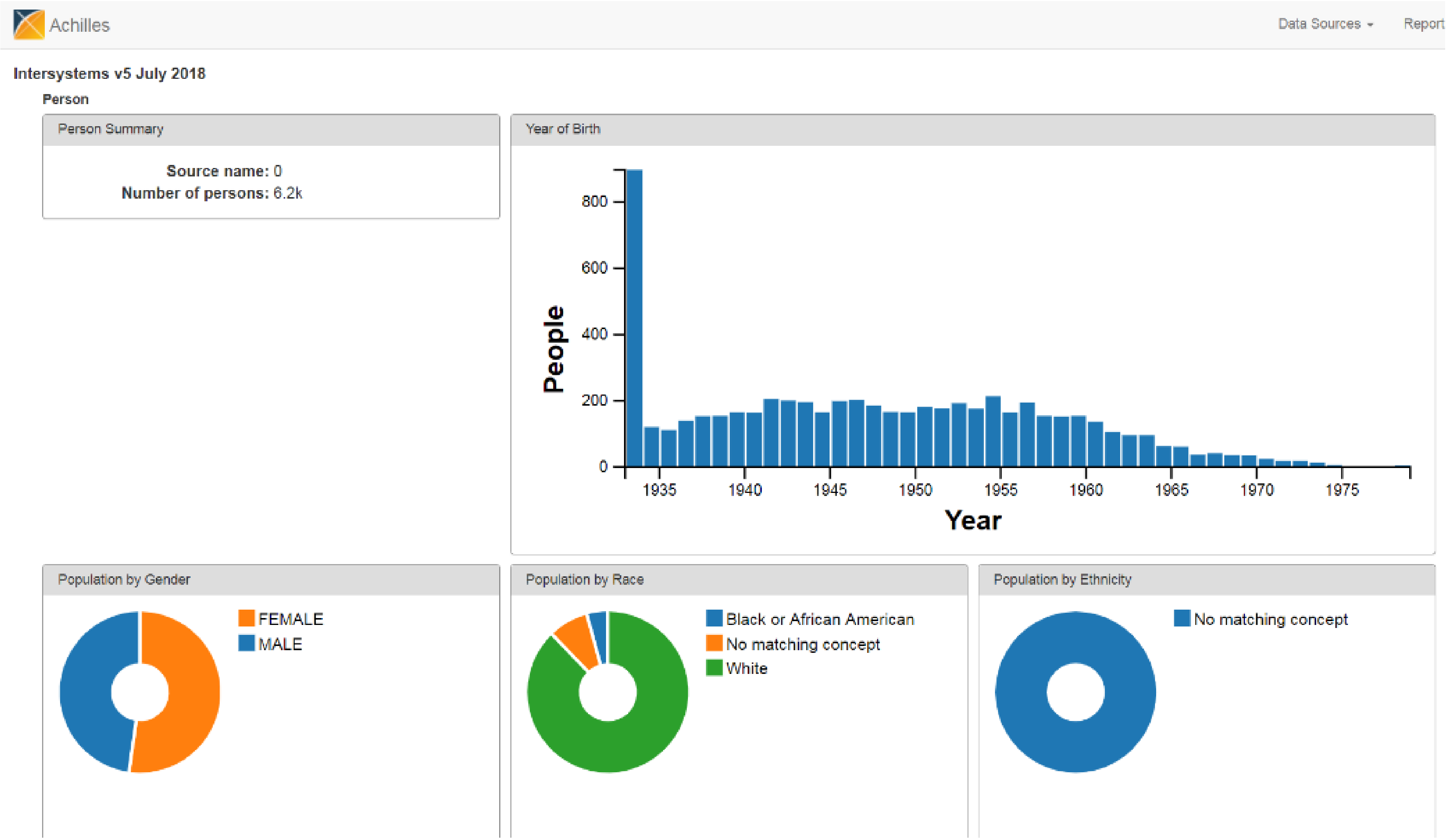
NOTIFICATION | Unmapped data over percentage threshold in Observation

NOTIFICATION | Count of unmapped source values exceeds threshold in measurement

NOTIFICATION | Count of unmapped source values exceeds threshold in condition occurrence

NOTIFICATION | 20+ percent of persons have exactly one observation period

Achilles



Intersystems v5 July 2018

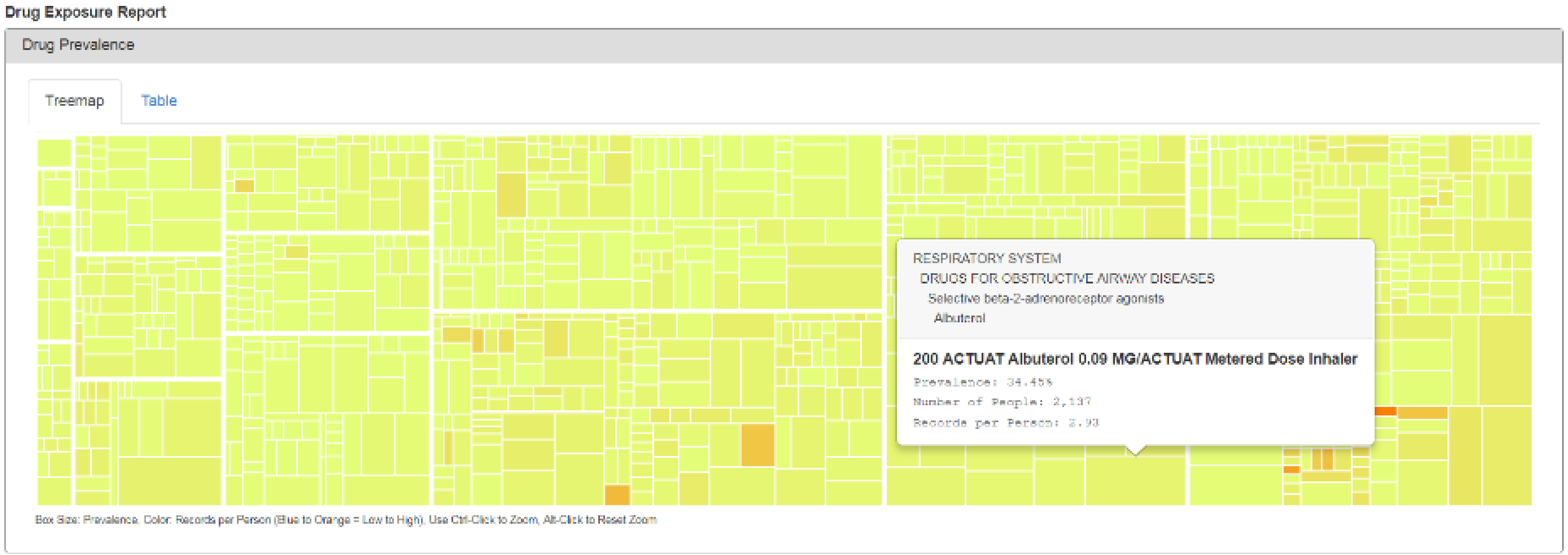
Person

Person Summary

Source name: 0
Number of persons: 6.2k

Achilles – Reports by Domain

- Heat map report

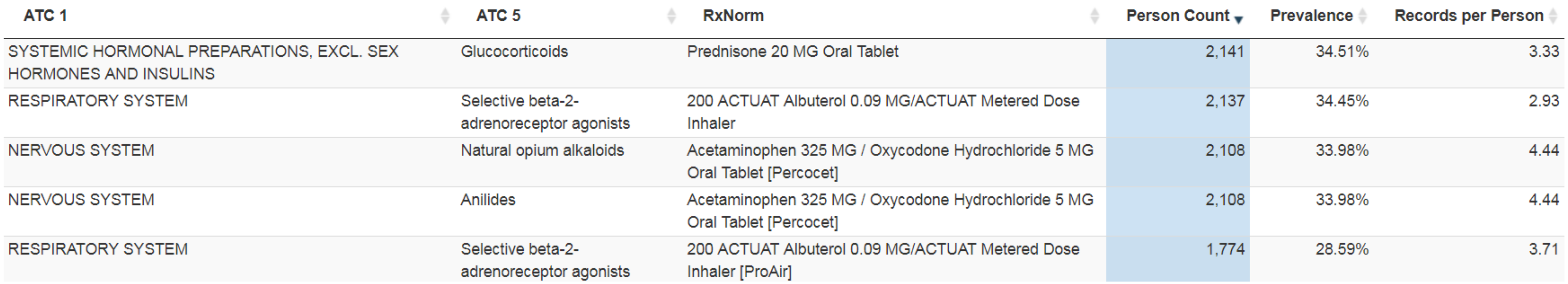


RESPIRATORY SYSTEM
DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES
Selective beta-2-adrenoreceptor agonists
Albuterol

200 ACTUAT Albuterol 0.09 MG/ACTUAT Metered Dose Inhaler
Prevalence: 34.45%
Number of People: 2,137
Records per Person: 2.93

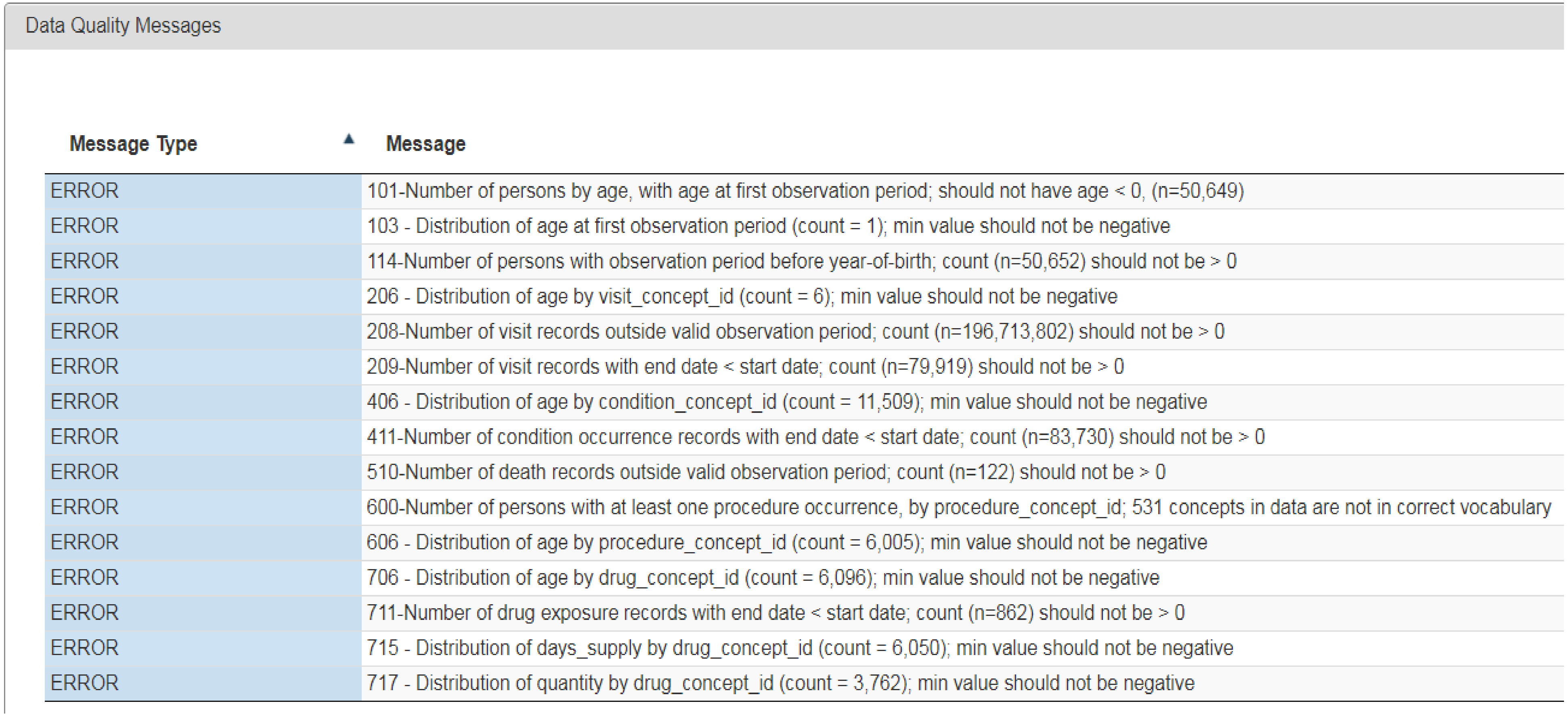
Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

- Tabular Report



ATC 1	ATC 5	RxNorm	Person Count	Prevalence	Records per Person
SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS	Glucocorticoids	Prednisone 20 MG Oral Tablet	2,141	34.51%	3.33
RESPIRATORY SYSTEM	Selective beta-2-adrenoreceptor agonists	200 ACTUAT Albuterol 0.09 MG/ACTUAT Metered Dose Inhaler	2,137	34.45%	2.93
NERVOUS SYSTEM	Natural opium alkaloids	Acetaminophen 325 MG / Oxycodone Hydrochloride 5 MG Oral Tablet [Percocet]	2,108	33.98%	4.44
NERVOUS SYSTEM	Anilides	Acetaminophen 325 MG / Oxycodone Hydrochloride 5 MG Oral Tablet [Percocet]	2,108	33.98%	4.44
RESPIRATORY SYSTEM	Selective beta-2-adrenoreceptor agonists	200 ACTUAT Albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [ProAir]	1,774	28.59%	3.71

Achilles – Achilles Heel Report



Data Quality Messages

Message Type ▲	Message
ERROR	101-Number of persons by age, with age at first observation period; should not have age < 0, (n=50,649)
ERROR	103 - Distribution of age at first observation period (count = 1); min value should not be negative
ERROR	114-Number of persons with observation period before year-of-birth; count (n=50,652) should not be > 0
ERROR	206 - Distribution of age by visit_concept_id (count = 6); min value should not be negative
ERROR	208-Number of visit records outside valid observation period; count (n=196,713,802) should not be > 0
ERROR	209-Number of visit records with end date < start date; count (n=79,919) should not be > 0
ERROR	406 - Distribution of age by condition_concept_id (count = 11,509); min value should not be negative
ERROR	411-Number of condition occurrence records with end date < start date; count (n=83,730) should not be > 0
ERROR	510-Number of death records outside valid observation period; count (n=122) should not be > 0
ERROR	600-Number of persons with at least one procedure occurrence, by procedure_concept_id; 531 concepts in data are not in correct vocabulary
ERROR	606 - Distribution of age by procedure_concept_id (count = 6,005); min value should not be negative
ERROR	706 - Distribution of age by drug_concept_id (count = 6,096); min value should not be negative
ERROR	711-Number of drug exposure records with end date < start date; count (n=862) should not be > 0
ERROR	715 - Distribution of days_supply by drug_concept_id (count = 6,050); min value should not be negative
ERROR	717 - Distribution of quantity by drug_concept_id (count = 3,762); min value should not be negative

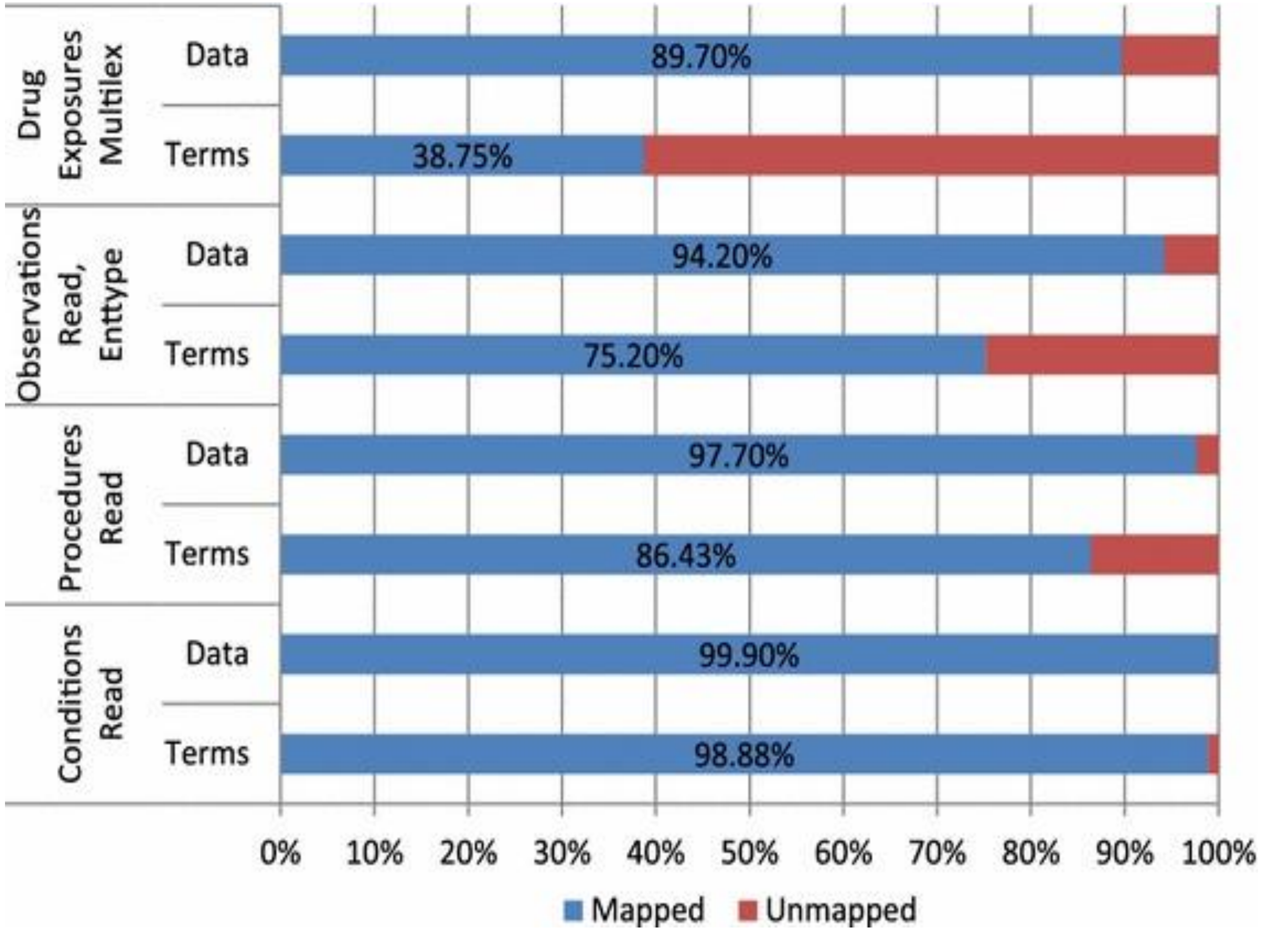
80/20 Rule

Conclusions

Raw data can be accurately transformed into the OMOP CDM with acceptable information loss across domains. CDM structure was adequate and vocabulary mappings were assessed to be high quality.

Lessons Learned

ETL helps standardize source data to research quality. The goal is to accurately transform the data into CDM format and standardized terms with acceptable information loss, and high-frequency source codes are mapped.



*Cited from “Fidelity assessment of a clinical practice research datalink conversion to the OMOP CDM model”

Data Quality Checks – Exercise

OHDSI

Observational Health Data Sciences and Informatics

Thank You!