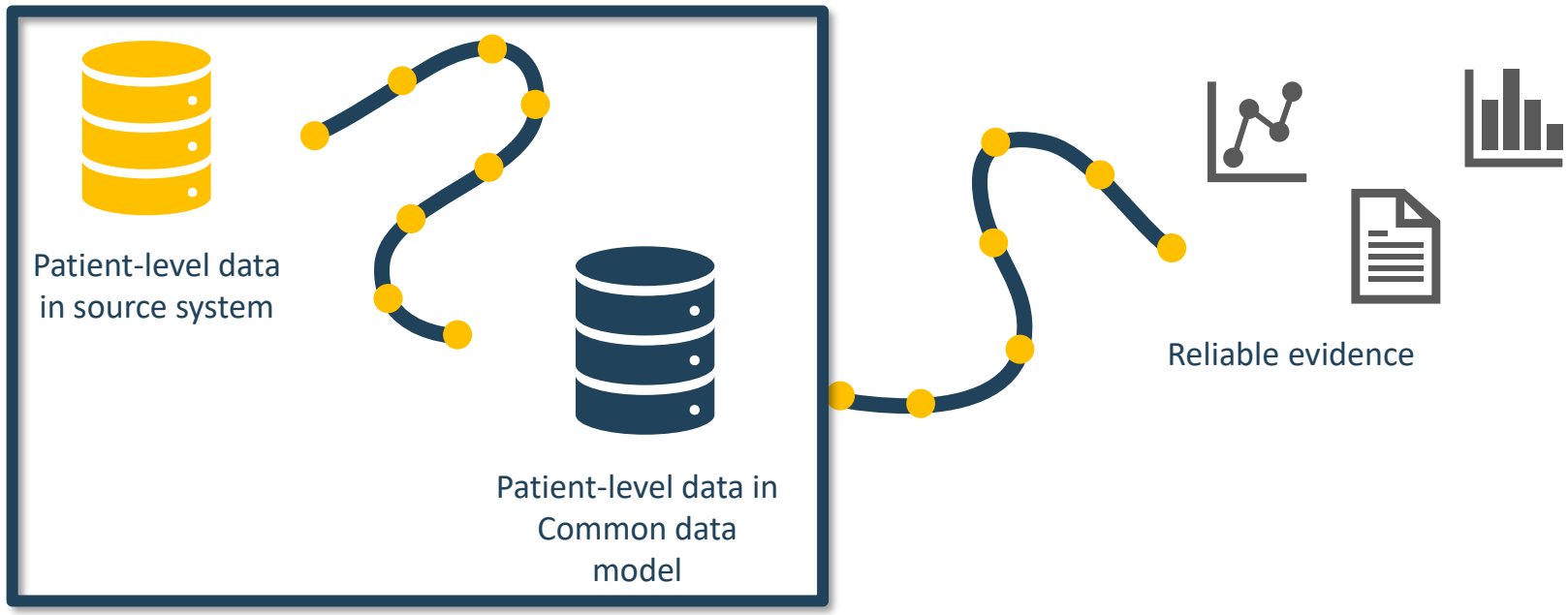
OMOP Conversion Process

ETL

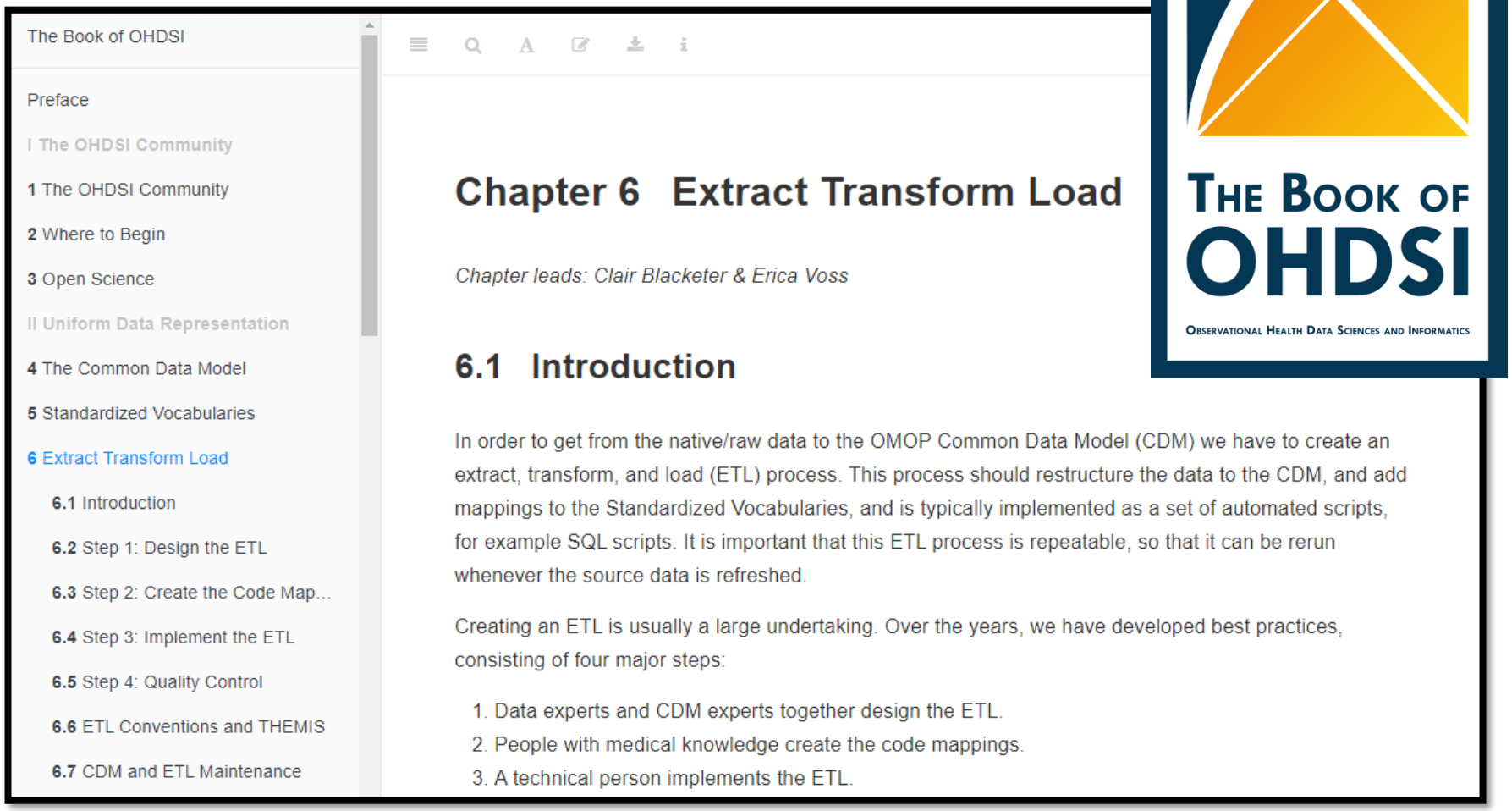
- Extract Transform Load

- In order to get from our native/raw data into the OMOP CDM we need to design and develop and ETL process



- Goal in ETLing is to standardize the format and terminology

ETL Process



https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html

ETL Process

Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

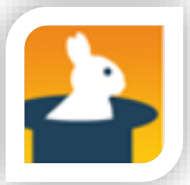
ETL Documentation

All are involved in quality control

A technical person implements the ETL

OHDSI Tools

White Rabbit



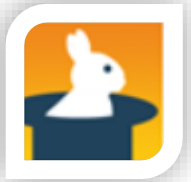
Rabbit In a Hat

Usagi

White Rabbit

ACHILLES

DQD



Rabbit In a Hat

Designing the ETL

Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

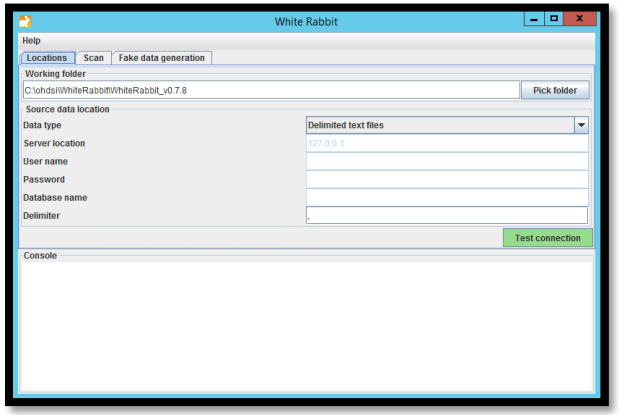
ETL Documentation

All are involved in quality control

A technical person implements the ETL

White Rabbit

- White Rabbit scans source data & creates a csv report on the source data



- The scan can be used to:
  - Learn about your source data
  - Help design the ETL
  - Used by Rabbit In a Hat

WR Output – ScanReport.xlsx

Table/Field Overview

Table	Field	Description	Type	Max length	N rows
pop	der_sex		character	1	16374539
pop	der_yob		double pre	6	16374539
pop	pat_id		character	64	16374539
pop	pat_hash_id		character	16	16374539
pop	pmtx_flag		numeric	1	16374539
pop	anon_ims_pat_id		character	11	16374539
pop	pat_region		character	2	16374539
pop	pat_state		character	2	16374539
pop	pat_zip3		character	3	16374539
pop	grp_indv_cd		character	1	16374539
pop	mh_cd		character	1	16374539
pop	enr_rel		character	2	16374539
pop	temp_col1		character	0	16374539
pop	temp_col2		character	0	16374539
pop	load_row_id		bigint	9	16374539
claims_diag_lk	person_source_valu		character	64	2992046684
claims_diag_lk	event_start_date		date	10	2992046684

Value counts

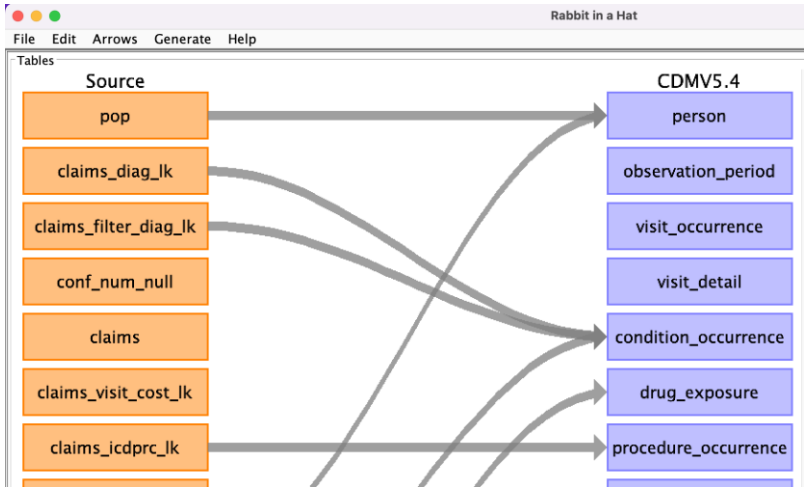
	A	B	C	D
1	der_sex	Frequency	der_yob	Frequency
2	F	50479	1991.0	2030
3	M	49514	1992.0	1970
4	U	7	1990.0	1947
5			1989.0	1908
6			1988.0	1873
7			1994.0	1872
8			1995.0	1806
9			1993.0	1805
10			1996.0	1716
11			1986.0	1676
12			1987.0	1643
13			1985.0	1633
14			1983.0	1588
15			1981.0	1581
16			1984.0	1576
17			1970.0	1555
18			1980.0	1553

pop | claims_diag_lk | claims

Rabbit in a Hat

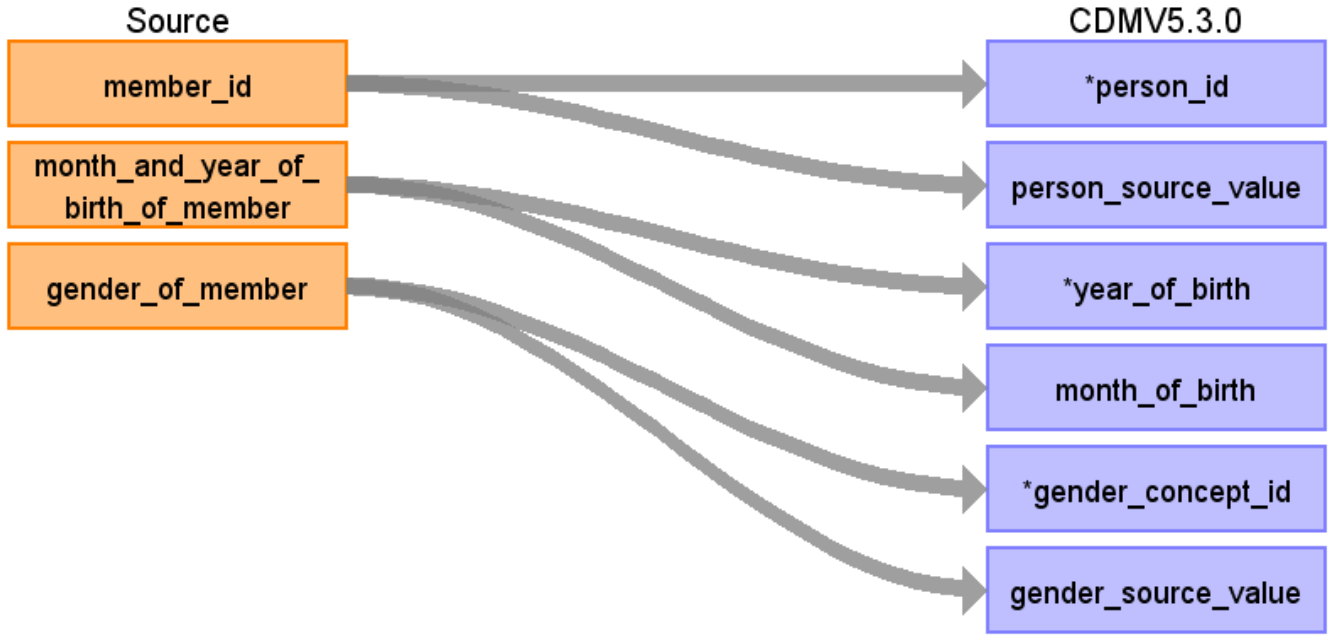
- Read and display a White Rabbit scan document

- Provides a graphical interface to allow a user to connect source data to CDM tables



RIAH – Column Mapping Example

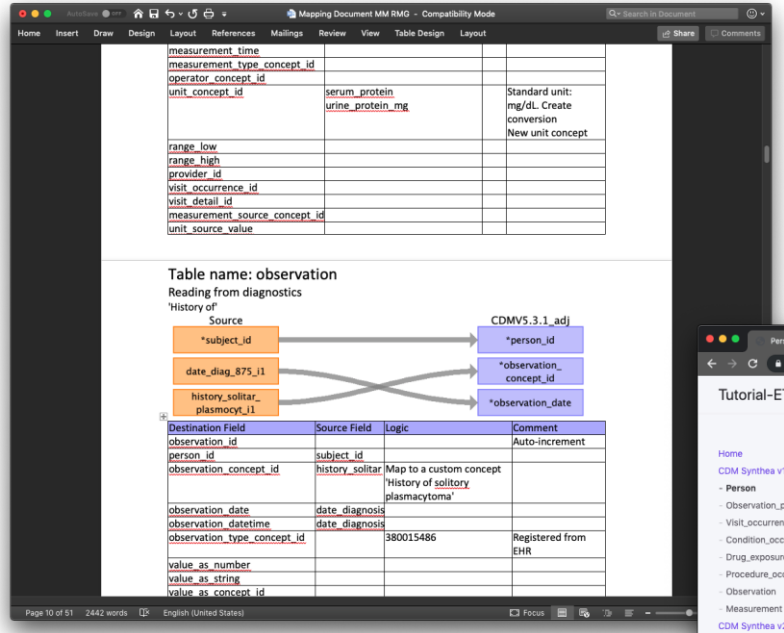
Reading from Enrollment



Destination Field	Source field	Logic
YEAR_OF_BIRTH	month_and_year_of_birth_ of_member	Take first 4 digits
MONTH_OF_BIRTH	month_and_year_of_birth_ of_member	Take last 2 digits (01 is January)

RiaH - Output

Word document



measurement_time				
measurement_type_concept_id				
operator_concept_id				
unit_concept_id	serum_protein urine_protein_mg			Standard unit: mg/dL. Create conversion New unit concept
range_low				
range_high				
provider_id				
visit_occurrence_id				
visit_detail_id				
measurement_source_concept_id				
unit_source_value				

Table name: observation

Reading from diagnostics

'History of'

Source → CDMV5.3.1_adj

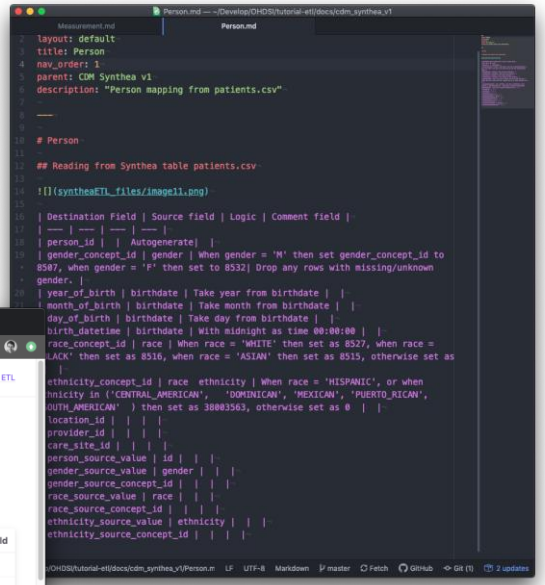
*subject_id → *person_id

date_diag_875_i1 → *observation_concept_id

history_solitar_plasmocyt_i1 → *observation_date

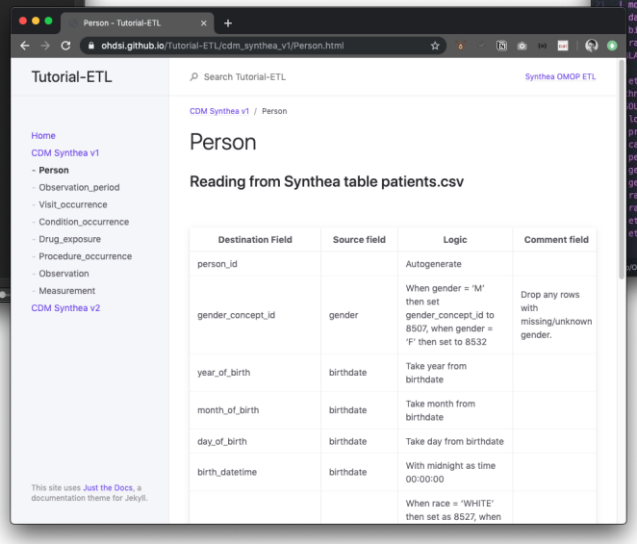
Destination Field	Source Field	Logic	Comment
observation_id			Auto-increment
person_id	subject_id		
observation_concept_id	history_solitar	Map to a custom concept 'History of solitory plasmacytoma'	
observation_date	date_diagnosis		
observation_datetime	date_diagnosis		
observation_type_concept_id		380015486	Registered from EHR
value_as_number			
value_as_string			
value_as_concept_id			

Markdown documents



```
layout: default
title: Person
nav_order: 1
parent: CDM Synthea v1
description: "Person mapping from patients.csv"

---

# Person

## Reading from Synthea table patients.csv

![](syntheaETL_files/image11.png)

| Destination Field | Source field | Logic | Comment field |
| --- | --- | --- | --- |
| person_id | | Autogenerate| |
| gender_concept_id | gender | When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532| Drop any rows with missing/unknown gender. |
| year_of_birth | birthdate | Take year from birthdate | |
| month_of_birth | birthdate | Take month from birthdate | |
| day_of_birth | birthdate | Take day from birthdate | |
| birth_datetime | birthdate | With midnight as time 00:00:00 | |
| race_concept_id | race | When race = 'WHITE' then set as 8527, when race = 'BLACK' then set as 8516, when race = 'ASIAN' then set as 8515, otherwise set as 0 | |
| ethnicity_concept_id | race  ethnicity | When race = 'HISPANIC', or when ethnicity in ('CENTRAL_AMERICAN',  'DOMINICAN', 'MEXICAN', 'PUERTO_RICAN', 'SOUTH_AMERICAN'  ) then set as 38003563, otherwise set as 0 | |
| location_id | | | |
| provider_id | | | |
| care_site_id | | | |
| person_source_value | id | | |
| gender_source_value | gender | | |
| gender_source_concept_id | | | |
| race_source_value | race | | |
| race_source_concept_id | | | |
| ethnicity_source_value | ethnicity | | |
| ethnicity_source_concept_id | | | |
```

Html



Tutorial-ETL

Search Tutorial-ETL

Synthea OMOP ETL

Home

CDM Synthea v1

- Person
  - Observation_period
  - Visit_occurrence
  - Condition_occurrence
  - Drug_exposure
  - Procedure_occurrence
  - Observation
  - Measurement

CDM Synthea v2

CDM Synthea v1 / Person

Person

Reading from Synthea table patients.csv

Destination Field	Source field	Logic	Comment field
person_id		Autogenerate	
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.
year_of_birth	birthdate	Take year from birthdate	
month_of_birth	birthdate	Take month from birthdate	
day_of_birth	birthdate	Take day from birthdate	
birth_datetime	birthdate	With midnight as time 00:00:00	
		When race = 'WHITE' then set as 8527, when	

Vocabulary Mapping

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

Using OMOP Vocabularies

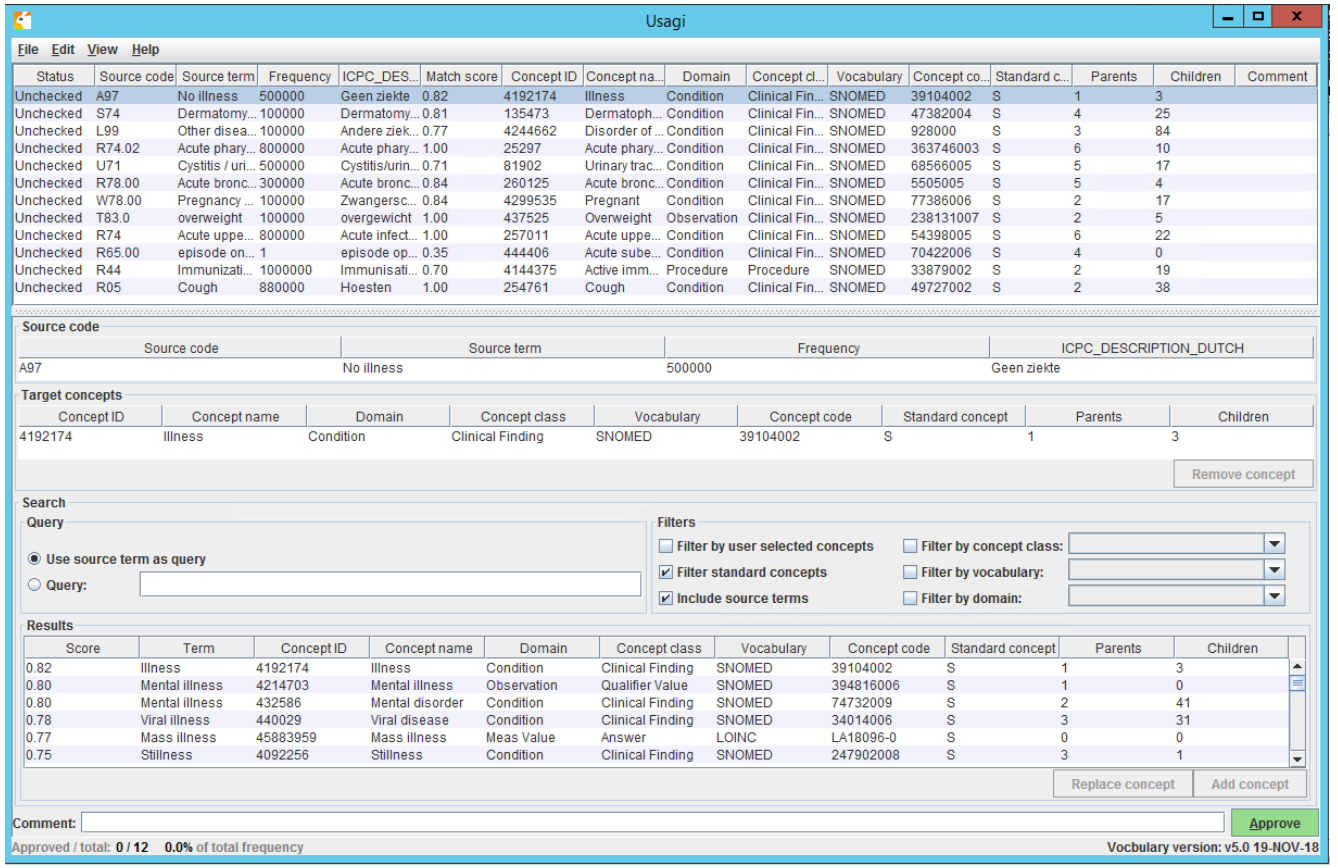
Destination Field	Source field	Logic	Comment field
person_id			
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.

Destination Field	Source field	Logic	Comment field
condition_concept_id	code	Use code to lookup target_concept_id in SOURCE_TO_STANDARD_VOCAB_MAP: select v.target_concept_id from conditions c join source_to_standard_vocab_map v on v.source_code = c.code and v.target_domain_id = 'Condition' and v.target_standard_concept = 'S' and **v.source_vocabulary_id in ('ICD10)**	

Usagi

- When the Vocabulary does not contain your source terms you will need to create a map to OMOP Vocabulary Concepts

- Usagi helps you to:
  - Find best matches, automatically and/or manually
  - Automatic matching based on text similarities (itf/df)
  - Create ‘source to concept map’



Implementing the ETL

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

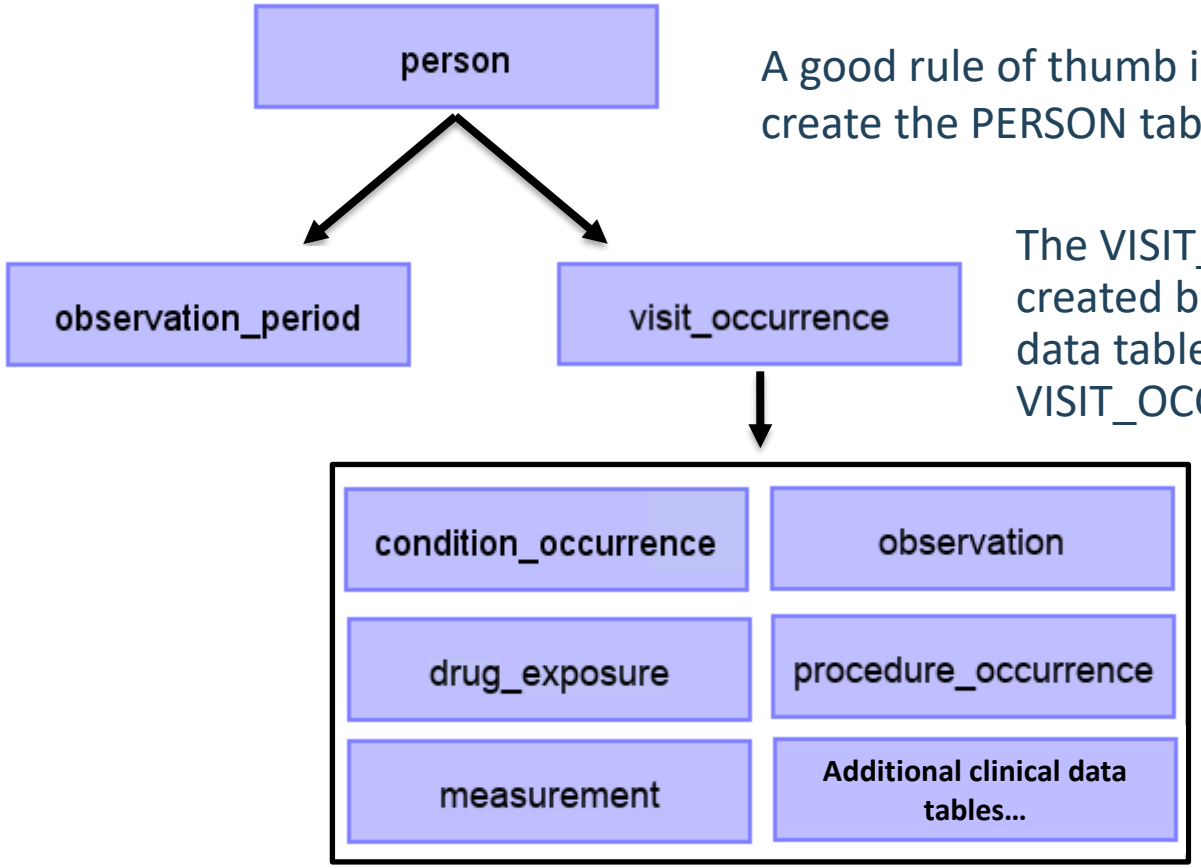
ETL Implementation

There are multiple tools available to implement your ETL

Your choice will largely depend on the size and complexity of the ETL design. And the tools available to you.

ETL Implementation

General Flow of 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

Quality Control

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

Quality

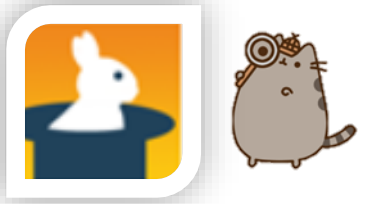
What tools are available to check that the CDM logic was implemented correctly?

Rabbit-in-a-Hat Test Case Framework

Achilles

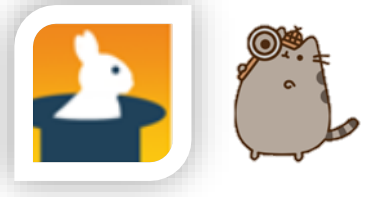
DataQualityDashboard (DQD)

Unit Test Cases



- Testing your CDM builder is important:
  - ETL is often complex, increasing the danger of making mistakes that go unnoticed
  - CDM can update
  - Source data structure/contents can change over time
- Rabbit-In-a-Hat can construct unit tests, or small pieces of code that can automatically check single aspects of the ETL design

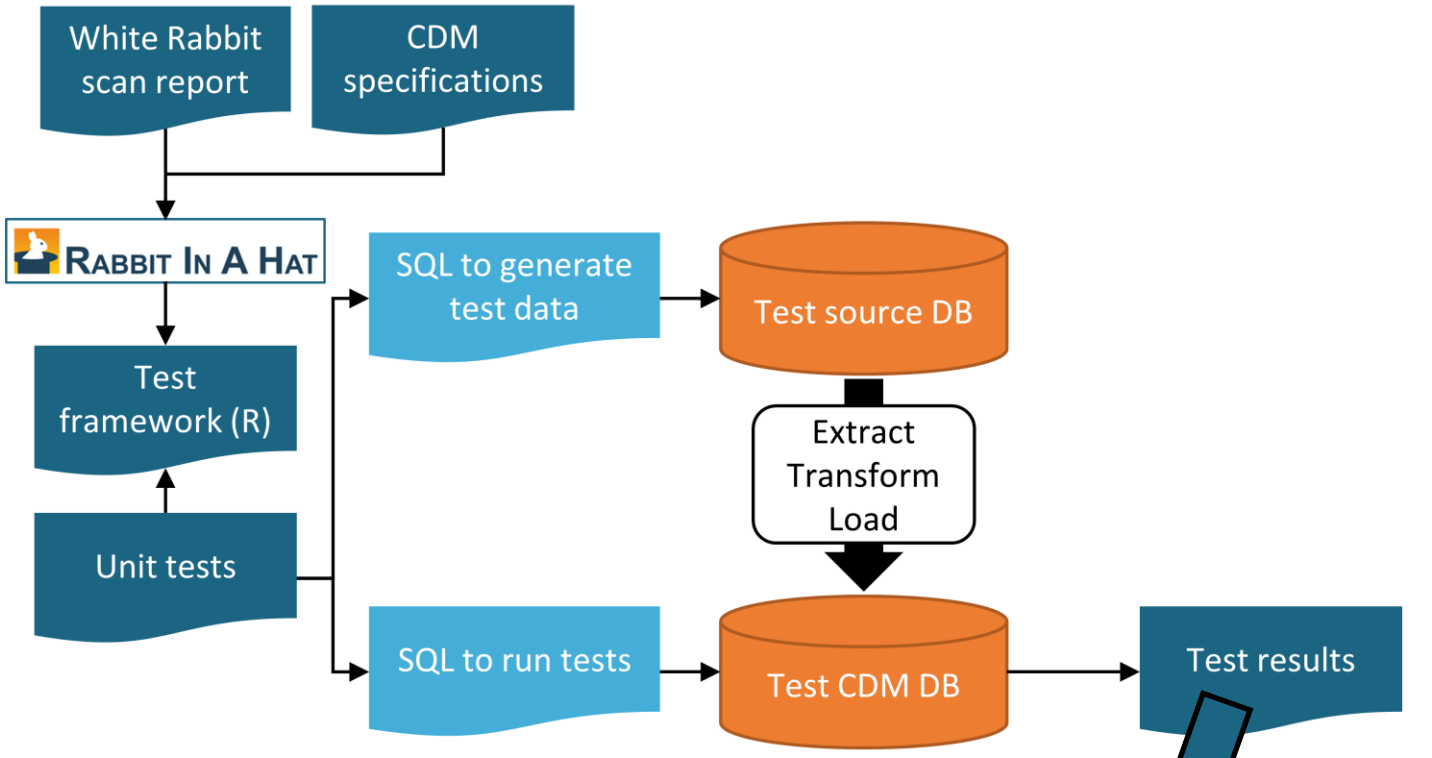
Unit Test Cases



The test framework creates a series of R functions that enables you to specify your ‘fake’ people and records in the same structure as your source data using the scan report as a guide.

```
source("Framework.R")
declareTest(101, "Person gender mappings")
add_enrollment(member_id = "M000000102", gender_of_member = "male")
add_enrollment(member_id = "M000000103", gender_of_member = "female")
expect_person(PERSON_ID = 102, GENDER_CONCEPT_ID = 8507
expect_person(PERSON_ID = 103, GENDER_CONCEPT_ID = 8532)
```

Unit Test Cases



ID	Description	Status
101	Person gender mappings	PASS
101	Person gender mappings	PASS

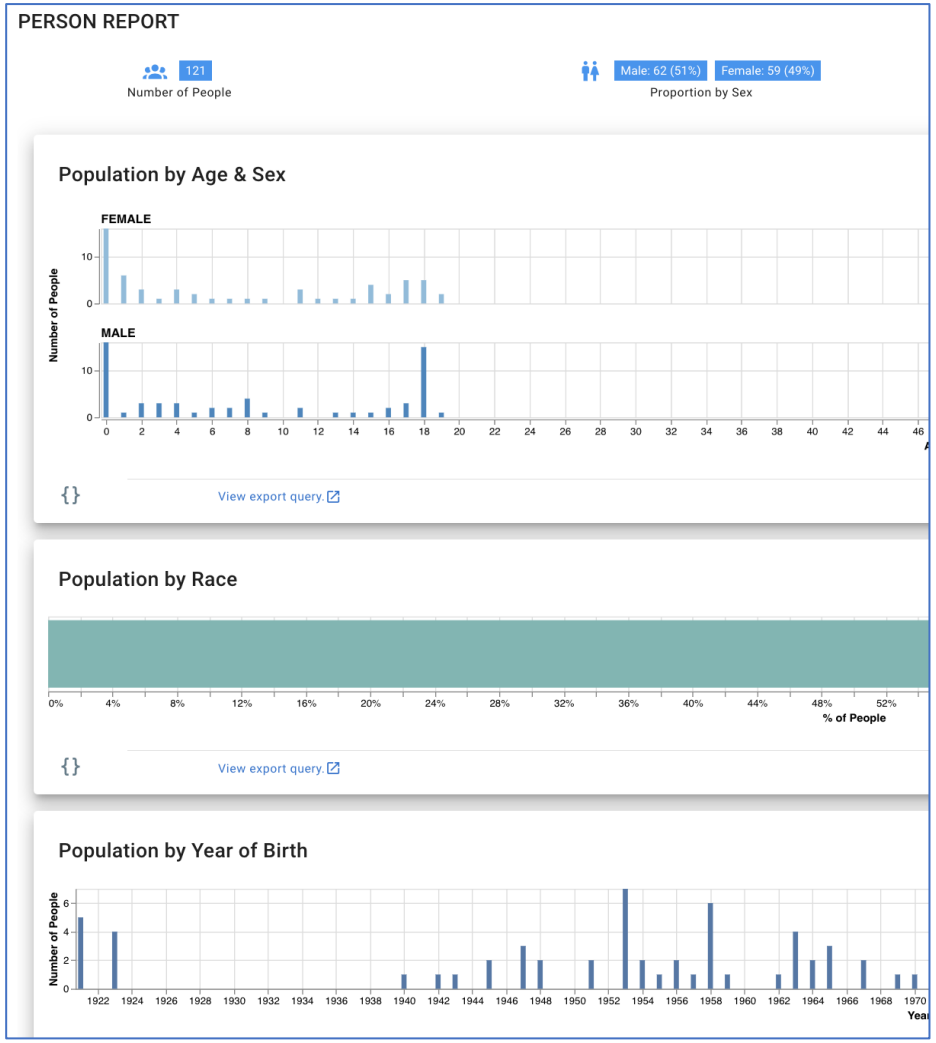
Achilles

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

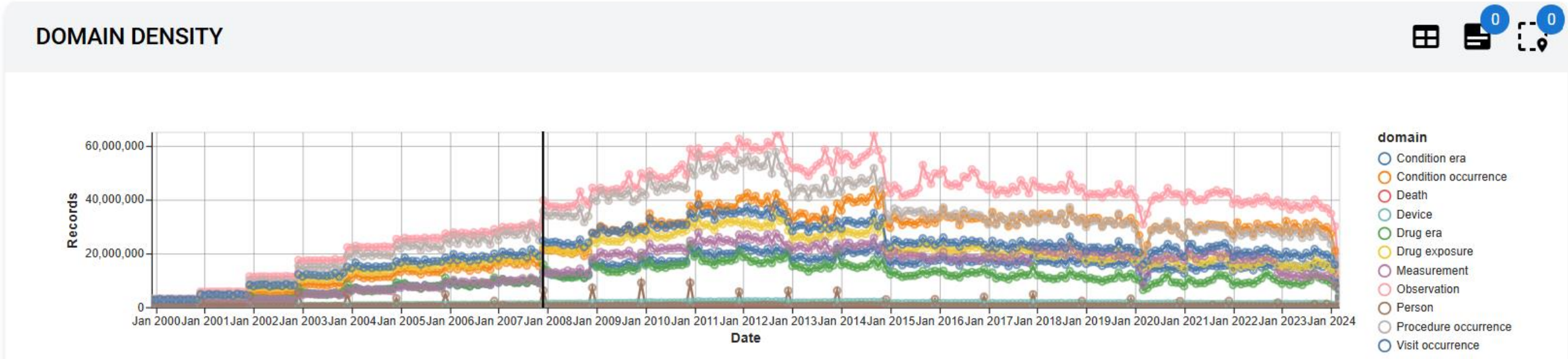
https://github.com/OHDSI/Achilles

Provides descriptive statistics on an OMOP CDM

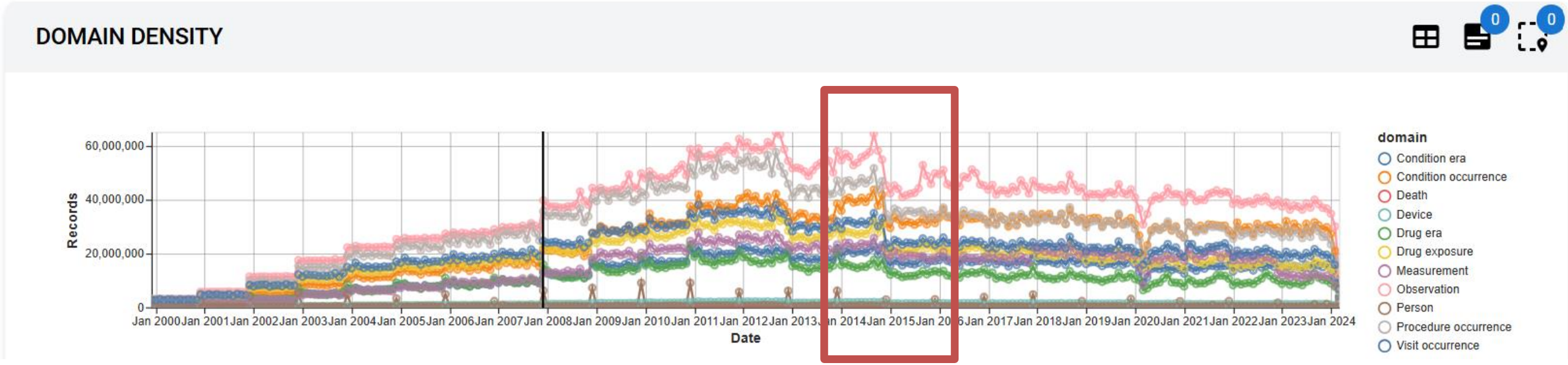
Results can be visualized in ARES or ATLAS



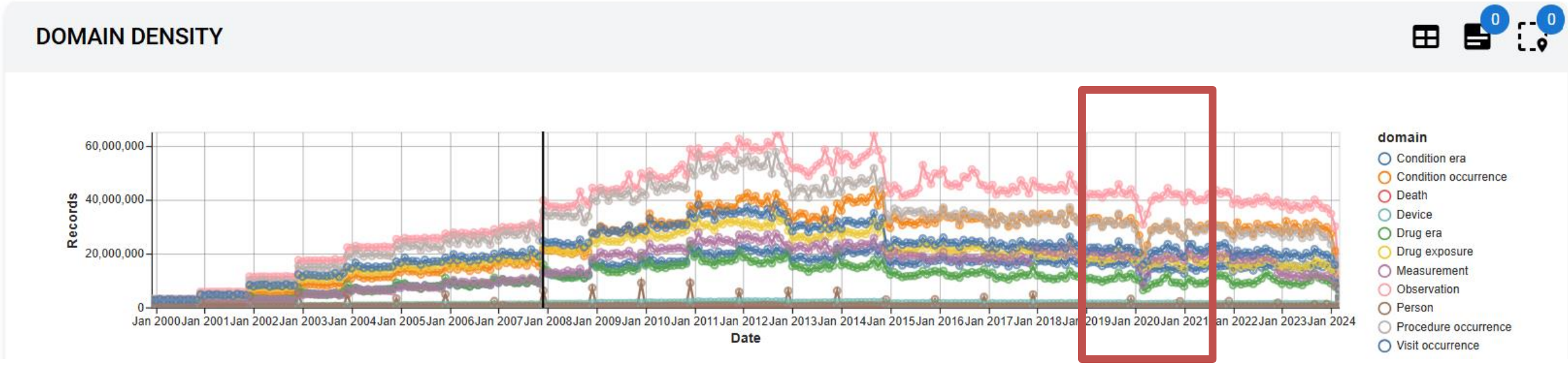
ARES: Data Density Plot



ARES: Data Density Plot



ARES: Data Density Plot



DataQualityDashboard (DQD)

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

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

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

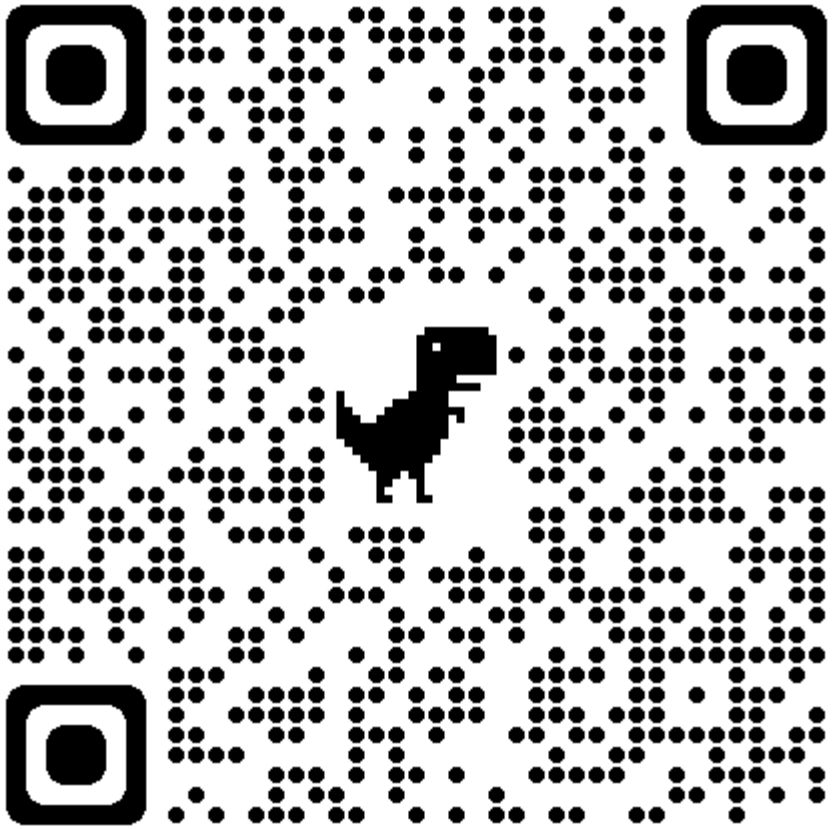
	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%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

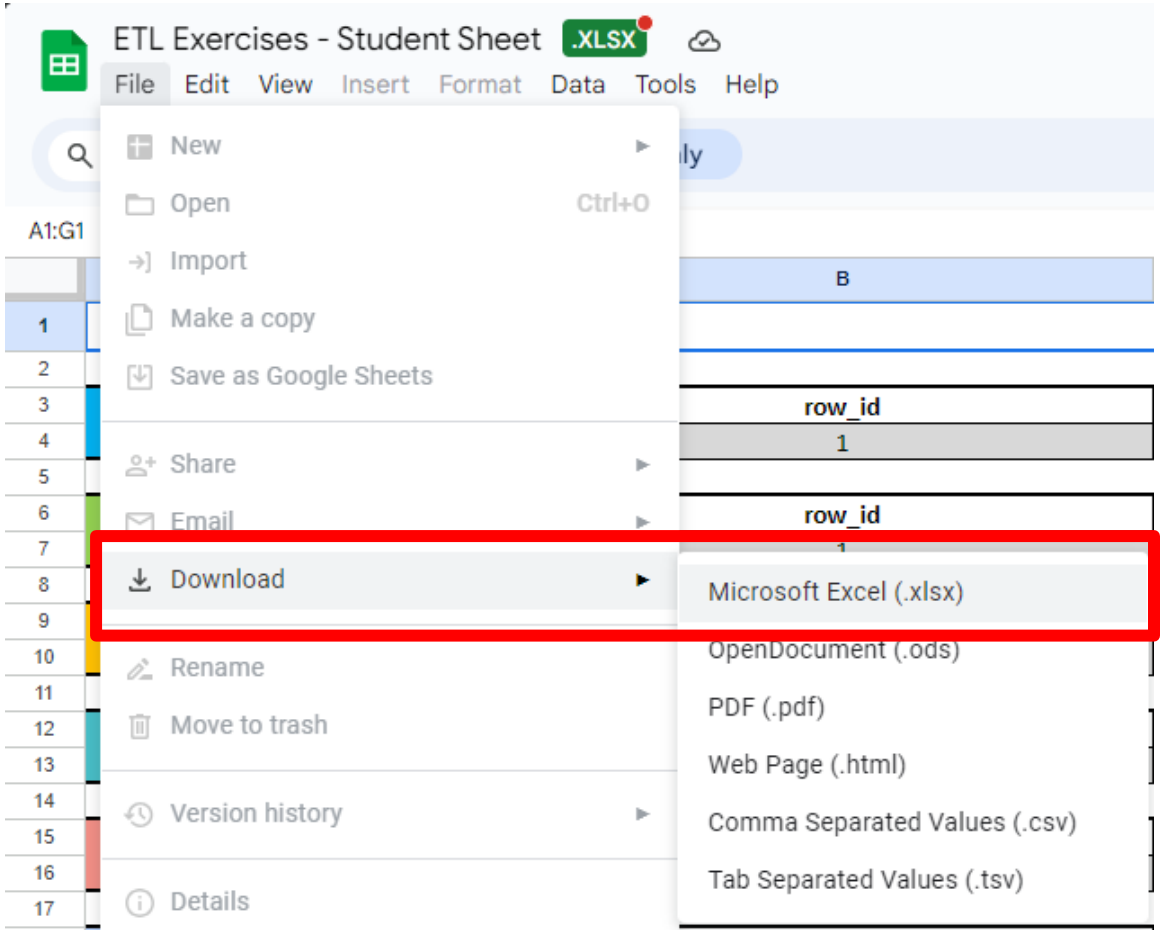
DQD Example Rules

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

Exercise Instructions

- Download a copy of the exercises at:





Exercise Instructions

- Together as a group, we will map the native data provided to the OMOP CDM using the template provided in the *ETL Development_1000* sheet
- You will then be given time to do the same on your own for the *ETL Development_1005* and *ETL Development_1010* sheets


OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank you!