OHDSI

Observational Health Data Sciences and Informatics

OMOP Common Data Model Extract, Transform & Load Tutorial

What this tutorial will provide . . .

- Suggested process for developing a CDM ETL
- OHDSI ETL tools:
  White Rabbit, Rabbit-In-A-Hat, and Usagi
- Resources like the Book of OHDSI and THEMIS
- Generation of a simple ETL examples

Agenda

Time	Agenda Item
9:00-9:30	Overview
9:30-10:45	ETL Step 1 – Design Your ETL
10:45-11:15	Break
11:15-12:00	ETL Step 2 – Mapping to the Vocabulary
12:00-1:00	Lunch
1:00-1:30	ETL Step 2 – Mapping to the Vocabulary (continued)
1:30-3:00	ETL Step 3 – Develop ETL
3:00-3:30	Break
3:30-4:15	ETL Step 4 – Quality Control
4:15-4:45	ETL Step 5 – ETL Maintenance
4:45-5:00	ETL Pain Points & Conclusions

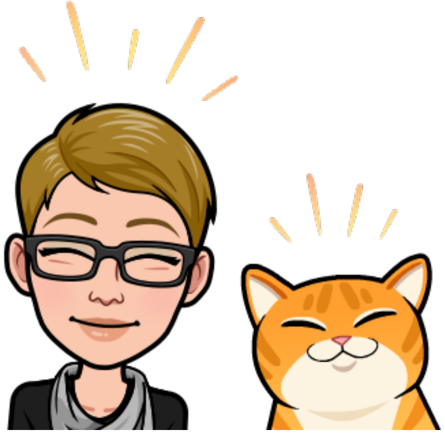
Ground Rules

- We have build in some decent sized breaks, please return before times up

- We are recording this presentation for future use

- We may take some questions off-line if too specific

Instructors

Clair Blacketer	Erica A. Voss
	
Evanette K. Burrows	**Maxim Moinat**

Connecting to the Hotel WIFI

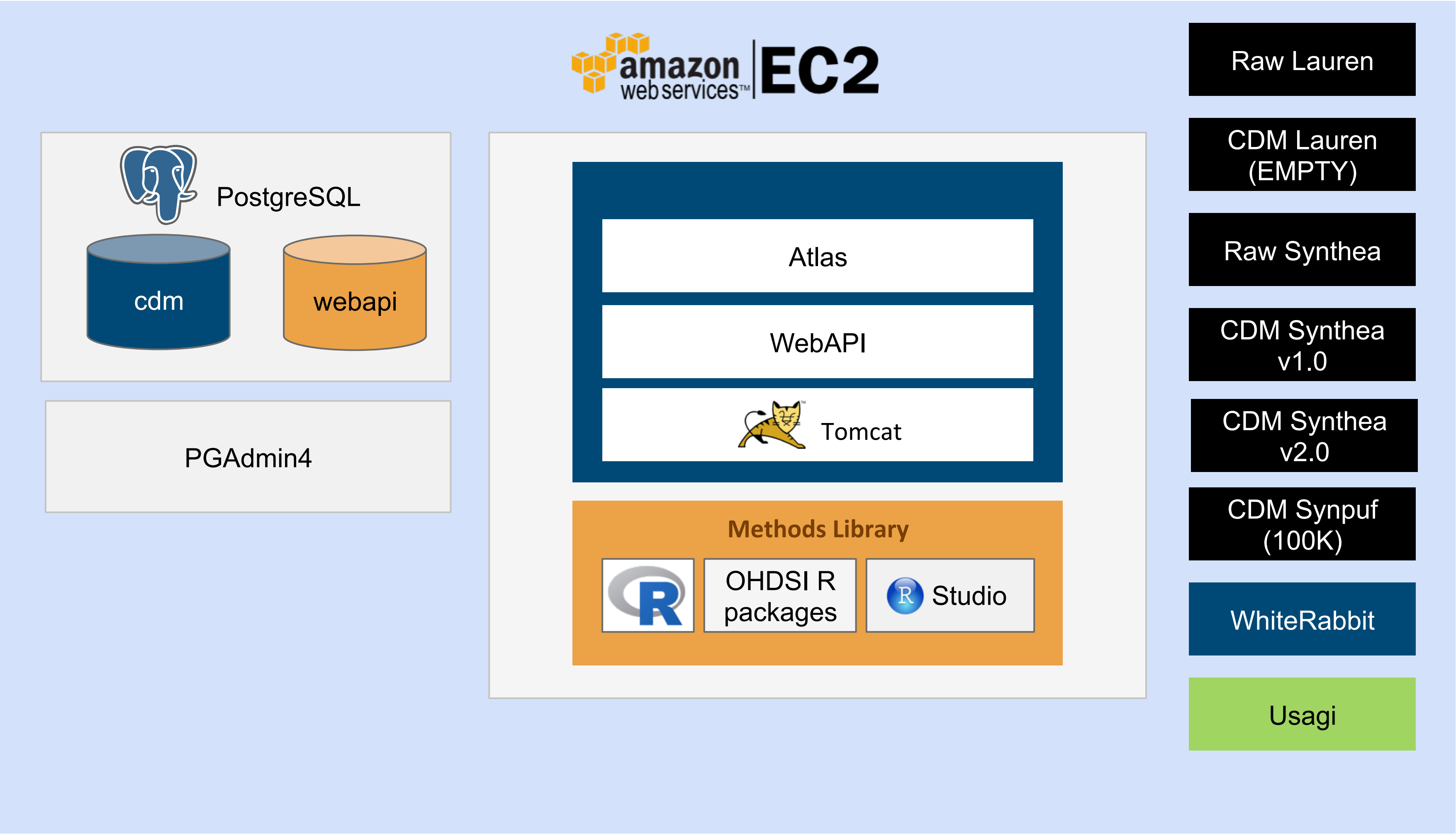
Network: OHDSISYMP

Password: OHDSI2019

Follow Along

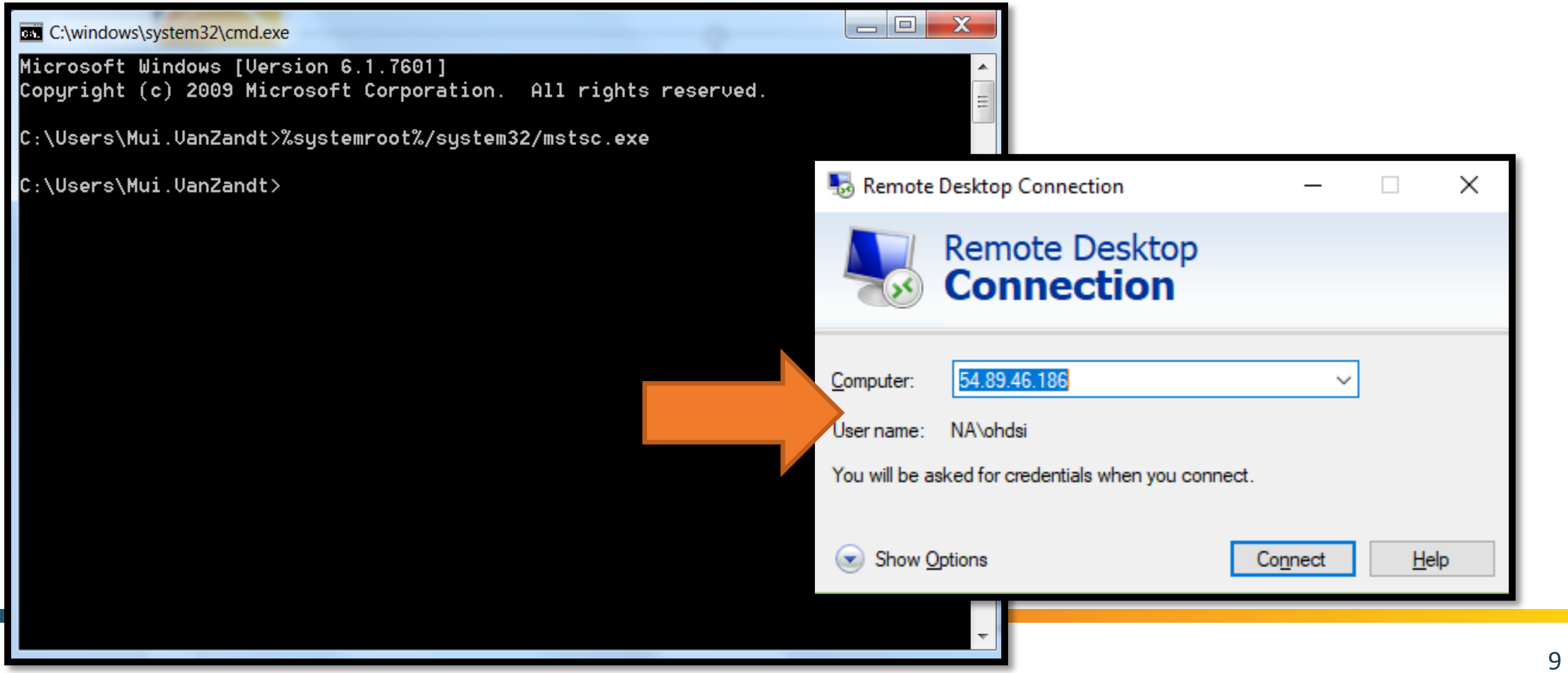
- This full deck can be found here:
  - https://github.com/OHDSI/Tutorial-ETL
  - Materials → OMOP Common Data Model Extract, Transform & Load.pptx

OHDSI in a Box



How to Sign into the Remote Desktop

From your command prompt, type
`%systemroot%/system32/mstsc.exe`
to launch Remote Desktop



C:\windows\system32\cmd.exe

```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\Mui.VanZandt>%systemroot%/system32/mstsc.exe

C:\Users\Mui.VanZandt>
```

Remote Desktop Connection

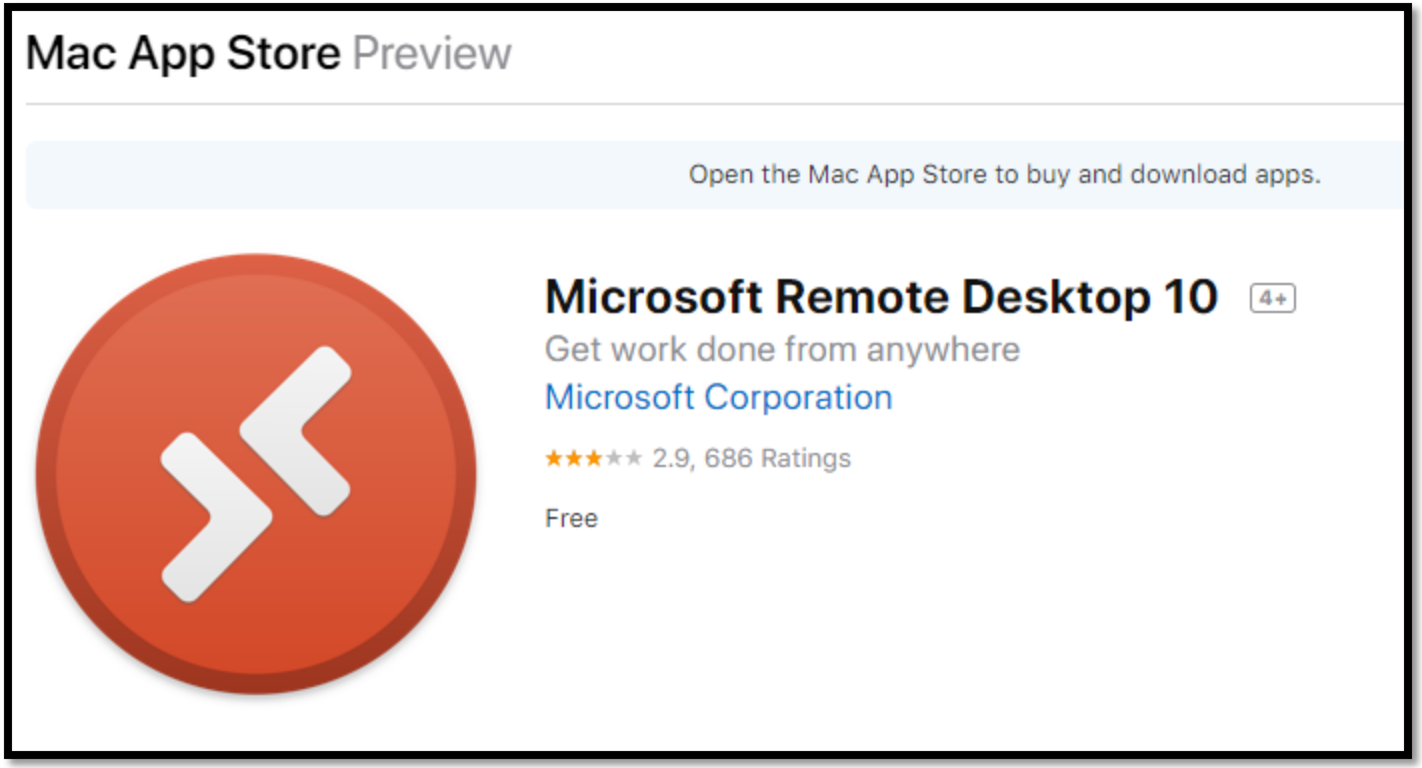
Remote Desktop **Connection**

Computer: 54.89.46.186

User name: NA\ohdsi

You will be asked for credentials when you connect.

How to Sign into the Remote Desktop



Mac App Store Preview

Open the Mac App Store to buy and download apps.

Microsoft Remote Desktop 10 4+

Get work done from anywhere

Microsoft Corporation

★★★★★ 2.9, 686 Ratings

Free

How to Sign into the Remote Desktop

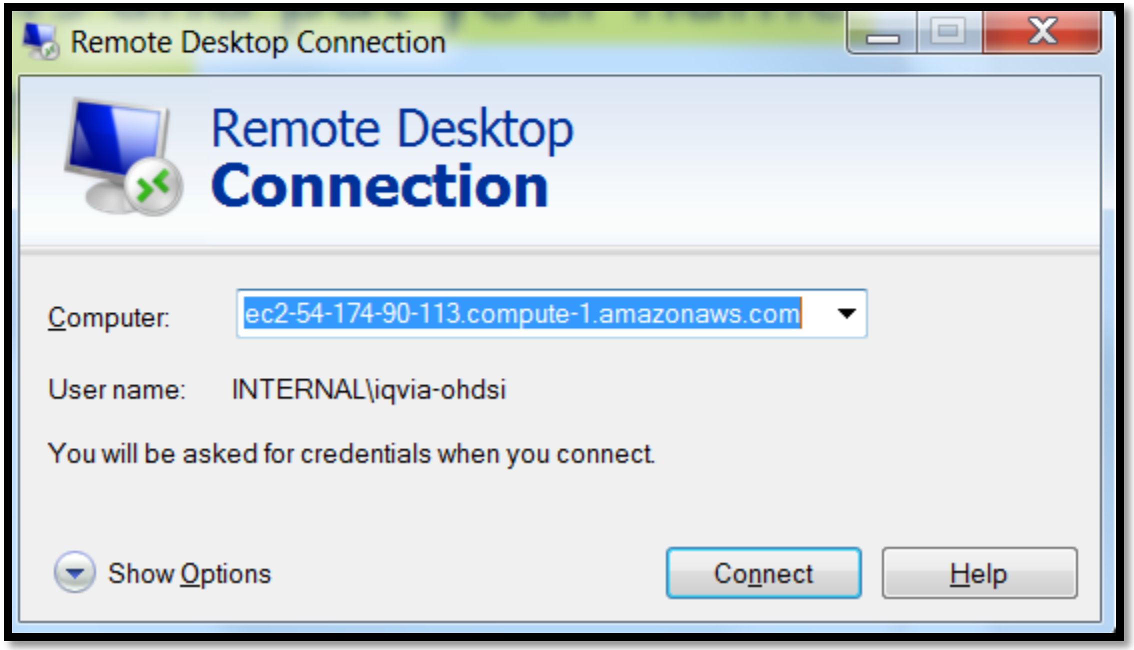
- Use the shortcut on the desktop named “Remote Desktop”

URL TBD

- Pick one of the rows and put your name on the second column

How to Sign into the Remote Desktop

- Take Column A from spreadsheet and copy into the “Computer” field



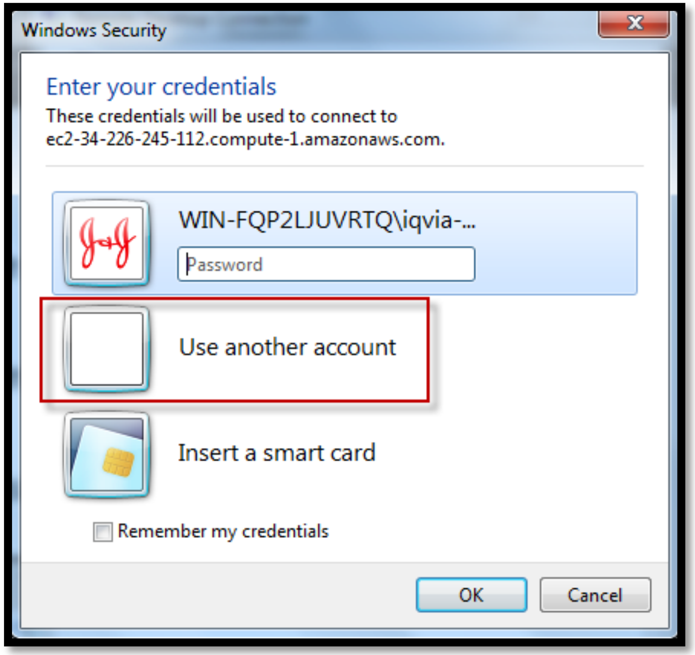
Computer: ec2-54-174-90-113.compute-1.amazonaws.com

User name: INTERNAL\iqvia-ohdsi

You will be asked for credentials when you connect.

How to Sign into the Remote Desktop

- Pick ‘Use Another Account’
- Copy username from Column C
- Copy password from Column D



Windows Security

Enter your credentials

These credentials will be used to connect to ec2-34-226-245-112.compute-1.amazonaws.com.

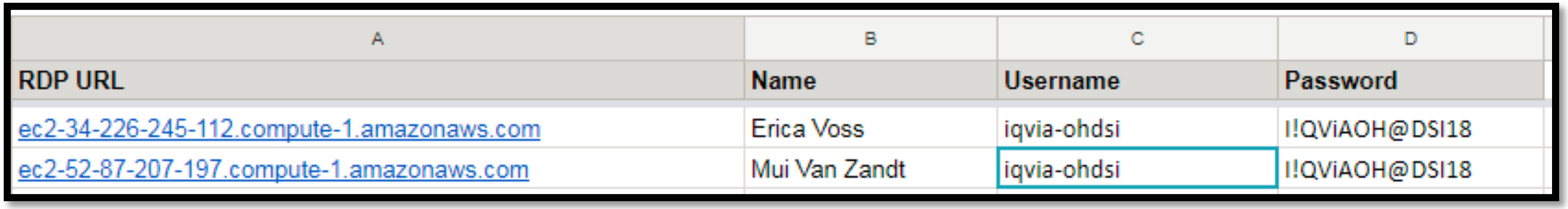
WIN-FQP2LJUVRTQ\iqvia-...

Password

Use another account

Insert a smart card

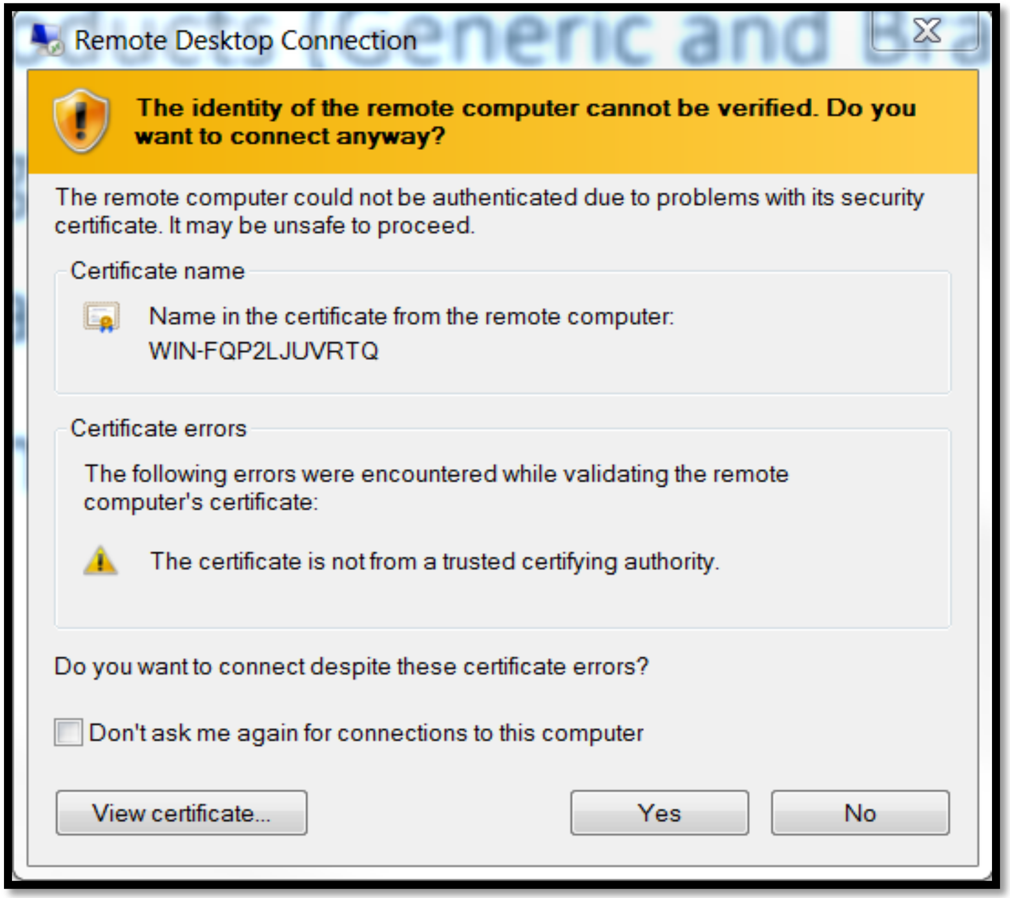
□ Remember my credentials



A	B	C	D
RDP URL	Name	Username	Password
ec2-34-226-245-112.compute-1.amazonaws.com	Erica Voss	iqvia-ohdsi	I!QViAOH@DSI18
ec2-52-87-207-197.compute-1.amazonaws.com	Mui Van Zandt	iqvia-ohdsi	I!QViAOH@DSI18

How to Sign into the Remote Desktop

- If you get this page, select "Yes"



Remote Desktop Connection

The identity of the remote computer cannot be verified. Do you want to connect anyway?

The remote computer could not be authenticated due to problems with its security certificate. It may be unsafe to proceed.

Certificate name

Name in the certificate from the remote computer:
WIN-FQP2LJUVRTQ

Certificate errors

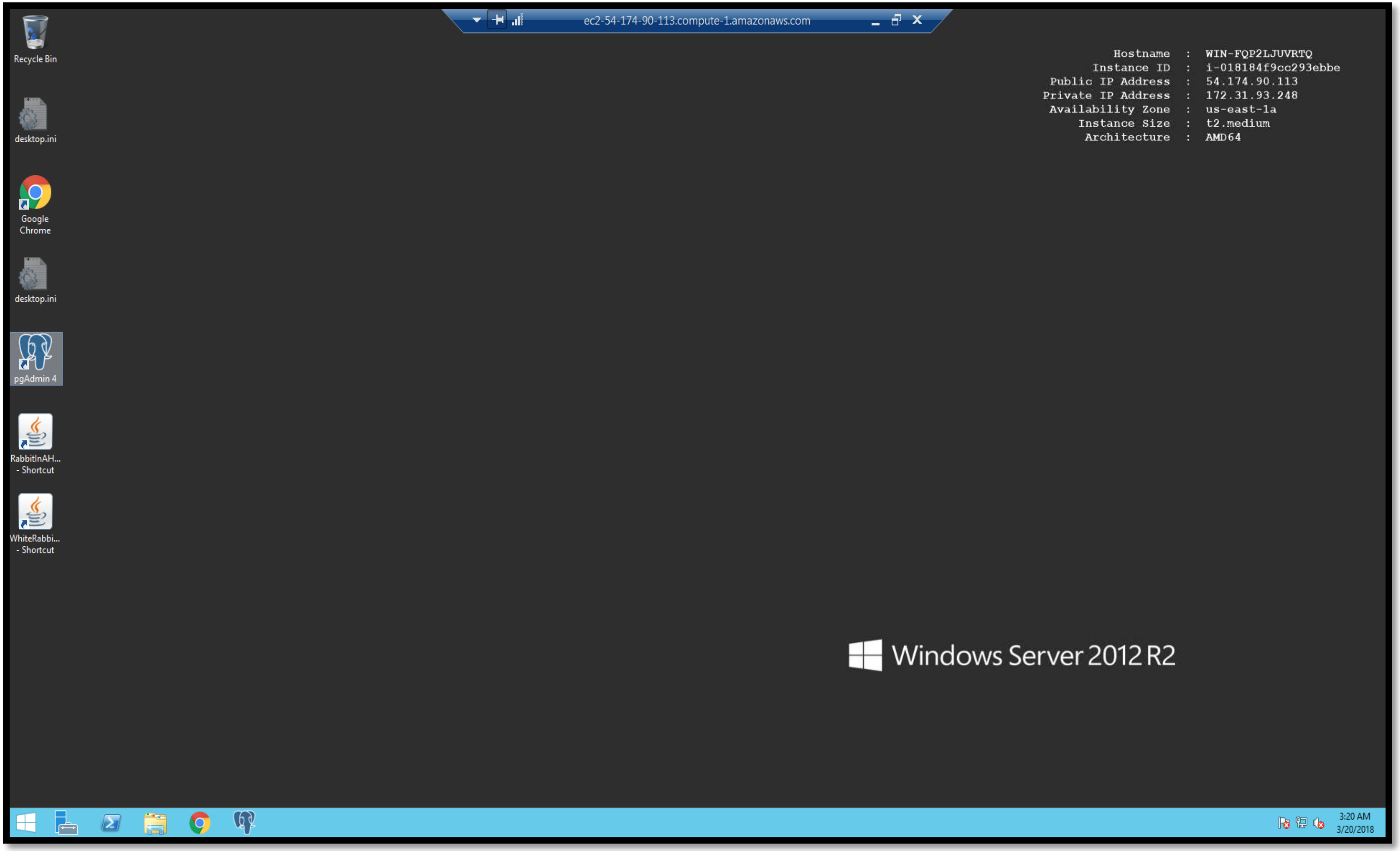
The following errors were encountered while validating the remote computer's certificate:

⚠ The certificate is not from a trusted certifying authority.

Do you want to connect despite these certificate errors?

☐ Don't ask me again for connections to this computer

OHDSI in a Box – Ready



```
          Hostname  :  WIN-FQP2LJUVRTQ
       Instance ID  :  i-018184f9cc293ebbe
 Public IP Address  :  54.174.90.113
Private IP Address  :  172.31.93.248
 Availability Zone  :  us-east-1a
     Instance Size  :  t2.medium
      Architecture  :  AMD64
```

OHDSI's Mission & Vision

To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

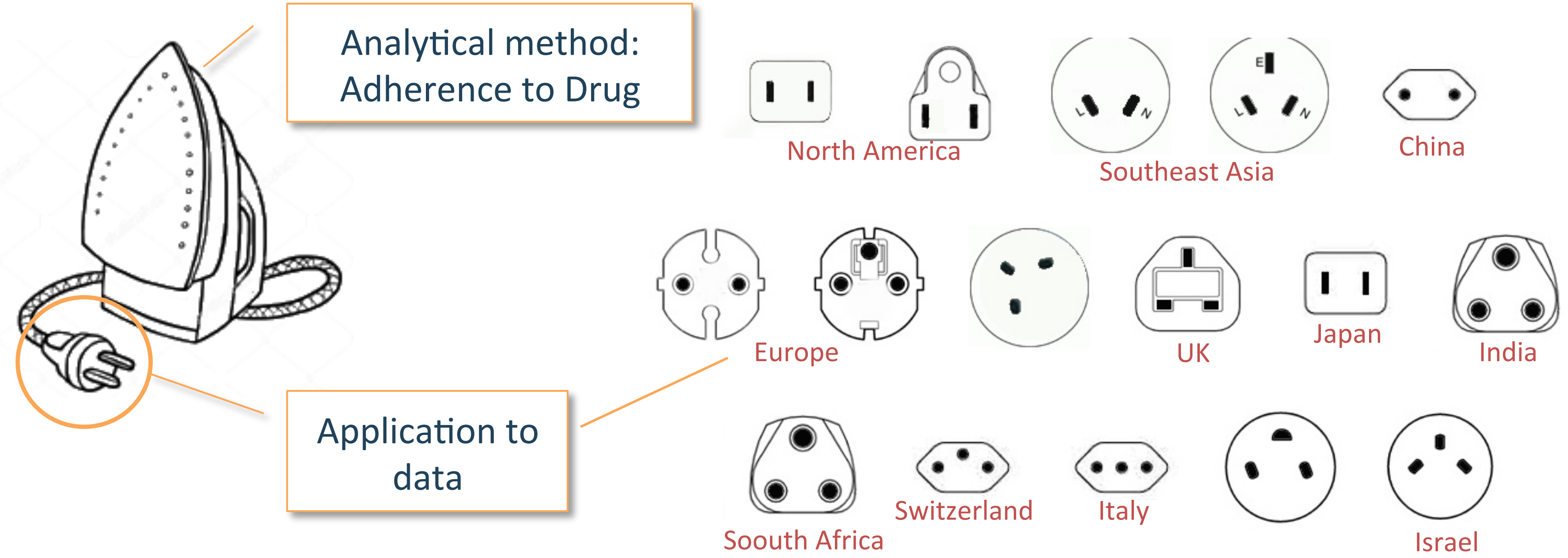
A world in which observational research produces a comprehensive understanding of health and disease.

Join us on the journey

http://ohdsi.org

Current Approach: “One Study – One Script“

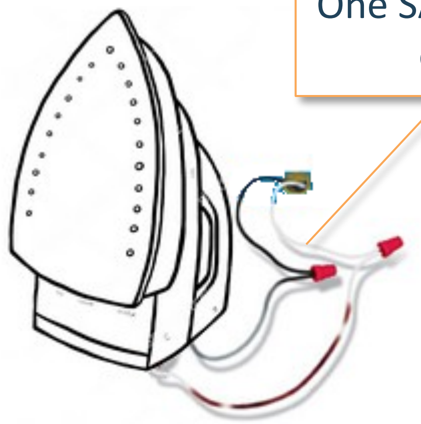
"What's the adherence to my drug in the data assets I own?"



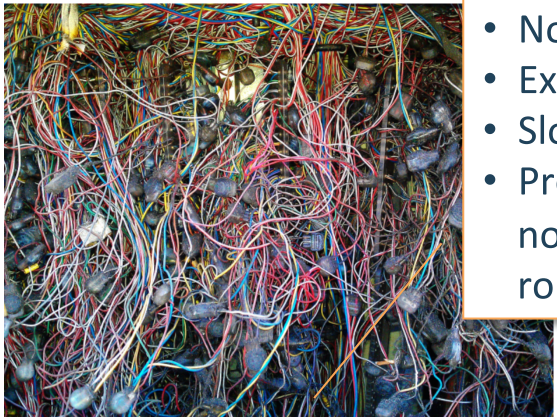
Analytical method: Adherence to Drug

Application to data

Current solution:



One SAS or R script for each study



- Not scalable
- Not transparent
- Expensive
- Slow
- Prohibitive to non-expert routine use

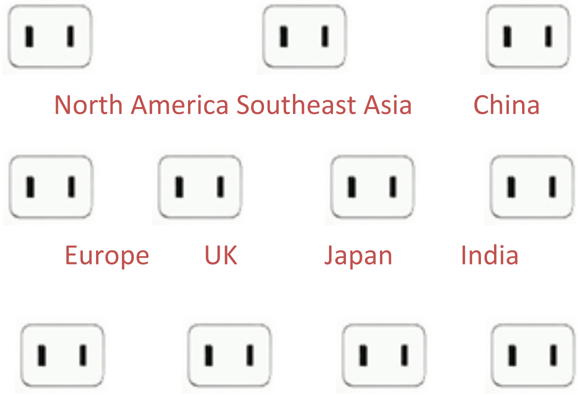
Solution: Data Standardization Enables Systematic Research

Adherence

Mortality

Source of Business

Safety Signals



North America Southeast Asia

China

Europe

UK

Japan

India

So Africa

Switzerland

Italy

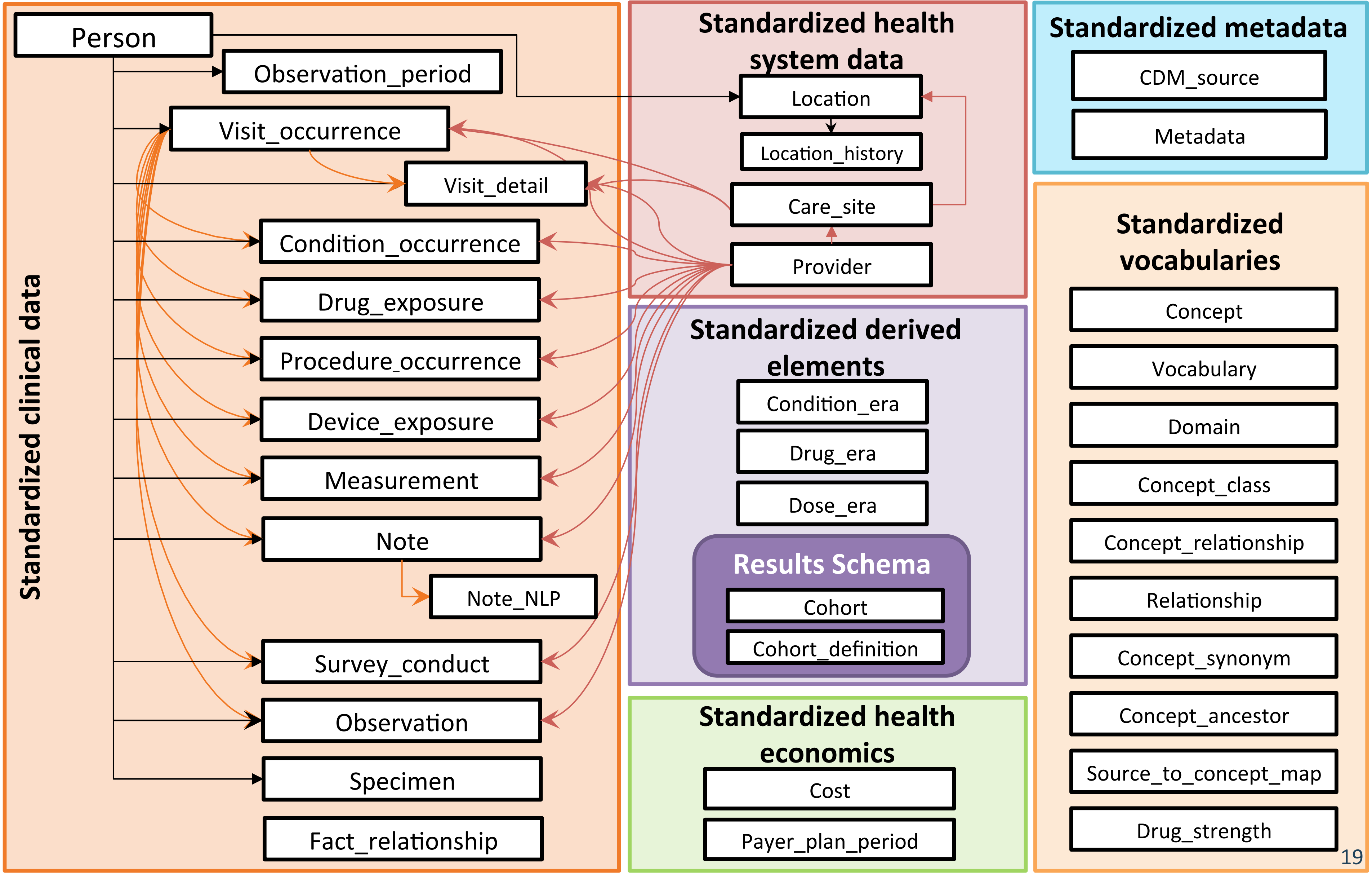
Israel

Standardized data

OHDSI Tools

OMOP CDM

CDM Version 6 Key Domains



Standardized clinical data

Person

Observation_period

Visit_occurrence

Visit_detail

Condition_occurrence

Drug_exposure

Procedure_occurrence

Device_exposure

Measurement

Note

Note_NLP

Survey_conduct

Observation

Specimen

Fact_relationship

Standardized health system data

Location

Location_history

Care_site

Provider

Standardized derived elements

Condition_era

Drug_era

Dose_era

Results Schema

Cohort

Cohort_definition

Standardized health economics

Cost

Payer_plan_period

Standardized metadata

CDM_source

Metadata

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_map

Drug_strength

Why the CDM?

Ability to pursue **cross-institutional collaborations**

Write **one program** to run on multiple data assets

OMOP Vocabularies has greatly increased our **ability to find relevant codes**

You truly **know your data** if you convert it to the CDM

If you know a problem with your data, you can use the **ETL to address it**

Whole community of researchers across diverse organizations and countries

You can use **standardized tools** developed by OHDSI like ATLAS and the Patient Level Prediction Package

The CDM brings **consistency** to observational research through standardization of many of its components

Buy vs Build: leverage an entire community of technical and scientific capability for **"free"**

Takes observational research towards **open science**

ETL

- Extract, Transform, Load

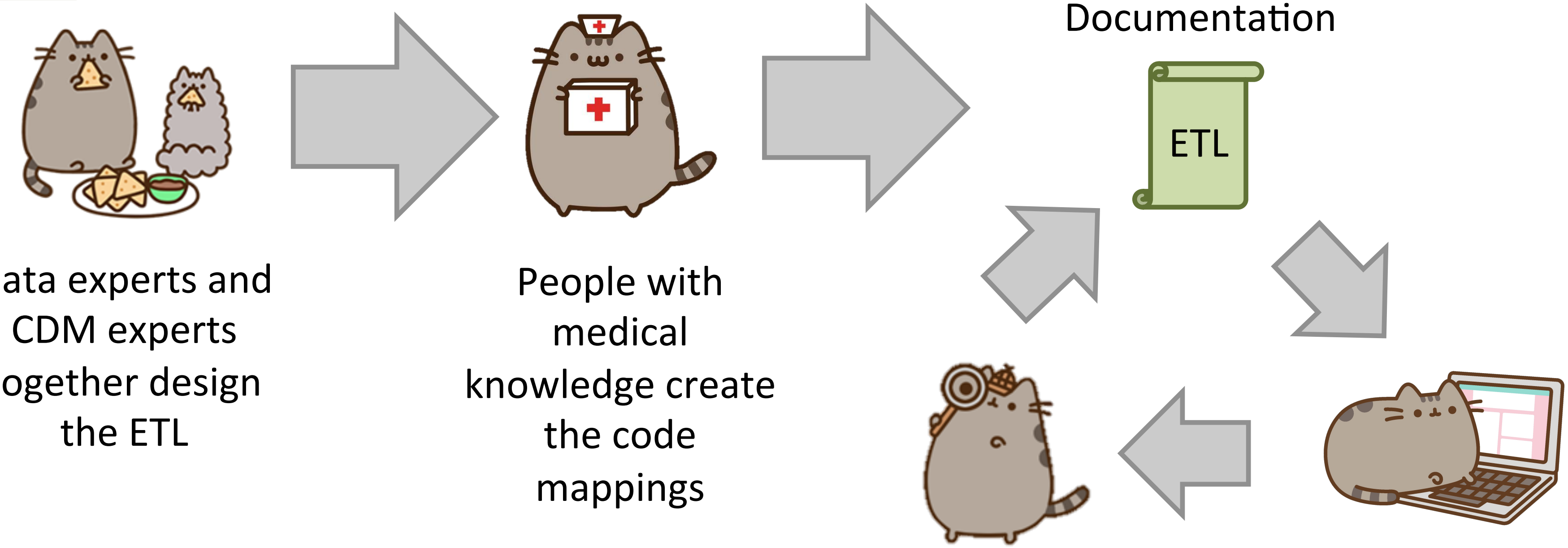
- In order to get from our native/raw data into the OMOP CDM we need to design and develop and ETL process

Patient-level data in source system/schema
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
Reliable evidence

- Goal in ETLing is to standardize the format and terminology

- This tutorial
  - Will teach you best practices around designing an ETL and CDM maintenance
  - Will not teach you how to program an ETL

ETL Process



Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

A technical person implements the ETL

All are involved in quality control

OHDSI Tools

White Rabbit

Rabbit In a Hat

Usagi

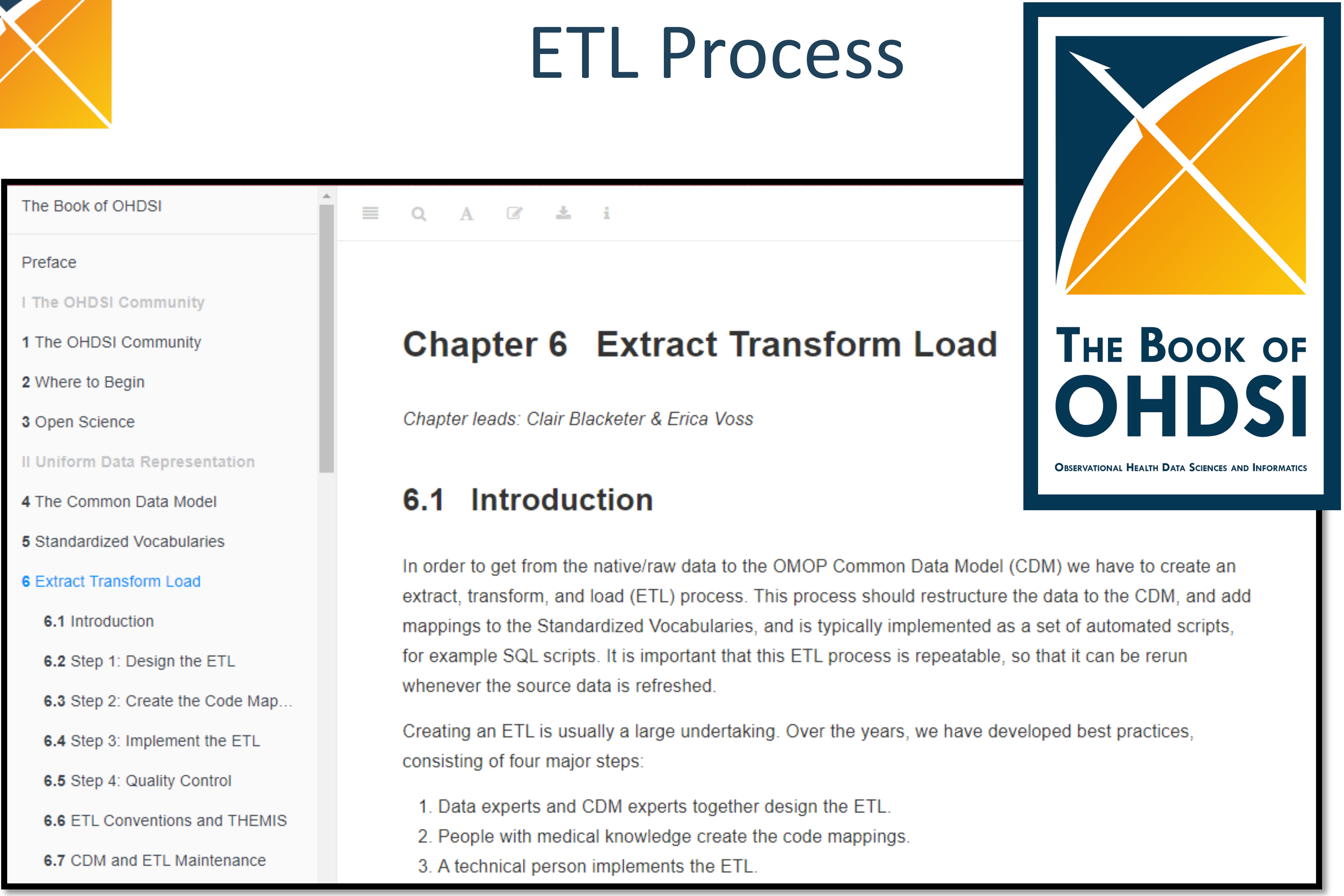
White Rabbit

ACHILLES

DQD

Rabbit In a Hat

ETL Process



The Book of OHDSI

Chapter 6 Extract Transform Load

Chapter leads: Clair Blacketer & Erica Voss

6.1 Introduction

In order to get from the native/raw data to the OMOP Common Data Model (CDM) we have to create an extract, transform, and load (ETL) process. This process should restructure the data to the CDM, and add mappings to the Standardized Vocabularies, and is typically implemented as a set of automated scripts, for example SQL scripts. It is important that this ETL process is repeatable, so that it can be rerun whenever the source data is refreshed.

Creating an ETL is usually a large undertaking. Over the years, we have developed best practices, consisting of four major steps:

1. Data experts and CDM experts together design the ETL.
2. People with medical knowledge create the code mappings.
3. A technical person implements the ETL.

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

Hands On Exercises for Today

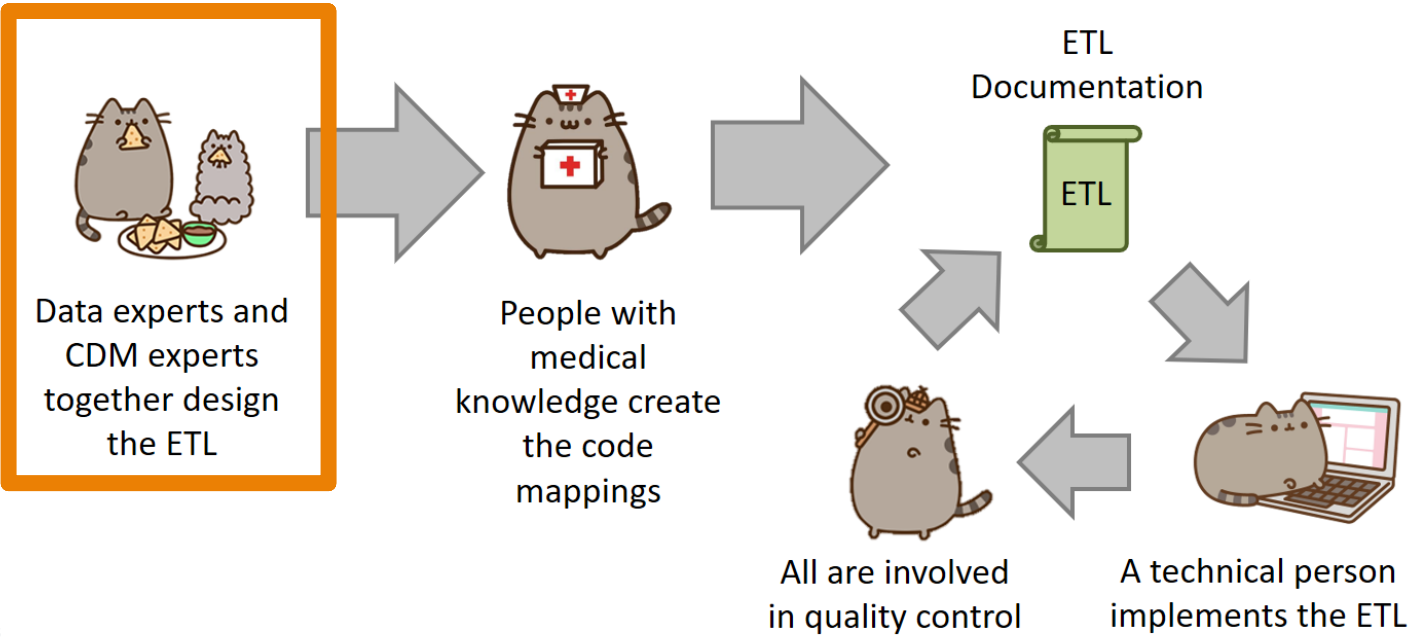
- Scan a database with White Rabbit

- Practice building an ETL document with Rabbit in a Hat

- Mapping Source Codes by with the OMOP Vocabulary and USAGI

OHDSI

Observational Health Data Sciences and Informatics



Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

A technical person implements the ETL

All are involved in quality control

A Patient's Story: Lauren

Lauren's story

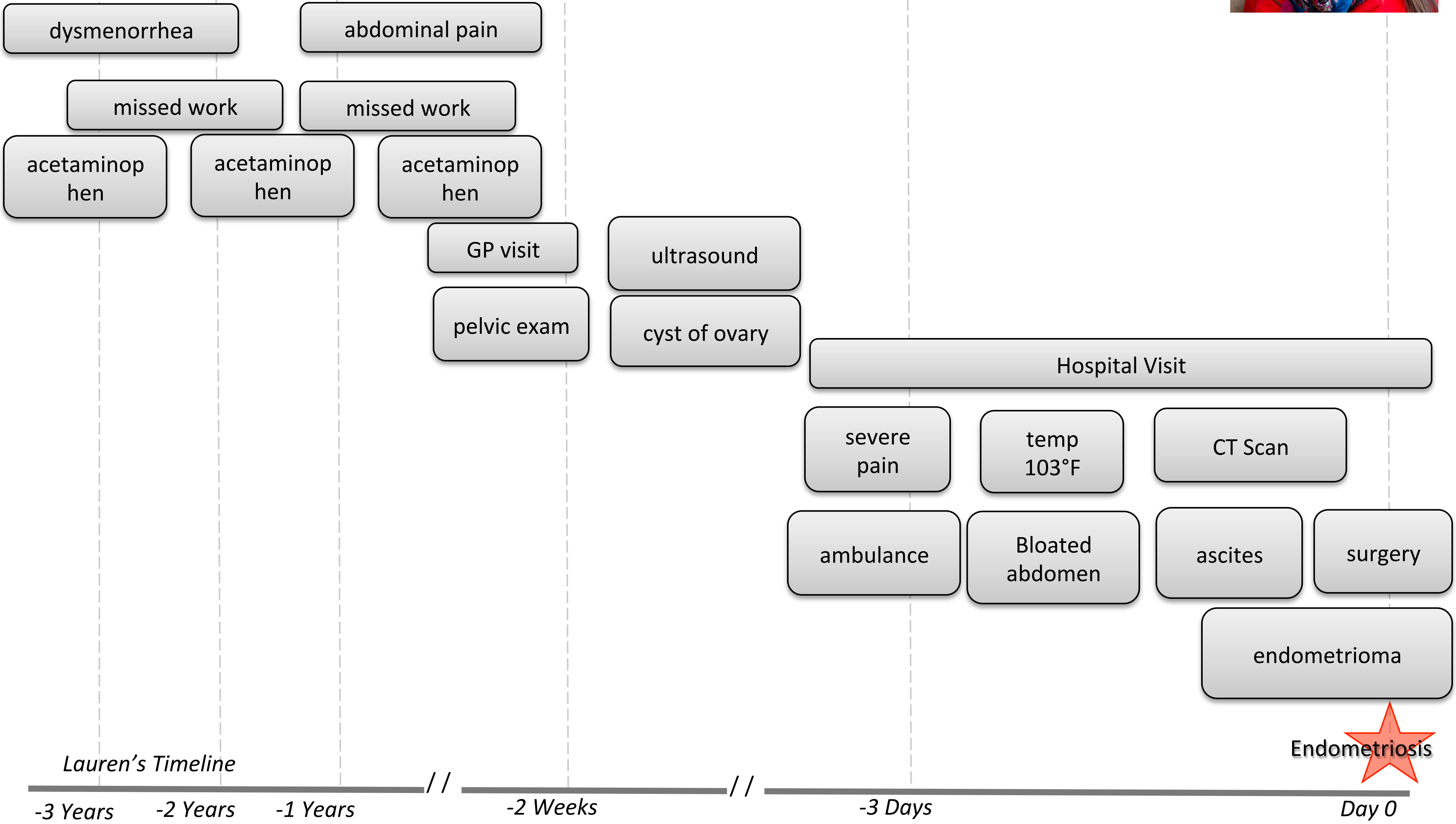
endometriosis uk

"Every step of this painful journey I've had to convince everyone how much pain I was in."

"My first surgery taught me that I had to be very patient with my recovery and very patient with myself in general."

https://www.endometriosis-uk.org/laurens-story

What data do we have?



dysmenorrhea

abdominal pain

missed work

missed work

acetaminophen

acetaminophen

acetaminophen

GP visit

ultrasound

pelvic exam

cyst of ovary

Hospital Visit

severe pain

temp 103°F

CT Scan

ambulance

Bloated abdomen

ascites

surgery

endometrioma

Endometriosis

Lauren's Timeline

-3 Years *-2 Years* *-1 Years* // *-2 Weeks* // *-3 Days* *Day 0*

Data Format

- Synthea^{TM} is a Synthetic Patient Population Simulator. The goal is to output synthetic, realistic (but not real), patient data and associated health records in a variety of formats.

- The resulting data is free from cost, privacy, and security restrictions. It can be used without restriction for a variety of secondary uses in academia, research, industry, and government (although a citation would be appreciated).

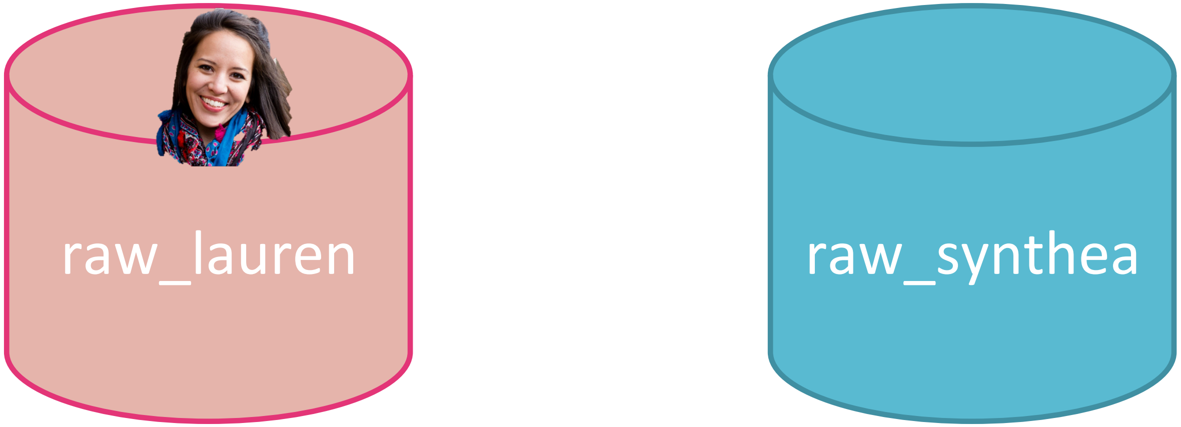
- https://github.com/synthetichealth/synthea

Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2017 Aug 30. doi: 10.1093/jamia/ocx079. [Epub ahead of print] PubMed PMID: 29025144.

Synthea Tables

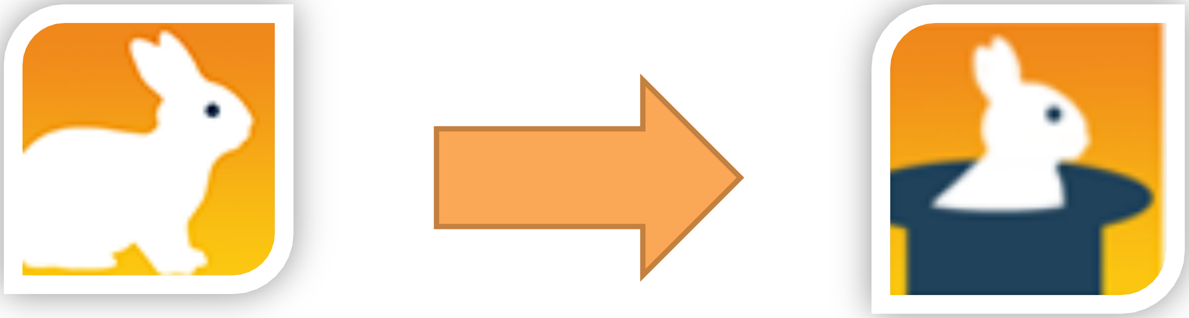
File	Description
allergies.csv	Patient allergy data.
careplans.csv	Patient care plan data, including goals.
conditions.csv	Patient conditions or diagnoses.
encounters.csv	Patient encounter data.
imaging_studies.csv	Patient imaging metadata.
immunizations.csv	Patient immunization data.
medications.csv	Patient medication data.
observations.csv	Patient observations including vital signs and lab reports.
organizations.csv	Provider organizations including hospitals.
patients.csv	Patient demographic data.
procedures.csv	Patient procedure data including surgeries.
providers	Clinicians that provide patient care.

Raw Data



1 Patient	1000 Patient
Lauren Data	Synthetic Data
Synthea Format	Synthea Format

Tools help us get started . . .



White Rabbit

- performs a scan of the source data, providing detailed information on the tables, fields, and values that appear in a field

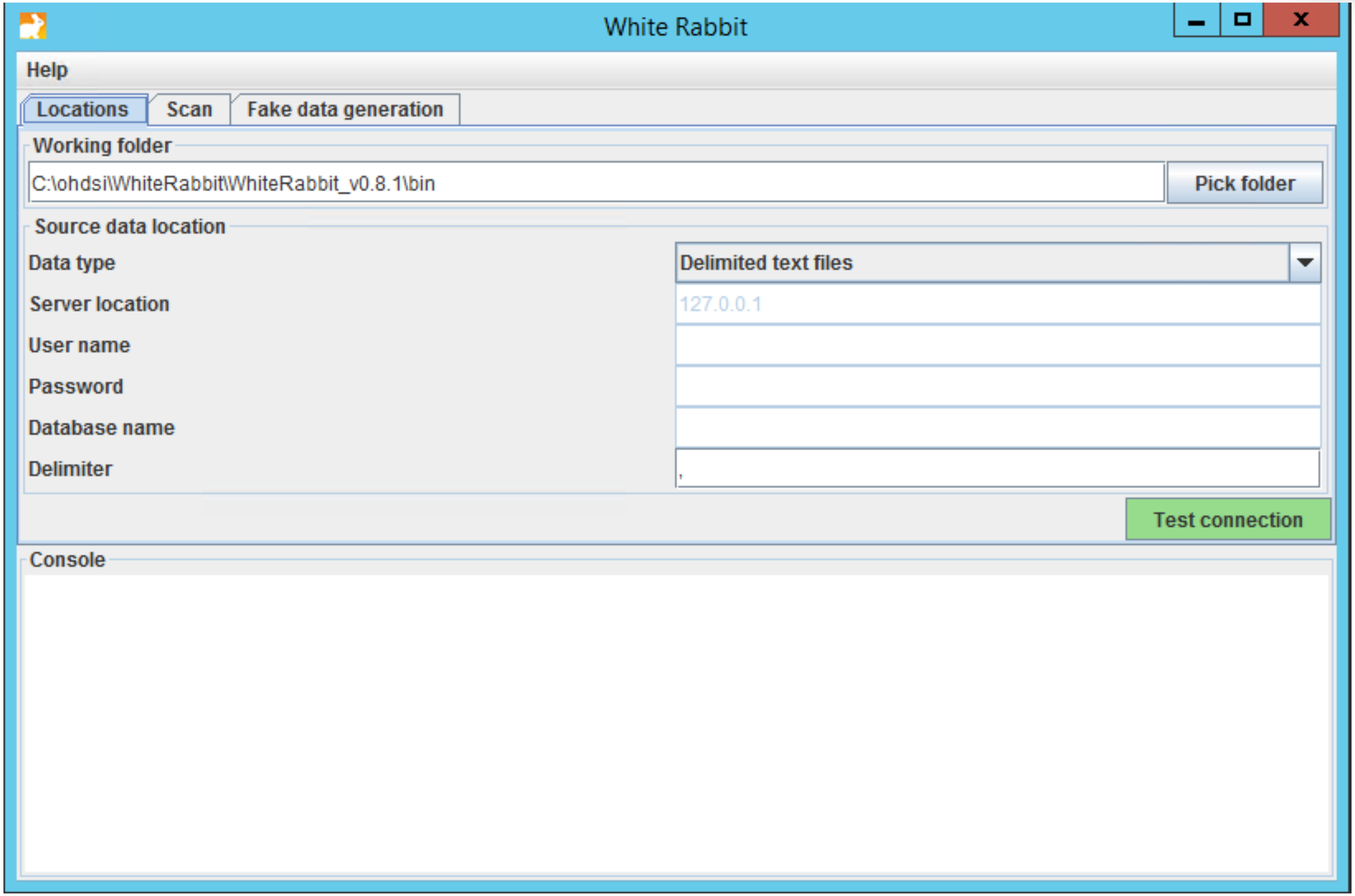
Rabbit In a Hat

- Uses White Rabbit scan to provide a graphical user interface to help build an ETL document

- Does not generate code*

**But people are test driving this*

White Rabbit - Location



Working folder: C:\ohdsi\WhiteRabbit\WhiteRabbit_v0.8.1\bin

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

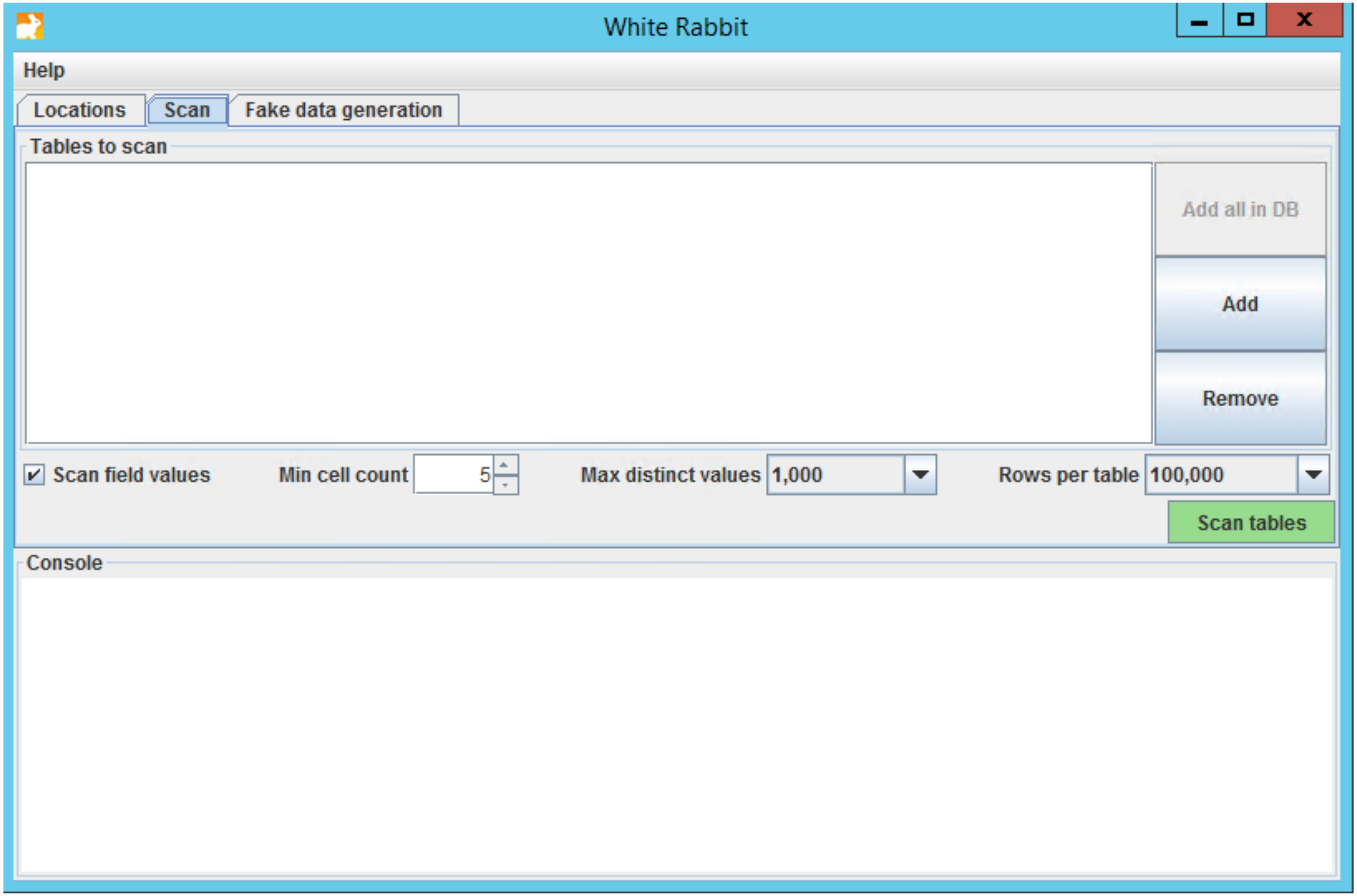
User name:

Password:

Database name:

Delimiter: ,

White Rabbit - Scan



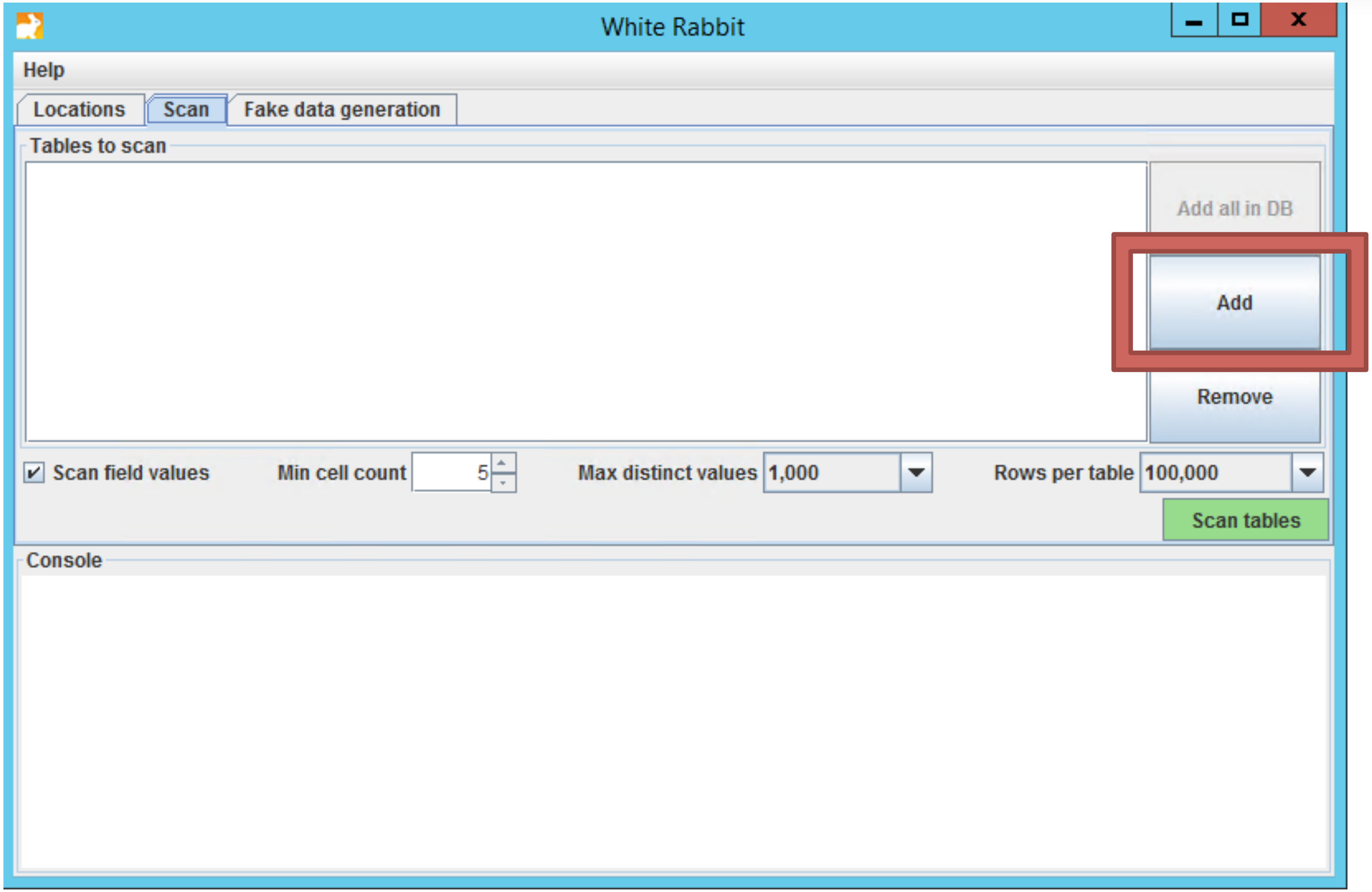
☑ Scan field values

Min cell count 5

Max distinct values 1,000

Rows per table 100,000

White Rabbit - Scan



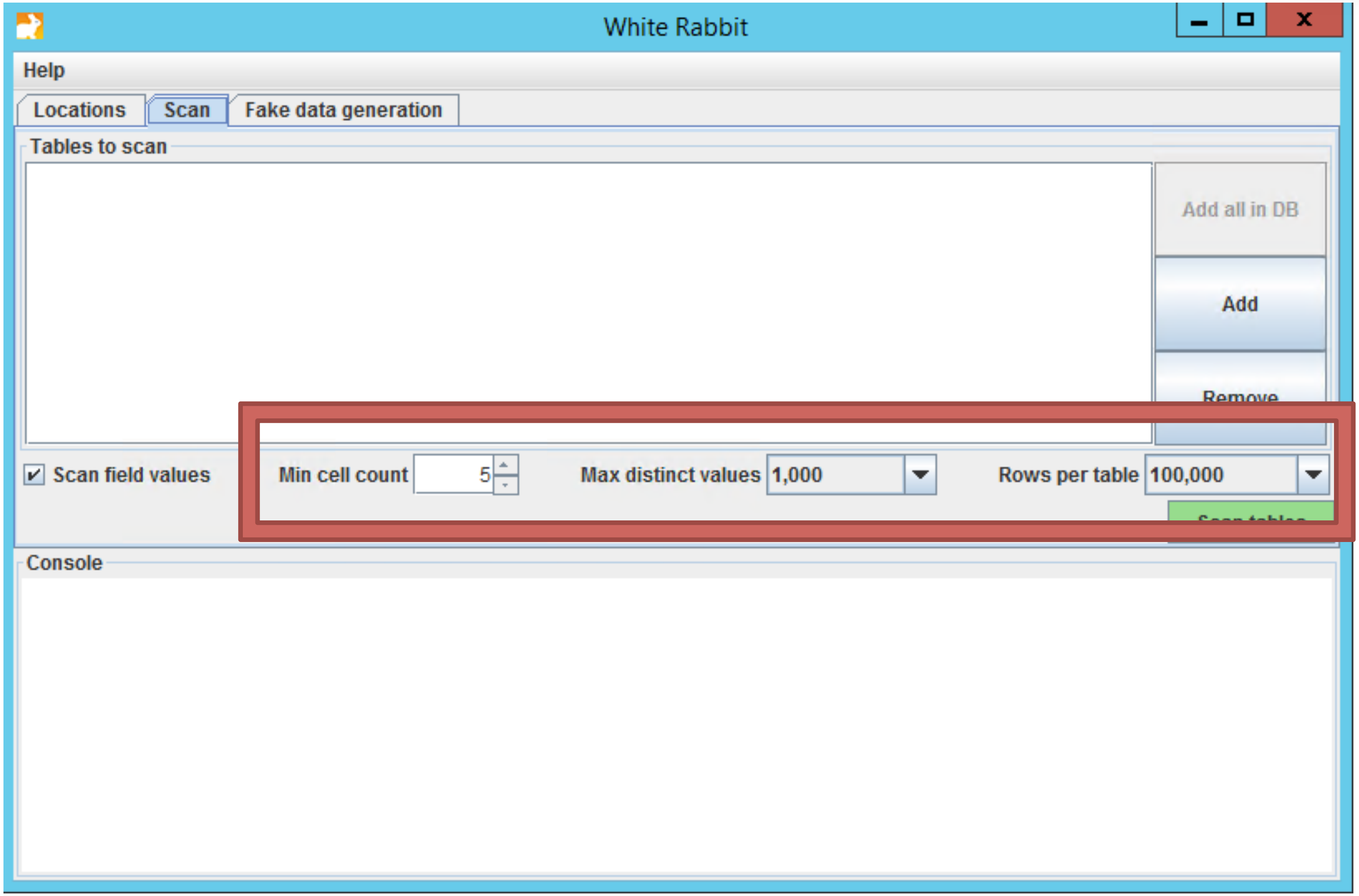
☑ Scan field values

Min cell count: 5

Max distinct values: 1,000

Rows per table: 100,000

White Rabbit - Scan



☑ Scan field values

Min cell count 5

Max distinct values 1,000

Rows per table 100,000

White Rabbit – Scan Report

raw_synthea

- We already ran the scan on raw_synthea

- To open the scan while we review:
  - https://github.com/OHDSI/Tutorial-ETL
  - Materials → WhiteRabbit → ScanReport_raw_synthea.xlsx
  - Click “View Raw” to download the XLSX

White Rabbit – Scan Report: raw_synthea

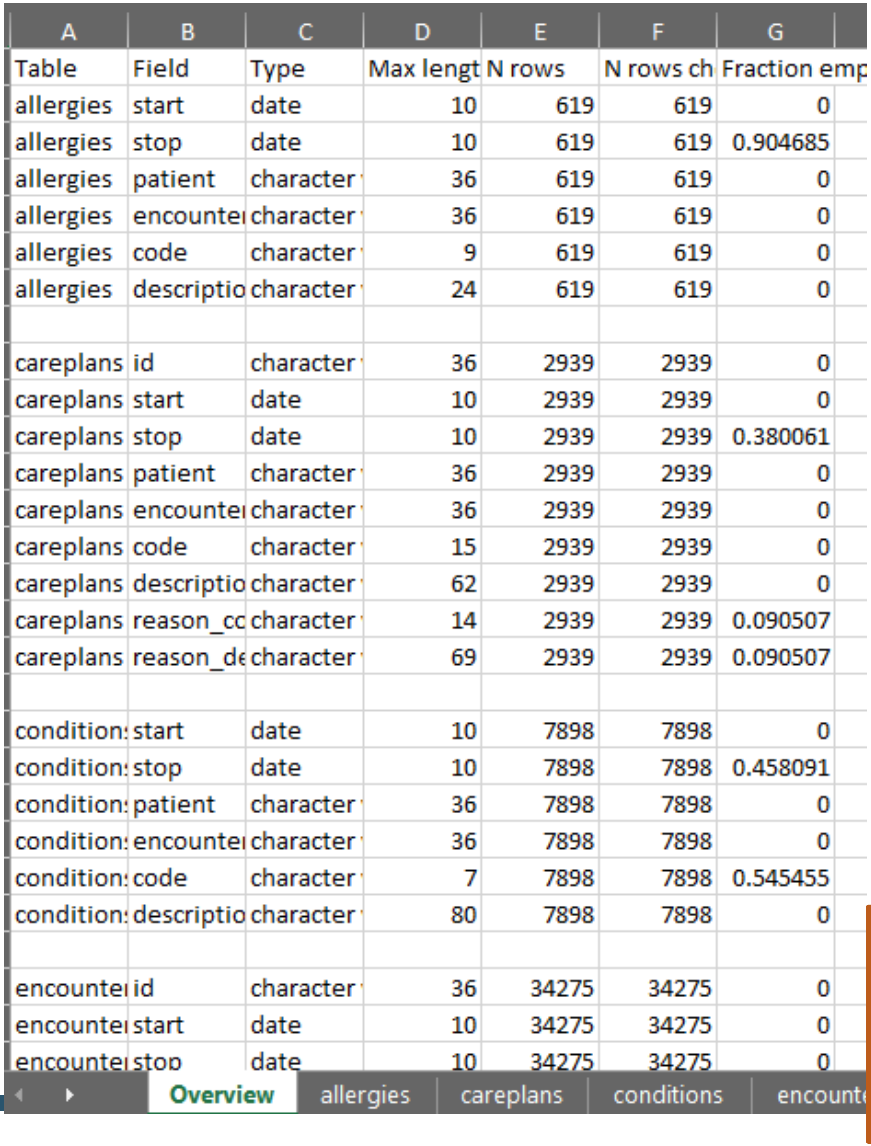
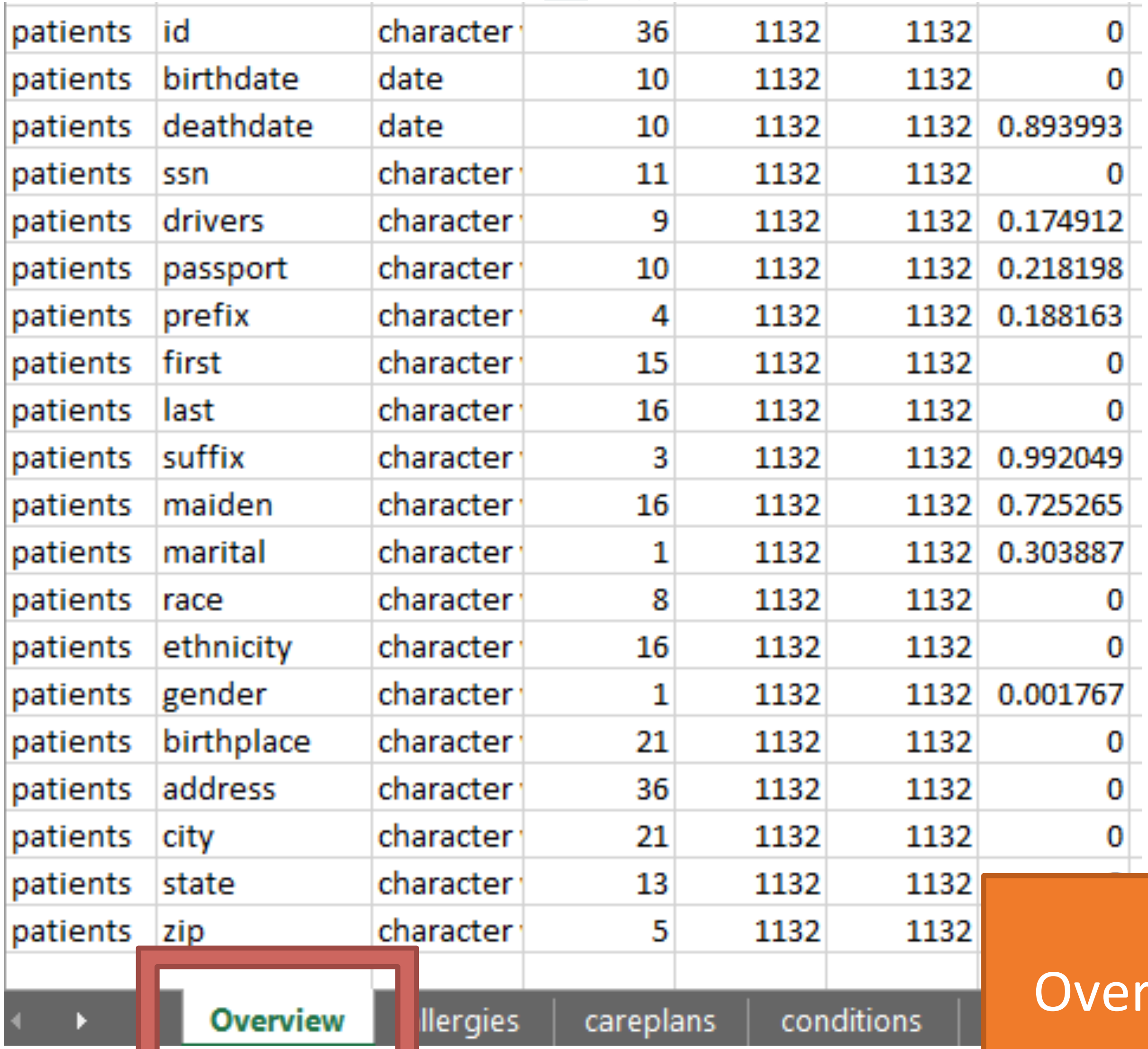


Table	Field	Type	Max lengt	N rows	N rows ch	Fraction emp
allergies	start	date	10	619	619	0
allergies	stop	date	10	619	619	0.904685
allergies	patient	character	36	619	619	0
allergies	encounte	character	36	619	619	0
allergies	code	character	9	619	619	0
allergies	descriptio	character	24	619	619	0
careplans	id	character	36	2939	2939	0
careplans	start	date	10	2939	2939	0
careplans	stop	date	10	2939	2939	0.380061
careplans	patient	character	36	2939	2939	0
careplans	encounte	character	36	2939	2939	0
careplans	code	character	15	2939	2939	0
careplans	descriptio	character	62	2939	2939	0
careplans	reason_cc	character	14	2939	2939	0.090507
careplans	reason_de	character	69	2939	2939	0.090507
condition	start	date	10	7898	7898	0
condition	stop	date	10	7898	7898	0.458091
condition	patient	character	36	7898	7898	0
condition	encounte	character	36	7898	7898	0
condition	code	character	7	7898	7898	0.545455
condition	descriptio	character	80	7898	7898	0
encounte	id	character	36	34275	34275	0
encounte	start	date	10	34275	34275	0
encounte	stop	date	10	34275	34275	0

Overview Tab

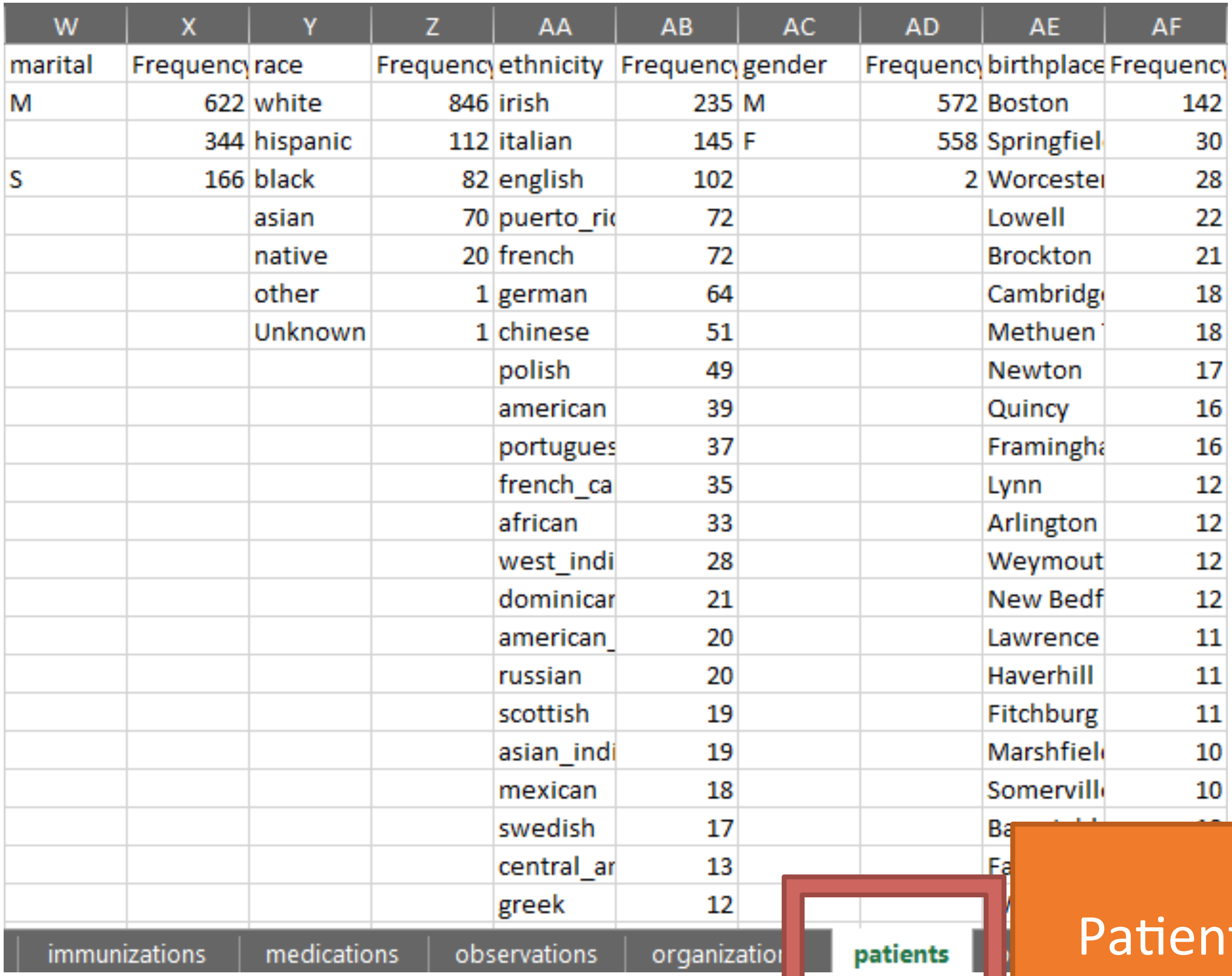
White Rabbit – Scan Report: raw_synthea



patients	id	character	36	1132	1132	0
patients	birthdate	date	10	1132	1132	0
patients	deathdate	date	10	1132	1132	0.893993
patients	ssn	character	11	1132	1132	0
patients	drivers	character	9	1132	1132	0.174912
patients	passport	character	10	1132	1132	0.218198
patients	prefix	character	4	1132	1132	0.188163
patients	first	character	15	1132	1132	0
patients	last	character	16	1132	1132	0
patients	suffix	character	3	1132	1132	0.992049
patients	maiden	character	16	1132	1132	0.725265
patients	marital	character	1	1132	1132	0.303887
patients	race	character	8	1132	1132	0
patients	ethnicity	character	16	1132	1132	0
patients	gender	character	1	1132	1132	0.001767
patients	birthplace	character	21	1132	1132	0
patients	address	character	36	1132	1132	0
patients	city	character	21	1132	1132	0
patients	state	character	13	1132	1132	
patients	zip	character	5	1132	1132	

Overview Tab

White Rabbit – Scan Report: raw_synthea

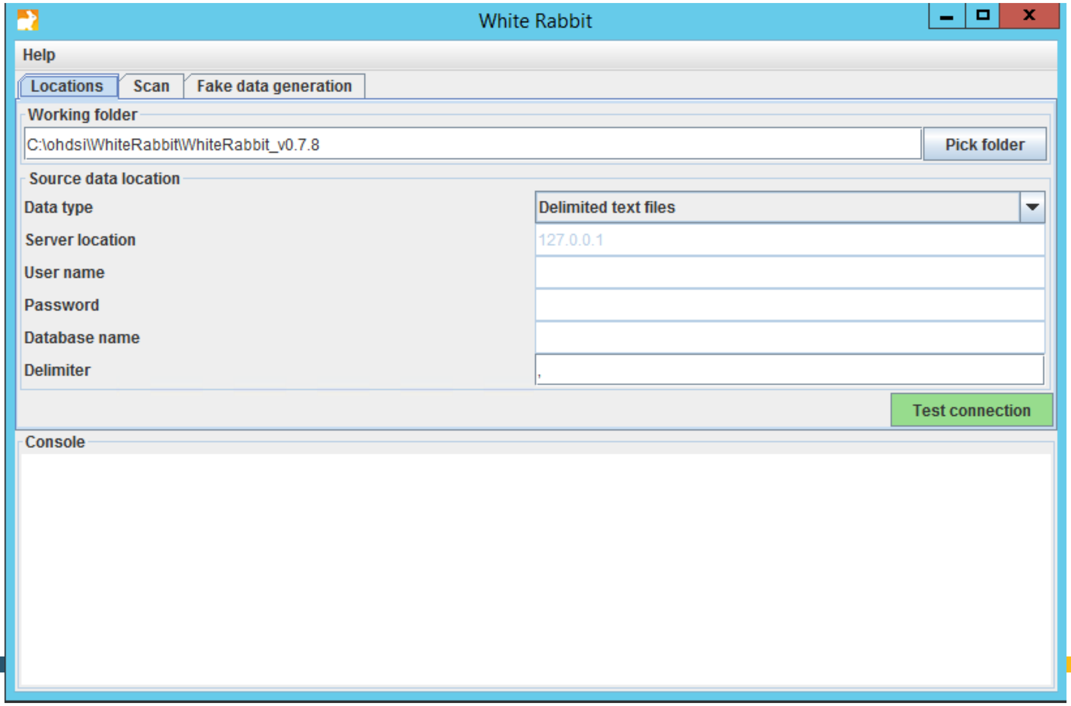


W	X	Y	Z	AA	AB	AC	AD	AE	AF
marital	Frequency	race	Frequency	ethnicity	Frequency	gender	Frequency	birthplace	Frequency
M	622	white	846	irish	235	M	572	Boston	142
	344	hispanic	112	italian	145	F	558	Springfiel	30
S	166	black	82	english	102		2	Worcester	28
		asian	70	puerto_ric	72			Lowell	22
		native	20	french	72			Brockton	21
		other	1	german	64			Cambridg	18
		Unknown	1	chinese	51			Methuen	18
				polish	49			Newton	17
				american	39			Quincy	16
				portugues	37			Framingha	16
				french_ca	35			Lynn	12
				african	33			Arlington	12
				west_indi	28			Weymout	12
				dominican	21			New Bedf	12
				american_	20			Lawrence	11
				russian	20			Haverhill	11
				scottish	19			Fitchburg	11
				asian_indi	19			Marshfiel	10
				mexican	18			Somervill	10
				swedish	17			Ba	
				central_ar	13			Fa	
				greek	12				

Patients Tab

Now Your Turn:
Scan Lauren’s Data

- Click on WhiteRabbit shortcut
- Go into the WhiteRabbit folder
- Open bin\whiteRabbit.bat



Working folder: C:\ohdsi\WhiteRabbit\WhiteRabbit_v0.7.8

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

User name:

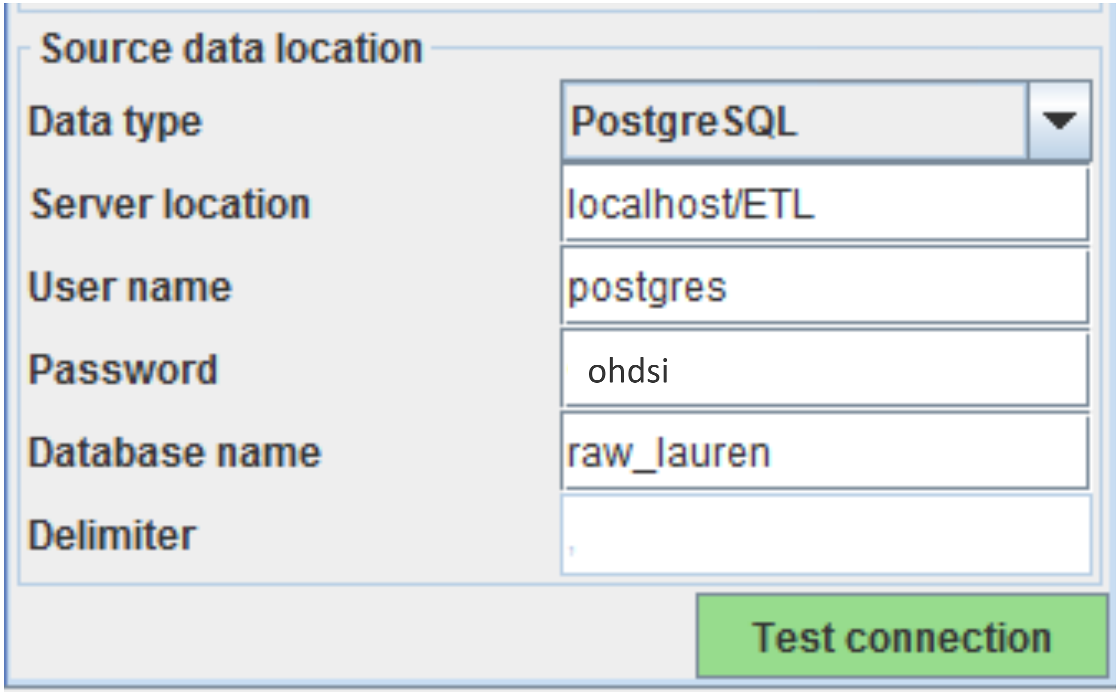
Password:

Database name:

Delimiter: ,

Now Your Turn:
Scan Lauren's Data

- Connect to Lauren's Data



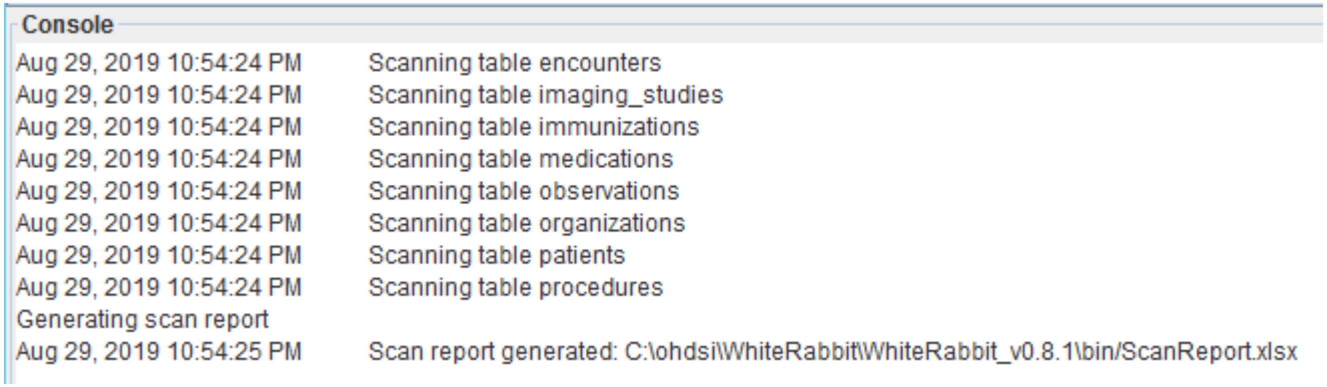
Source data location

Data type	PostgreSQL
Server location	localhost/ETL
User name	postgres
Password	ohdsi
Database name	raw_lauren
Delimiter	,

- Test connection

Now Your Turn:
Scan Lauren's Data

- Go to the "Scan" tab
- Press "Add all in DB" button, set "Min cell count" to 0, and then "Scan tables"

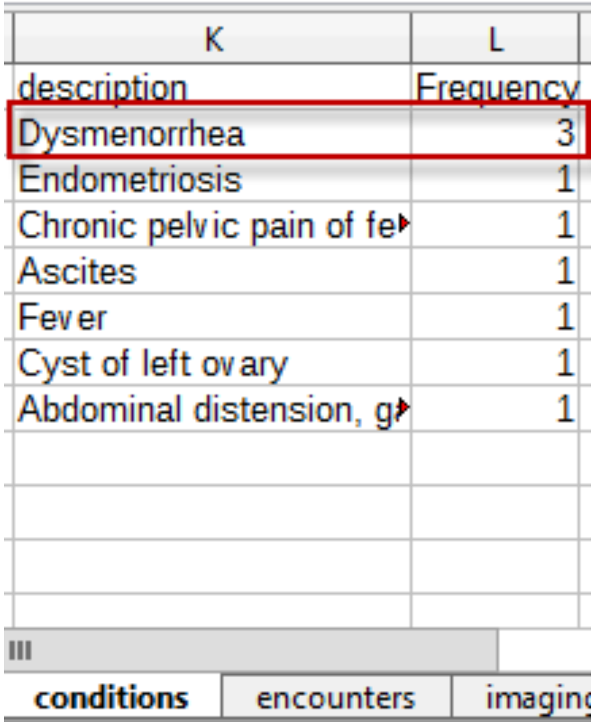


```
Console
Aug 29, 2019 10:54:24 PM     Scanning table encounters
Aug 29, 2019 10:54:24 PM     Scanning table imaging_studies
Aug 29, 2019 10:54:24 PM     Scanning table immunizations
Aug 29, 2019 10:54:24 PM     Scanning table medications
Aug 29, 2019 10:54:24 PM     Scanning table observations
Aug 29, 2019 10:54:24 PM     Scanning table organizations
Aug 29, 2019 10:54:24 PM     Scanning table patients
Aug 29, 2019 10:54:24 PM     Scanning table procedures
Generating scan report
Aug 29, 2019 10:54:25 PM     Scan report generated: C:\ohdsi\WhiteRabbit\WhiteRabbit_v0.8.1\bin/ScanReport.xlsx
```

- Open ScanReport.xlsx

Now Your Turn:
Scan Lauren's Data

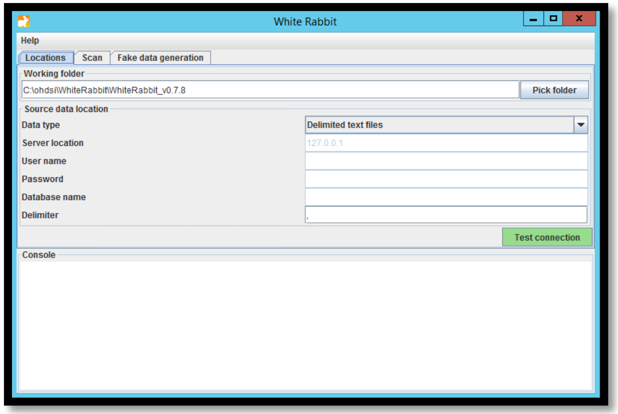
- What is the most common condition Lauren has?



K	L
description	Frequency
Dysmenorrhea	3
Endometriosis	1
Chronic pelvic pain of fe	1
Ascites	1
Fever	1
Cyst of left ovary	1
Abdominal distension, g	1

White Rabbit

- White Rabbit creates an export of information about the source data



Working folder: C:\ohdsi\WhiteRabbit\WhiteRabbit_v0.7.8

Source data location

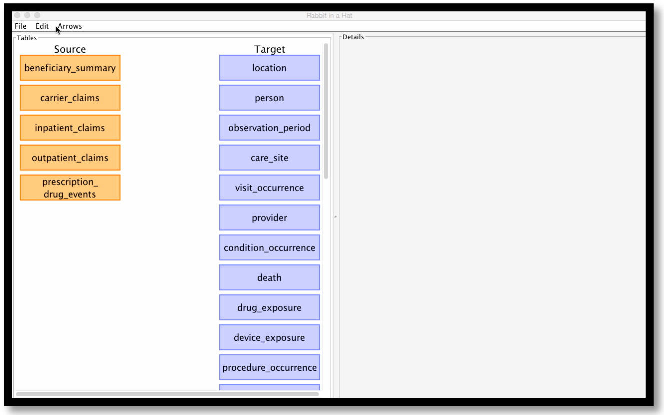
Data type	Delimited text files
Server location	127.0.0.1
User name	
Password	
Database name	
Delimiter	,

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

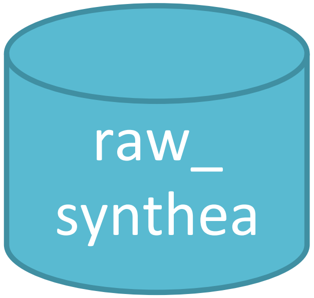
Rabbit in a Hat

- Can read and display a White Rabbit scan document

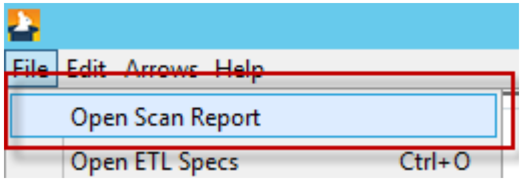
- Provides a graphical interface to allow a user to connect source data to tables



Rabbit in a Hat



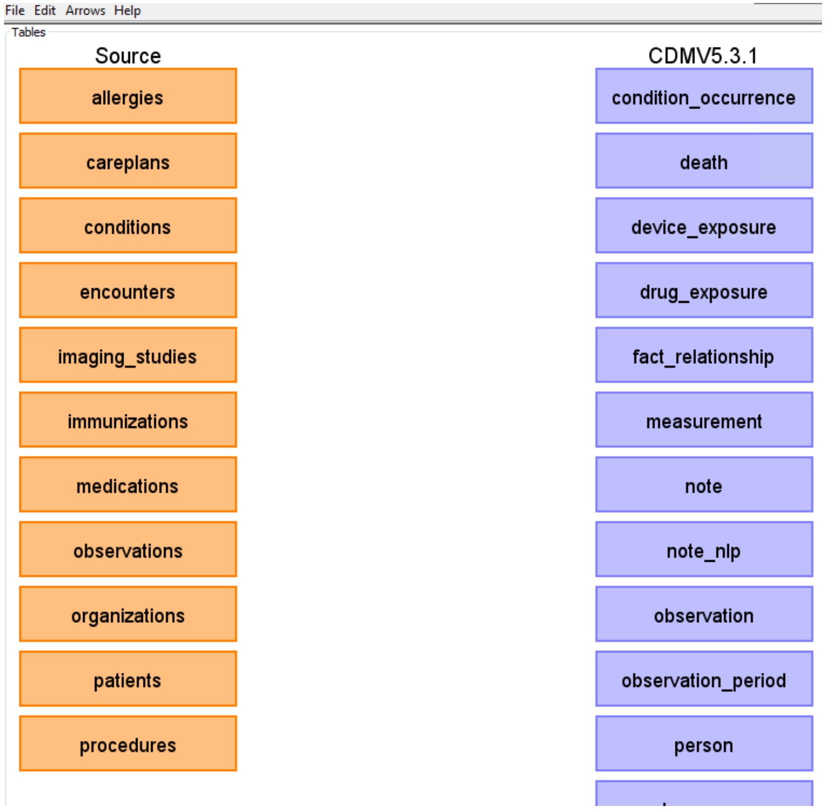
raw_ synthea



- We will use the ScanReport_raw_synthea.xlsx for this:
  - https://github.com/OHDSI/Tutorial-ETL
  - Materials → WhiteRabbit → ScanReport_raw_synthea.xlsx
  - Press the “Download” button
  - Save it to the desktop
  - Open it Rabbit in a Hat

Rabbit in a Hat

- The scan tells Rabbit in a Hat what is in the raw database
  - Orange Tables = Raw
  - Blue Tables = CDM



Source

- allergies
- careplans
- conditions
- encounters
- imaging_studies
- immunizations
- medications
- observations
- organizations
- patients
- procedures

CDMV5.3.1

- condition_occurrence
- death
- device_exposure
- drug_exposure
- fact_relationship
- measurement
- note
- note_nlp
- observation
- observation_period
- person

Rabbit in a Hat

Together

person

observation_period

condition_occurrence

On your Own

drug_exposure

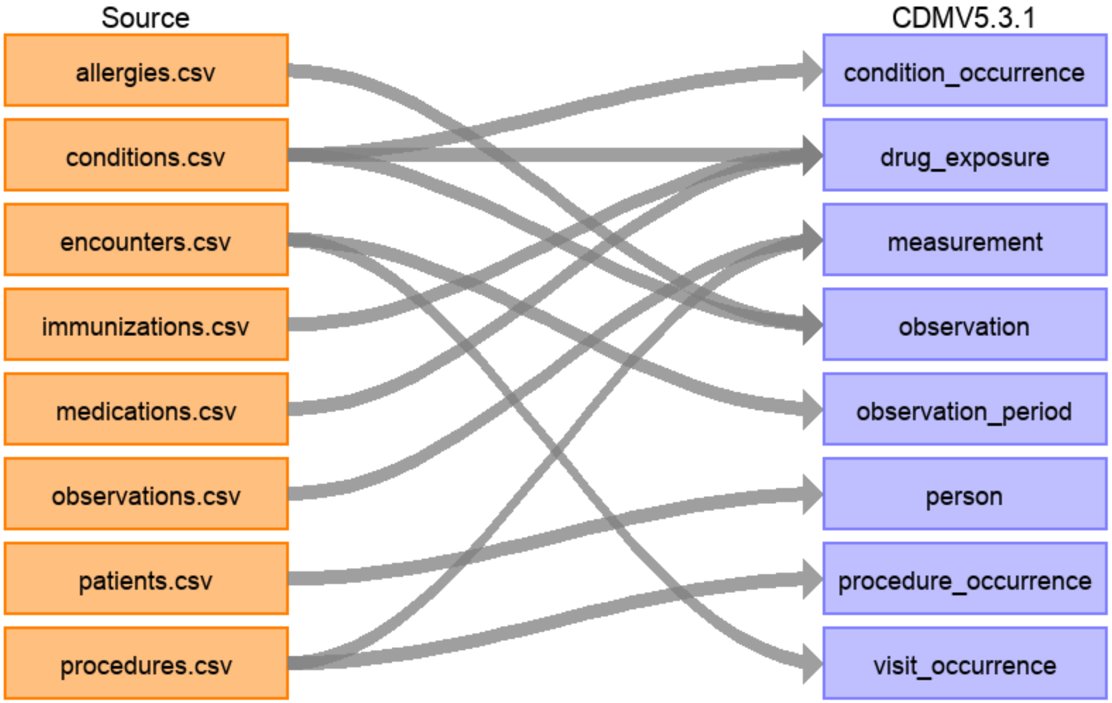
Generate document

Resources

- Important links to keep in mind when working on an ETL:
  - CDM Wiki
    https://github.com/OHDSI/CommonDataModel/wiki
    Information about the CDM structure and conventions to follow can be found here
  - OHDSI Forums
    http://forums.ohdsi.org/
    http://forums.ohdsi.org/c/cdm-builders
    OHDSI is an active community, your questions may have already been asked on the forum however if not do not be afraid to ask it yourself!
  - Book of OHDSI: ETL Chapter
    https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html
    The OHSDI community wrote the book to serve as a central knowledge repository for all things OHDSI.
  - THEMIS Working Group
    https://github.com/OHDSI/Themis

Rabbit in a Hat

- The full ETL document: https://ohdsi.github.io/ETL-Synthea/



Some Parting Thoughts On ETL

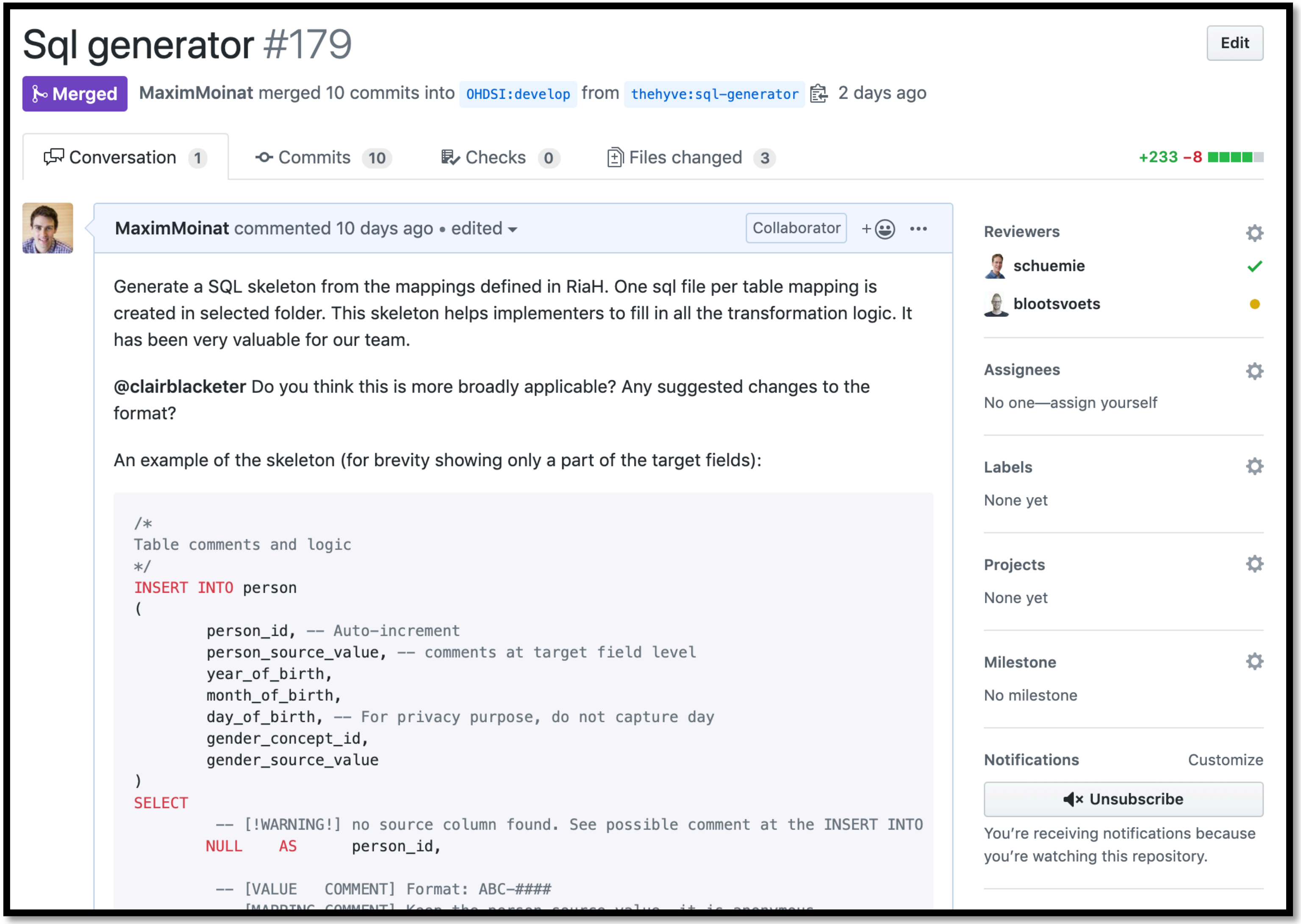
- Vocabulary will tell a source record where to go.
  - Example, just because it is a condition code and in a condition table does not mean it will end up in CONDITION_OCCURRENCE

    ICD9 783.1 - Abnormal weight gain

- STEM Table in Rabbit In a Hat

stem_table

Upcoming enhancements



Sql generator #179

Merged MaximMoinat merged 10 commits into `OHDSI:develop` from `thehyve:sql-generator` 2 days ago

MaximMoinat commented 10 days ago • edited

Generate a SQL skeleton from the mappings defined in RiaH. One sql file per table mapping is created in selected folder. This skeleton helps implementers to fill in all the transformation logic. It has been very valuable for our team.

@clairblacketer Do you think this is more broadly applicable? Any suggested changes to the format?

An example of the skeleton (for brevity showing only a part of the target fields):

```
/*
Table comments and logic
*/
INSERT INTO person
(
        person_id, -- Auto-increment
        person_source_value, -- comments at target field level
        year_of_birth,
        month_of_birth,
        day_of_birth, -- For privacy purpose, do not capture day
        gender_concept_id,
        gender_source_value
)
SELECT
         -- [!WARNING!] no source column found. See possible comment at the INSERT INTO
        NULL    AS      person_id,

         -- [VALUE   COMMENT] Format: ABC-####
            [MAPPING COMMENT] Keep the person_source_value, it is anonymous
```

Reviewers

schuemie ✓

blootsvoets ●

Assignees

No one—assign yourself

Labels

None yet

Projects

None yet

Milestone

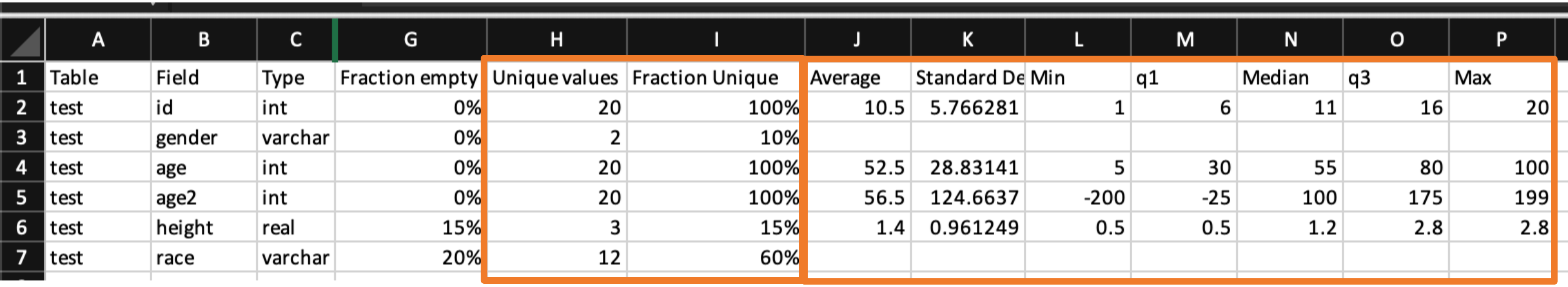
No milestone

Notifications

You're receiving notifications because you're watching this repository.

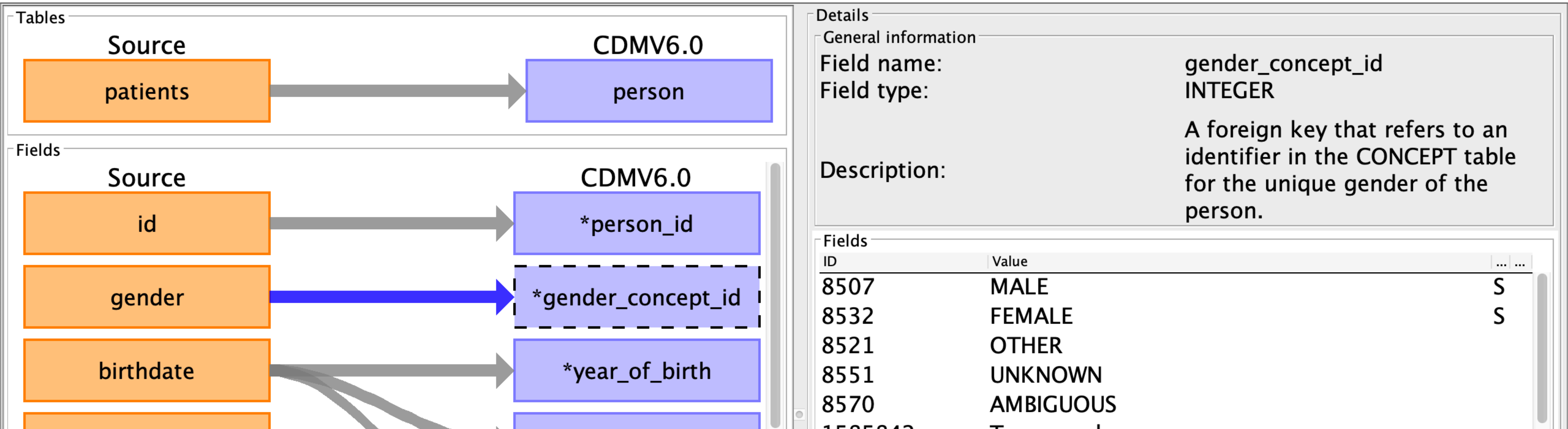
Upcoming enhancements

Additional scan report metrics



	A	B	C	G	H	I	J	K	L	M	N	O	P
1	Table	Field	Type	Fraction empty	Unique values	Fraction Unique	Average	Standard De	Min	q1	Median	q3	Max
2	test	id	int	0%	20	100%	10.5	5.766281	1	6	11	16	20
3	test	gender	varchar	0%	2	10%							
4	test	age	int	0%	20	100%	52.5	28.83141	5	30	55	80	100
5	test	age2	int	0%	20	100%	56.5	124.6637	-200	-25	100	175	199
6	test	height	real	15%	3	15%	1.4	0.961249	0.5	0.5	1.2	2.8	2.8
7	test	race	varchar	20%	12	60%							

Concept id hints



Tables

Source: patients → CDMV6.0: person

Fields

Source → CDMV6.0

id → *person_id

gender → *gender_concept_id

birthdate → *year_of_birth

Details

General information

Field name: gender_concept_id

Field type: INTEGER

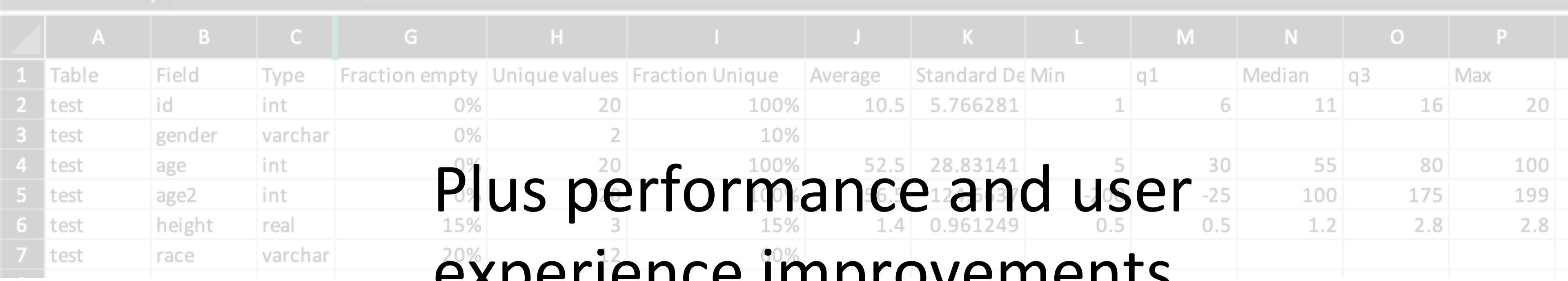
Description: A foreign key that refers to an identifier in the CONCEPT table for the unique gender of the person.

Fields

ID	Value	...	...
8507	MALE	S	
8532	FEMALE	S	
8521	OTHER		
8551	UNKNOWN		
8570	AMBIGUOUS		

Upcoming enhancements

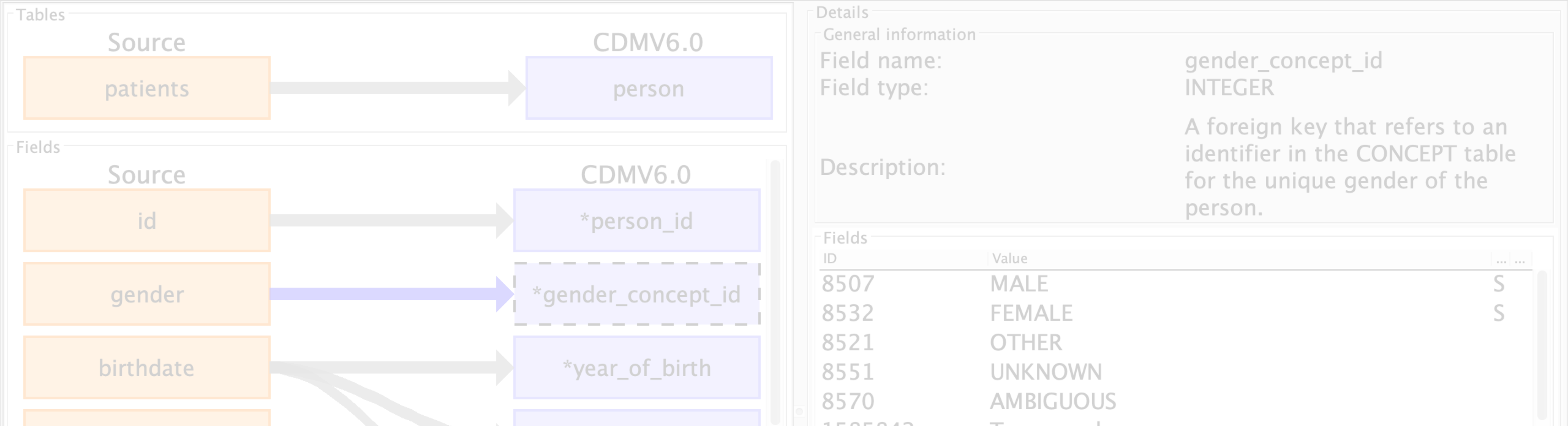
Additional scan report metrics



	A	B	C	G	H	I	J	K	L	M	N	O	P
1	Table	Field	Type	Fraction empty	Unique values	Fraction Unique	Average	Standard De	Min	q1	Median	q3	Max
2	test	id	int	0%	20	100%	10.5	5.766281	1	6	11	16	20
3	test	gender	varchar	0%	2	10%							
4	test	age	int	0%	20	100%	52.5	28.83141	5	30	55	80	100
5	test	age2	int	0%	[illegible]	100%	56.5	[illegible]	-20	-25	100	175	199
6	test	height	real	15%	3	15%	1.4	0.961249	0.5	0.5	1.2	2.8	2.8
7	test	race	varchar	20%	12	60%							

Plus performance and user experience improvements

Concept id hints



Details

General information

Field name: gender_concept_id

Field type: INTEGER

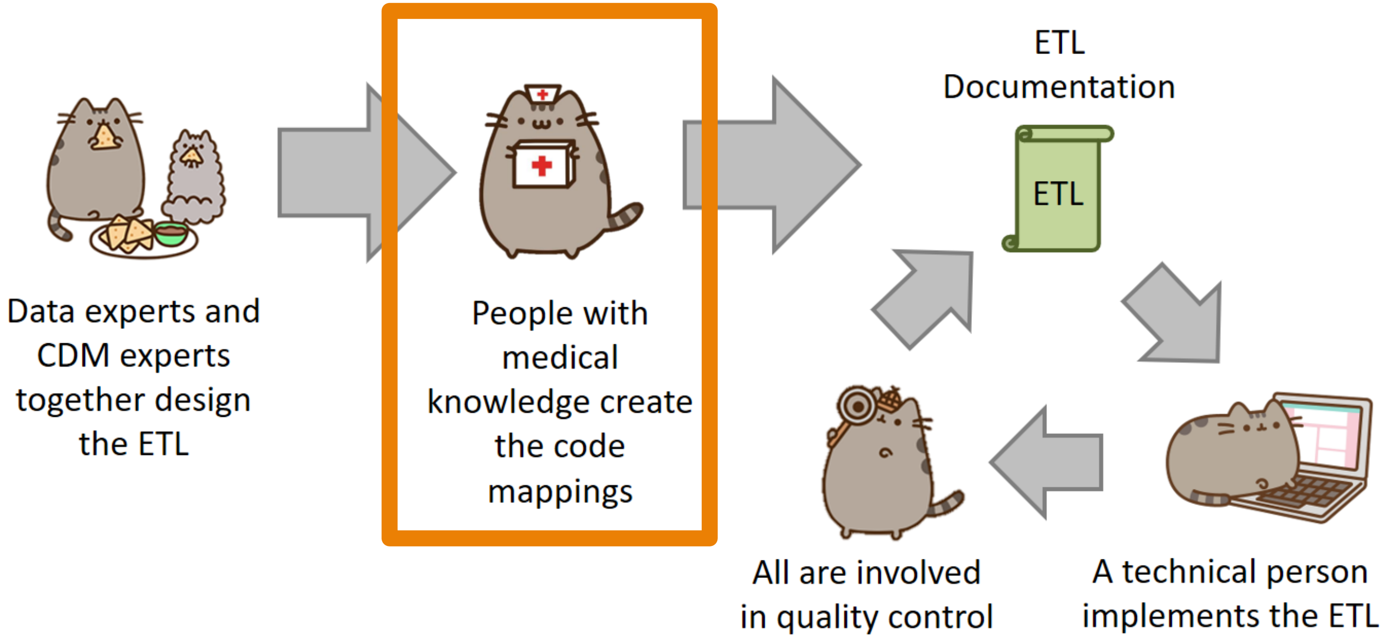
Description: A foreign key that refers to an identifier in the CONCEPT table for the unique gender of the person.

Fields

ID	Value	
8507	MALE	S
8532	FEMALE	S
8521	OTHER	
8551	UNKNOWN	
8570	AMBIGUOUS	

OHDSI

Observational Health Data Sciences and Informatics



Data experts and CDM experts together design the ETL

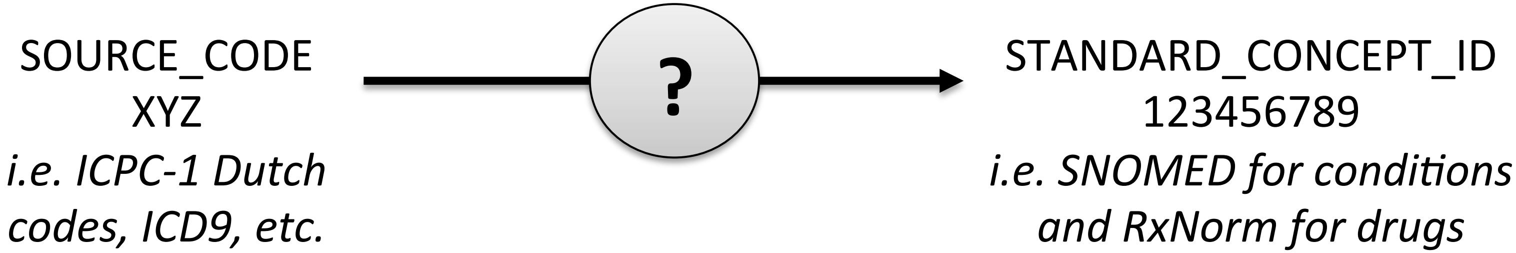
People with medical knowledge create the code mappings

ETL Documentation

A technical person implements the ETL

All are involved in quality control

Standardizing Terminologies



SOURCE_CODE
XYZ
i.e. ICPC-1 Dutch codes, ICD9, etc.

STANDARD_CONCEPT_ID
123456789
i.e. SNOMED for conditions and RxNorm for drugs

- What is standardize:
  1. TABLE_CONCEPT_ID
     standard concept the source code maps to, **used for analysis**
  2. TABLE_SOURCE_CONCEPT_ID
     concept representation of the source code, **helps maintain tie to raw data**

- Ways to get a source code to standard code:
  1. OMOP Vocabulary
  2. USAGI

OMOP Vocab

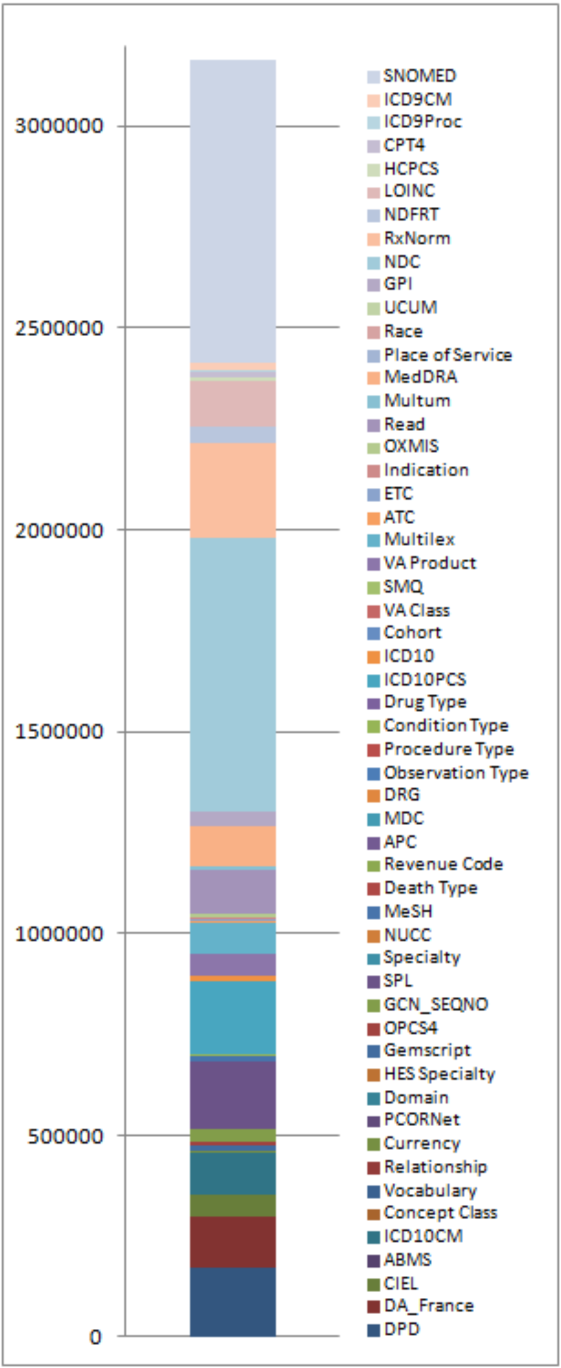
- There are two standard queries to help us use the OMOP Vocabulary:
  - `SOURCE_TO_STANDARD.sql`
  - `SOURCE_TO_SOURCE.sql`

- https://github.com/OHDSI/Tutorial-ETL
  - Materials → Queries

OMOP Vocab

- If your source data's codes are in the OMOP Vocab you can use it to translate to a standard

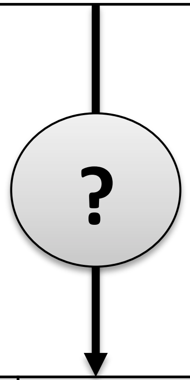
- For example:
  - ICD9 → SNOMED
  - NDC → RxNORM



Mapping a Lauren Row to CONCEPT_ID

```
SELECT *
FROM RAW_LAUREN.CONDITIONS
WHERE ENCOUNTER = '70'
```

START	STOP	PATIENT	ENCOUNTER	CODE	DESCRIPTION
1/6/2010		1	70	N94.6	Dysmenorrhea



CONDITION_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID

Source to Standard

```
WITH CTE_VOCAB_MAP AS (
      SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS SOURCE_CODE_DESCRIPTION,
c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS SOURCE_INVALID_REASON,
c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS TARGET_VOCABUALRY_ID,
c1.domain_id AS TARGET_DOMAIN_ID, c1.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c1.INVALID_REASON AS TARGET_INVALID_REASON,
c1.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM CONCEPT C
             JOIN CONCEPT_RELATIONSHIP CR
                        ON C.CONCEPT_ID = CR.CONCEPT_ID_1
                        AND CR.invalid_reason IS NULL
                        AND cr.relationship_id = 'Maps to'
              JOIN CONCEPT C1
                        ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID
                        AND C1.INVALID_REASON IS NULL
       UNION

SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS
SOURCE_CONCEPT_CLASS_ID,c1.VALID_START_DATE AS SOURCE_VALID_START_DATE,
c1.VALID_END_DATE AS SOURCE_VALID_END_DATE, stcm.INVALID_REASON AS SOURCE_INVALID_REASON,target_concept_id,
c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME, target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID,
c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c2.INVALID_REASON AS TARGET_INVALID_REASON,
c2.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM source_to_concept_map stcm
              LEFT OUTER JOIN CONCEPT c1
                     ON c1.concept_id = stcm.source_concept_id
              LEFT OUTER JOIN CONCEPT c2
                     ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT TARGET_CONCEPT_ID, TARGET_CONCEPT_NAME, TARGET_DOMAIN_ID
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
AND TARGET_STANDARD_CONCEPT = 'S'
```

Source to Standard

```
WITH CTE_VOCAB_MAP AS (
      SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS SOURCE_CODE_DESCRIPTION,
c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS SOURCE_INVALID_REASON,
c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS TARGET_VOCABUALRY_ID,
c1.domain_id AS TARGET_DOMAIN_ID, c1.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c1.INVALID_REASON AS TARGET_INVALID_REASON,
c1.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM CONCEPT C
             JOIN CONCEPT_RELATIONSHIP CR
                        ON C.CONCEPT_ID = CR.CONCEPT_ID_1
                        AND CR.invalid_reason IS NULL
                        AND cr.relationship_id = 'Maps to'
              JOIN CONCEPT C1
                        ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID
                        AND C1.INVALID_REASON IS NULL

SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CO                                                      OURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS
SOURCE_CONCEPT_CLASS_ID,c1.VALID_START_DATE AS
c1.VALID_END_DATE AS SOURCE_VALID_END_DATE, stc                                                       id,
c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME, target_
c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID,
c2.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM source_to_concept_map stcm
              LEFT OUTER JOIN CONCEPT c1
                     ON c1.concept_id = stcm.so
              LEFT OUTER JOIN CONCEPT c2
                     ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT TARGET_CONCEPT_ID, TARGET_CONCEPT_NAME, TARGET_DOMAIN_ID
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
AND TARGET_STANDARD_CONCEPT = 'S'
```

Look in the
OMOP Vocabulary for a map

Source to Standard

```
WITH CTE_VOCAB_MAP AS (
      SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS SOURCE_CODE_DESCRIPTION,
c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS SOURCE_INVALID_REASON,
c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS TARGET_VOCABUALRY_ID,
c1.domain_id AS TARGET_DOMAIN_ID, c1.concept_c                                    S TARGET_INVALID_REASON,
c1.standard_concept AS TARGET_STANDARD_CONCEP
       FROM CONCEPT C
             JOIN CONCEPT_RELATIONSHIP CR
                        ON C.CONCEPT_ID = CR.C
                        AND CR.invalid_reason
                        AND cr.relationship_i
              JOIN CONCEPT C1
                        ON CR.CONCEPT_ID_2 = C
                        AND C1.INVALID_REASON IS NULL
       UNION

SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS
SOURCE_CONCEPT_CLASS_ID,c1.VALID_START_DATE AS SOURCE_VALID_START_DATE,
c1.VALID_END_DATE AS SOURCE_VALID_END_DATE, stcm.INVALID_REASON AS SOURCE_INVALID_REASON,target_concept_id,
c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME, target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID,
c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c2.INVALID_REASON AS TARGET_INVALID_REASON,
c2.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM source_to_concept_map stcm
              LEFT OUTER JOIN CONCEPT c1
                     ON c1.concept_id = stcm.source_concept_id
              LEFT OUTER JOIN CONCEPT c2
                     ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL

SELECT TARGET_CONCEPT_ID, TARGET_CONCEPT_NAME, TARGET_DOMAIN_ID
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
AND TARGET_STANDARD_CONCEPT = 'S'
```

Look in the Source to Concept Map table for a map

Source to Standard

```
WITH CTE_VOCAB_MAP AS (
      SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS SOURCE_CODE_DESCRIPTION,
c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS SOURCE_INVALID_REASON,
c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS TARGET_VOCABUALRY_ID,
c1.domain_id AS TARGET_DOMAIN_ID, c1.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c1.INVALID_REASON AS TARGET_INVALID_REASON,
c1.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM CONCEPT C
             JOIN CONCEPT_RELATIONSHIP CR
                        ON C.CONCEPT_ID = CR.CONCEPT_ID_1
                        AND CR.invalid_reason IS NULL
                        AND cr.relationship_id = 'Maps to'
              JOIN CONCEPT C1
                        ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID
                        AND C1.INVALID_REASON IS NULL
       UNION

SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS
SOUR                                                         _START_DATE,
c1.                                                          SON AS SOURCE_INVALID_REASON,target_concept_id,
c2.                                                          , c2.domain_id AS TARGET_DOMAIN_ID,
c2.                                                          EASON AS TARGET_INVALID_REASON,
c2.



                                                             id

                        ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT TARGET_CONCEPT_ID, TARGET_CONCEPT_NAME, TARGET_DOMAIN_ID
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
AND TARGET_STANDARD_CONCEPT = 'S'
```

Look up your source Code here

Mapping a Lauren Row to CONCEPT_ID:
Source to Standard

START	STOP	PATIENT	ENCOUNTER	CODE	DESCRIPTION
1/6/2010		1	70	N94.6	Dysmenorrhea

TARGET_ CONCEPT_ID	TARGET_ CONCEPT_NAME	TARGET_ DOMAIN_ID
194696	Dysmenorrhea	Condition

CONDITION_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID
194696	

Source to Source

```
WITH CTE_VOCAB_MAP AS (
       SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID,
             c.CONCEPT_NAME AS SOURCE_CODE_DESCRIPTION, c.vocabulary_id AS SOURCE_VOCABULARY_ID,
             c.domain_id AS SOURCE_DOMAIN_ID, c.concept_class_id AS SOURCE_CONCEPT_CLASS_ID,
             c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE,
             c.invalid_reason AS SOURCE_INVALID_REASON, c.concept_ID as TARGET_CONCEPT_ID,
             c.concept_name AS TARGET_CONCEPT_NAME, c.vocabulary_id AS TARGET_VOCABULARY_ID,
             c.domain_id AS TARGET_DOMAIN_ID, c.concept_class_id AS TARGET_CONCEPT_CLASS_ID,
             c.INVALID_REASON AS TARGET_INVALID_REASON, c.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
       FROM CONCEPT c

       UNION

       SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id,
             c1.domain_id AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
             c1.VALID_START_DATE AS SOURCE_VALID_START_DATE, c1.VALID_END_DATE AS SOURCE_VALID_END_DATE,
             stcm.INVALID_REASON AS SOURCE_INVALID_REASON,target_concept_id,
             c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME, target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID,
             c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c2.INVALID_REASON AS TARGET_INVALID_REASON,
             c2.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM source_to_concept_map stcm
              LEFT OUTER JOIN CONCEPT c1
                     ON c1.concept_id = stcm.source_concept_id
              LEFT OUTER JOIN CONCEPT c2
                     ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT *
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
```

Source to Source

```
WITH CTE_VOCAB_MAP AS (
       SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID,
             c.CONCEPT_NAME AS SOURCE_CODE_DESCRIPTION, c.vocabulary_id AS SOURCE_VOCABULARY_ID,
             c.domain_id AS SOURCE_DOMAIN_ID, c.concept_class_id AS SOURCE_CONCEPT_CLASS_ID,
             c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE,
             c.invalid_reason AS SOURCE_INVALID_REASON, c.concept_ID as TARGET_CONCEPT_ID,
             c.concept_name AS TARGET_CONCEPT_NAME, c.vocabulary_id AS TARGET_VOCABULARY_ID,
             c.domain_id AS TARGET_DOMAIN_ID, c.concept_class_id AS TARGET_CONCEPT_CLASS_ID,
             c.INVALID_REASON AS TARGET_INVALID_REASON, c.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
       FROM CONCEPT c

       UNION

       SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id,
             c1.domain_id AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
             c1.VALID_START_DATE AS SOURCE_VALID_START_DATE, c1.VALID_END_DATE AS SOURCE_VALID_END_DATE,
                                                      ALID_REASON,target_concept_id,
                                                      AME, target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID,
                                                      PT_CLASS_ID, c2.INVALID_REASON AS TARGET_INVALID_REASON,
                                                      ARD_CONCEPT


                                                      rce_concept_id


                                                      get_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT *
FROM CTE_VOCAB_MAP
WHERE SOURCE_CODE = 'N94.6'
AND SOURCE_VOCABULARY_ID = 'ICD10CM'
```

Look up your source Code here

Mapping a Lauren Row to CONCEPT_ID: Source to Source

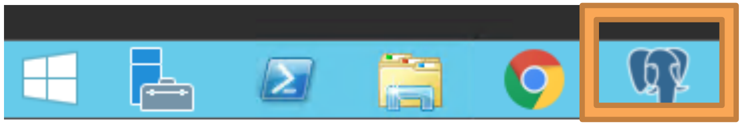
START	STOP	PATIENT	ENCOUNTER	CODE	DESCRIPTION
1/6/2010		1	70	N94.6	Dysmenorrhea

TARGET_CONCEPT_ID	TARGET_CONCEPT_NAME	TARGET_DOMAIN_ID
35209488	Dysmenorrhea, unspecified	Condition

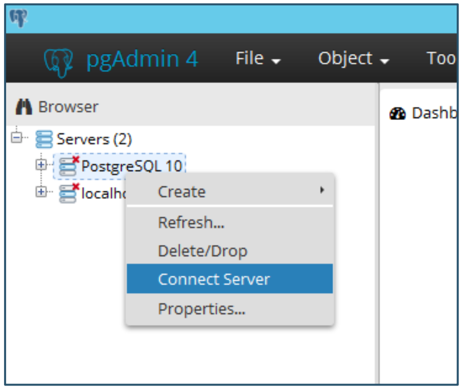
CONDITION_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID
194696	35209488

Mapping Source Codes – Your turn

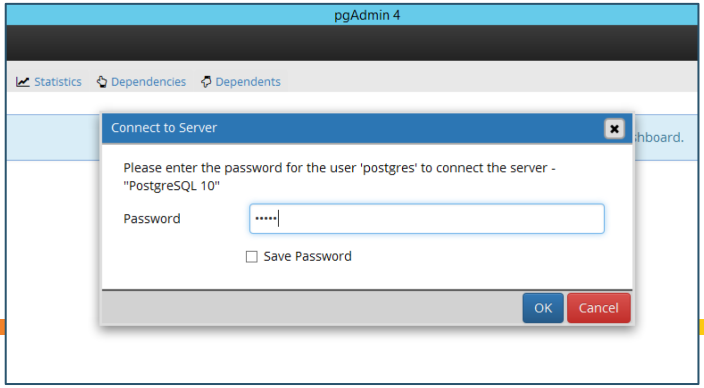
- Let’s open PostgreSQL
  - Open up pgAdmin4 using the icon on the task bar



  - Expand the server list and right-click on PostgreSQL 10 and choose Connect Server from the drop-down menu



  - When it asks for a password, type in ohdsi



Please enter the password for the user 'postgres' to connect the server - "PostgreSQL 10"

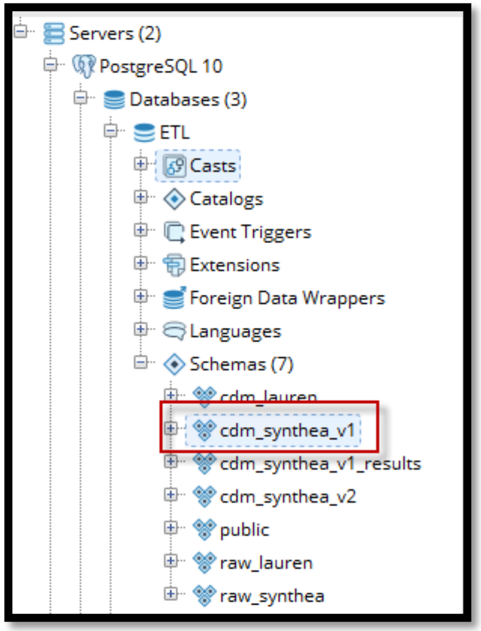
Password •••••

□ Save Password

Mapping Source Codes – Your turn

- Open up to and select the CDM (which has a copy of the vocab)
- Tools → Query Tool
- Type the following and hit F5 to run:

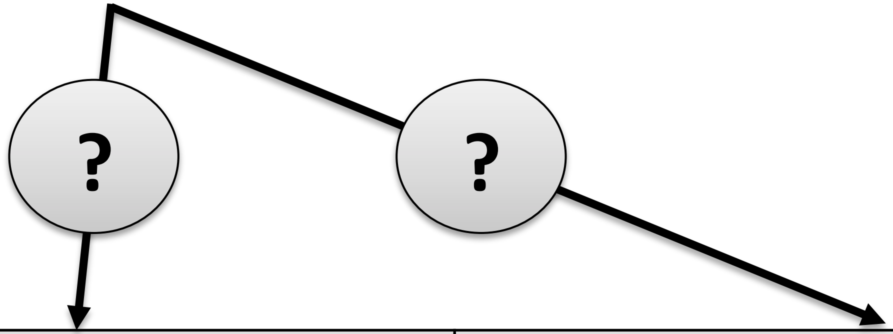
```
SET SEARCH_PATH TO
CDM_SYNTHEA_V1;
```



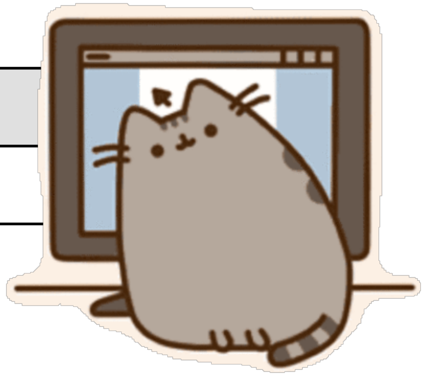
- Servers (2)
  - PostgreSQL 10
    - Databases (3)
      - ETL
        - Casts
        - Catalogs
        - Event Triggers
        - Extensions
        - Foreign Data Wrappers
        - Languages
        - Schemas (7)
          - cdm_lauren
          - cdm_synthea_v1
          - cdm_synthea_v1_results
          - cdm_synthea_v2
          - public
          - raw_lauren
          - raw_synthea

Mapping Source Codes – Your turn

CODE	DESCRIPTION	CODE TYPE
C83.3	Diffuse large B-cell lymphoma	ICD10 (not ICD10CM)



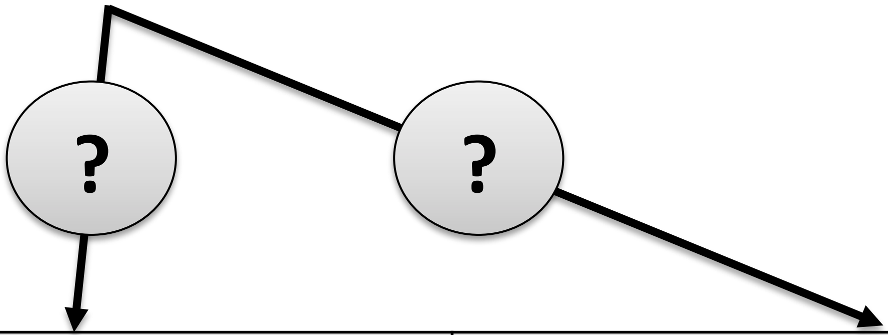
CONDITION_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID



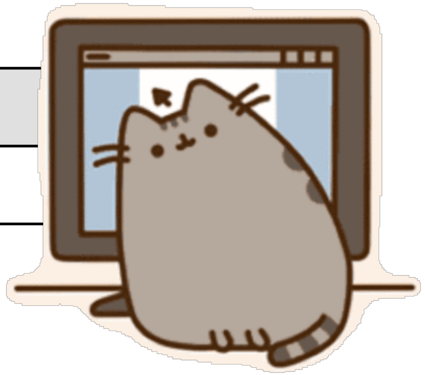
https://github.com/OHDSI/Tutorial-ETL/tree/master/materials/Queries

Mapping Source Codes – Your turn

CODE	DESCRIPTION	CODE TYPE
C83.3	Diffuse large B-cell lymphoma	ICD10 (not ICD10CM)

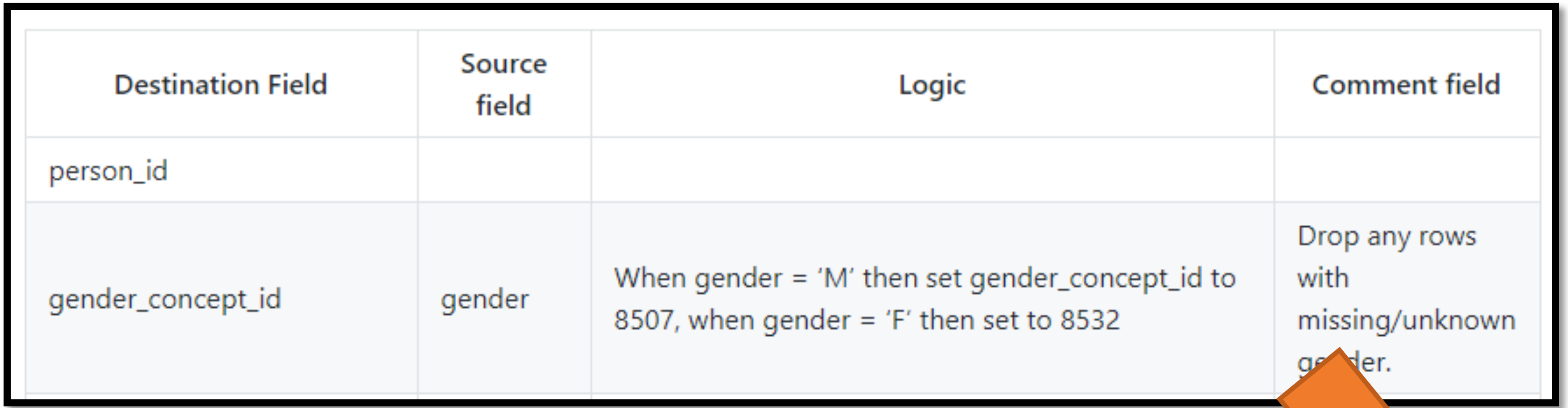


CONDITION_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID
4300704	1567654

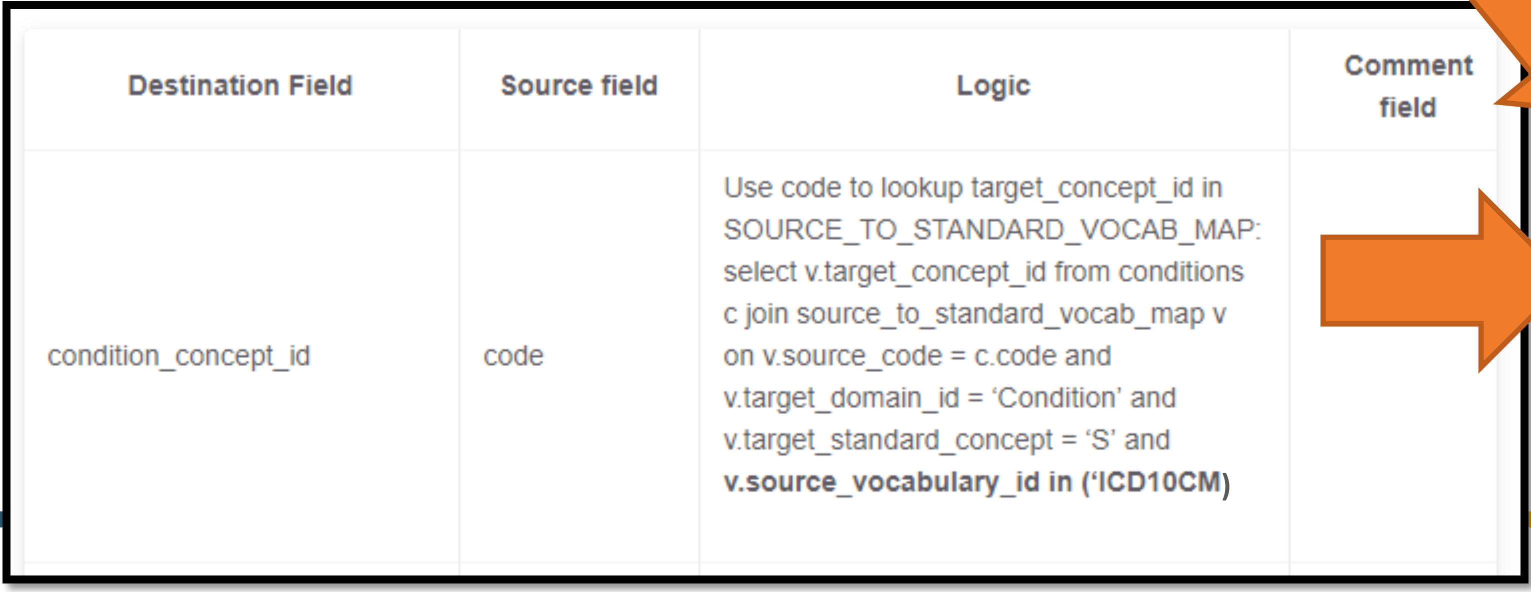


https://github.com/OHDSI/Tutorial-ETL/tree/master/materials/Queries

What do you do with the mapping information?



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 ('ICD10CM)**	

Usagi

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

- Usagi is Japanese for rabbit and was named after the first mapping exercise it was used for; mapping source codes used in a Japanese dataset into OMOP Vocabulary concepts

- Usagi software tool to help with process of mapping source codes to OMOP concepts

- Usagi performs text similarity between your source codes and what is in the OMOP Vocabulary

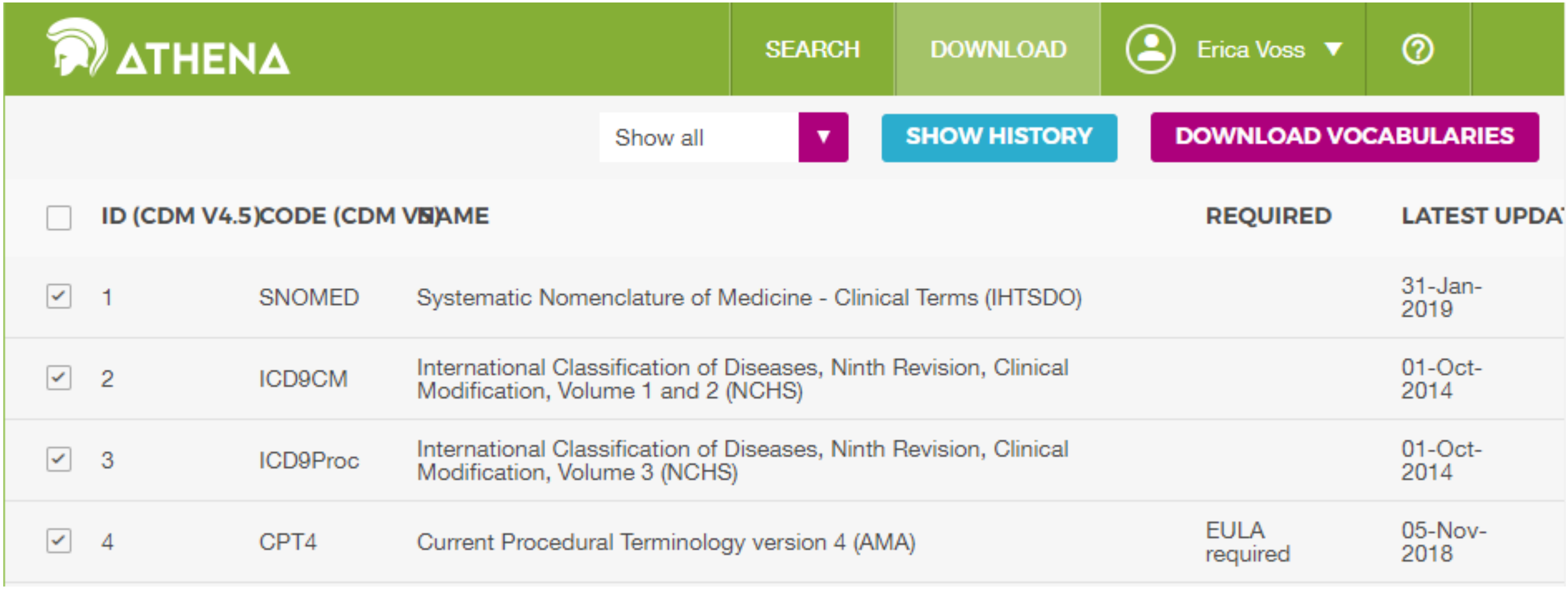
Usagi Process

1. Get a copy of the **Vocabulary** from ATHENA
2. Download **Usagi**
3. Have Usagi **build an index** on the Vocabulary
4. **Load your source codes** and let Usagi process them
5. **Review and update suggest mappings** with someone who has medical knowledge
6. **Export codes** into the SOURCE_TO_CONCEPT_MAP

Usagi Process

*1. Get a copy of the **Vocabulary** from ATHENA*

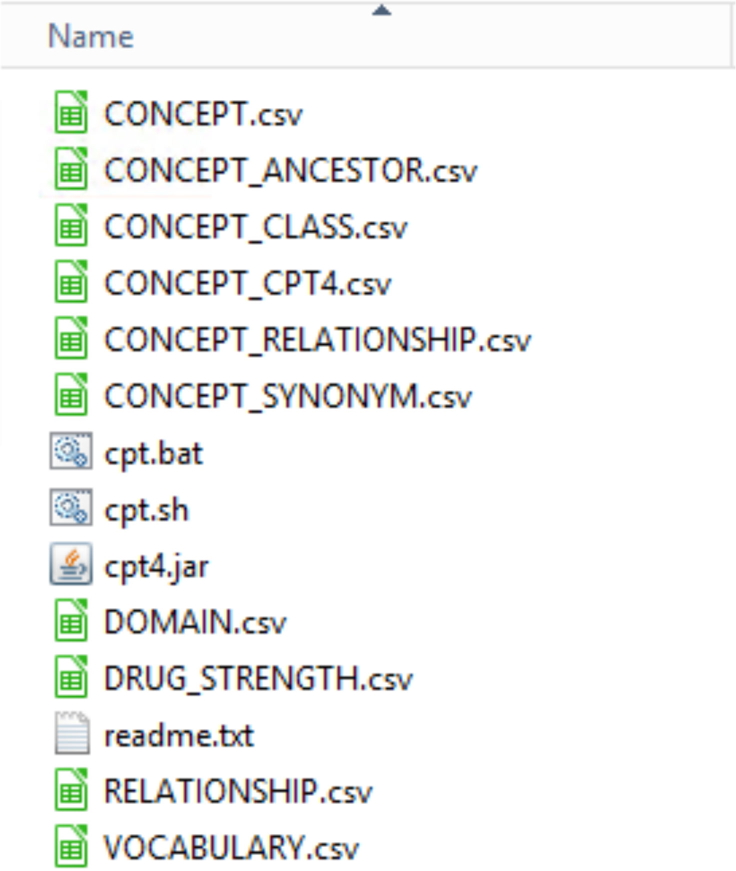
http://athena.ohdsi.org



□	ID (CDM V4.5)	CODE (CDM V5)	NAME	REQUIRED	LATEST UPDA
☑	1	SNOMED	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO)		31-Jan-2019
☑	2	ICD9CM	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 1 and 2 (NCHS)		01-Oct-2014
☑	3	ICD9Proc	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 3 (NCHS)		01-Oct-2014
☑	4	CPT4	Current Procedural Terminology version 4 (AMA)	EULA required	05-Nov-2018

Usagi Process

*1. Get a copy of the **Vocabulary** from ATHENA*



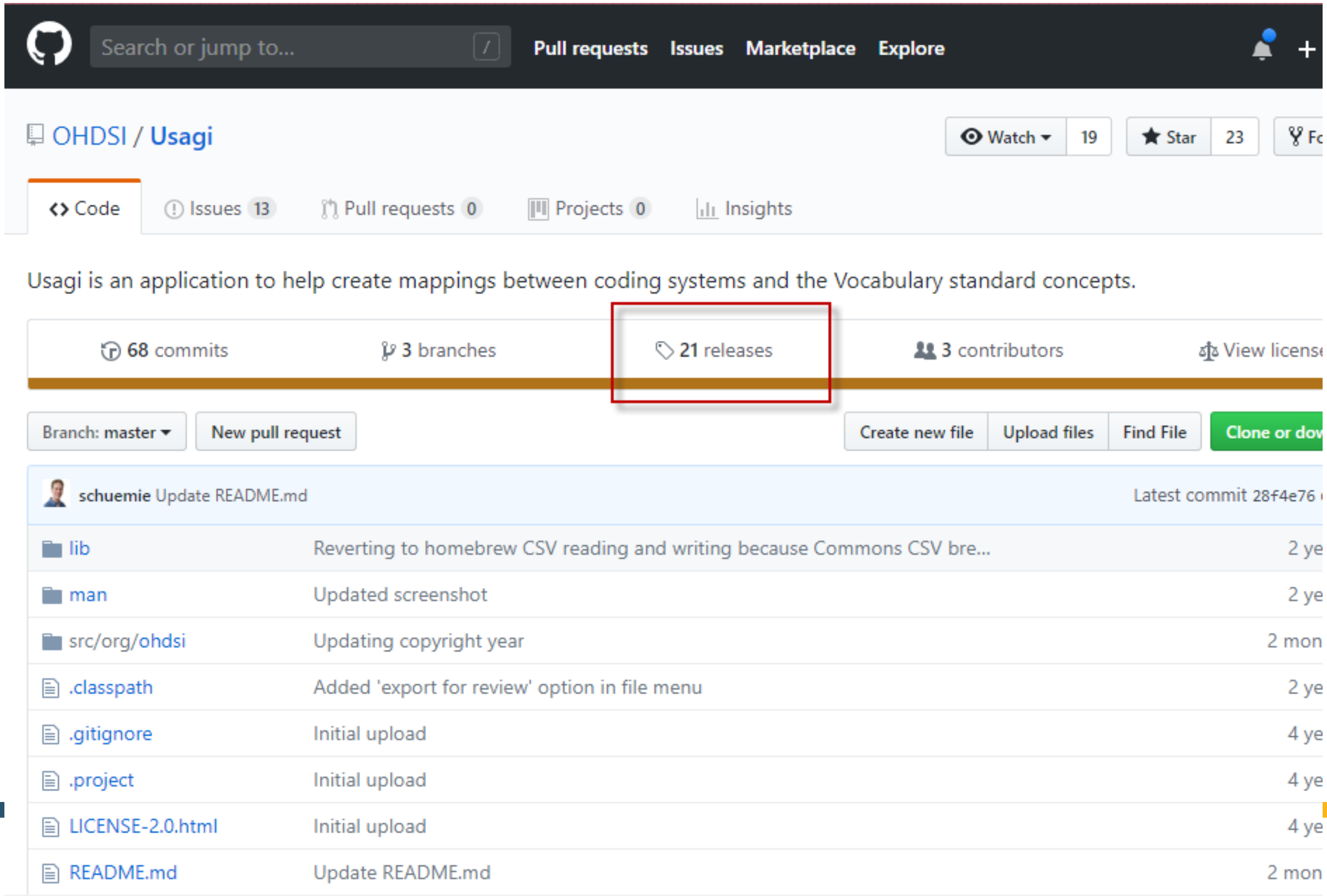
Name

- CONCEPT.csv
- CONCEPT_ANCESTOR.csv
- CONCEPT_CLASS.csv
- CONCEPT_CPT4.csv
- CONCEPT_RELATIONSHIP.csv
- CONCEPT_SYNONYM.csv
- cpt.bat
- cpt.sh
- cpt4.jar
- DOMAIN.csv
- DRUG_STRENGTH.csv
- readme.txt
- RELATIONSHIP.csv
- VOCABULARY.csv

Usagi Process

*2. Download **Usagi***

https://github.com/OHDSI/Usagi



OHDSI / Usagi

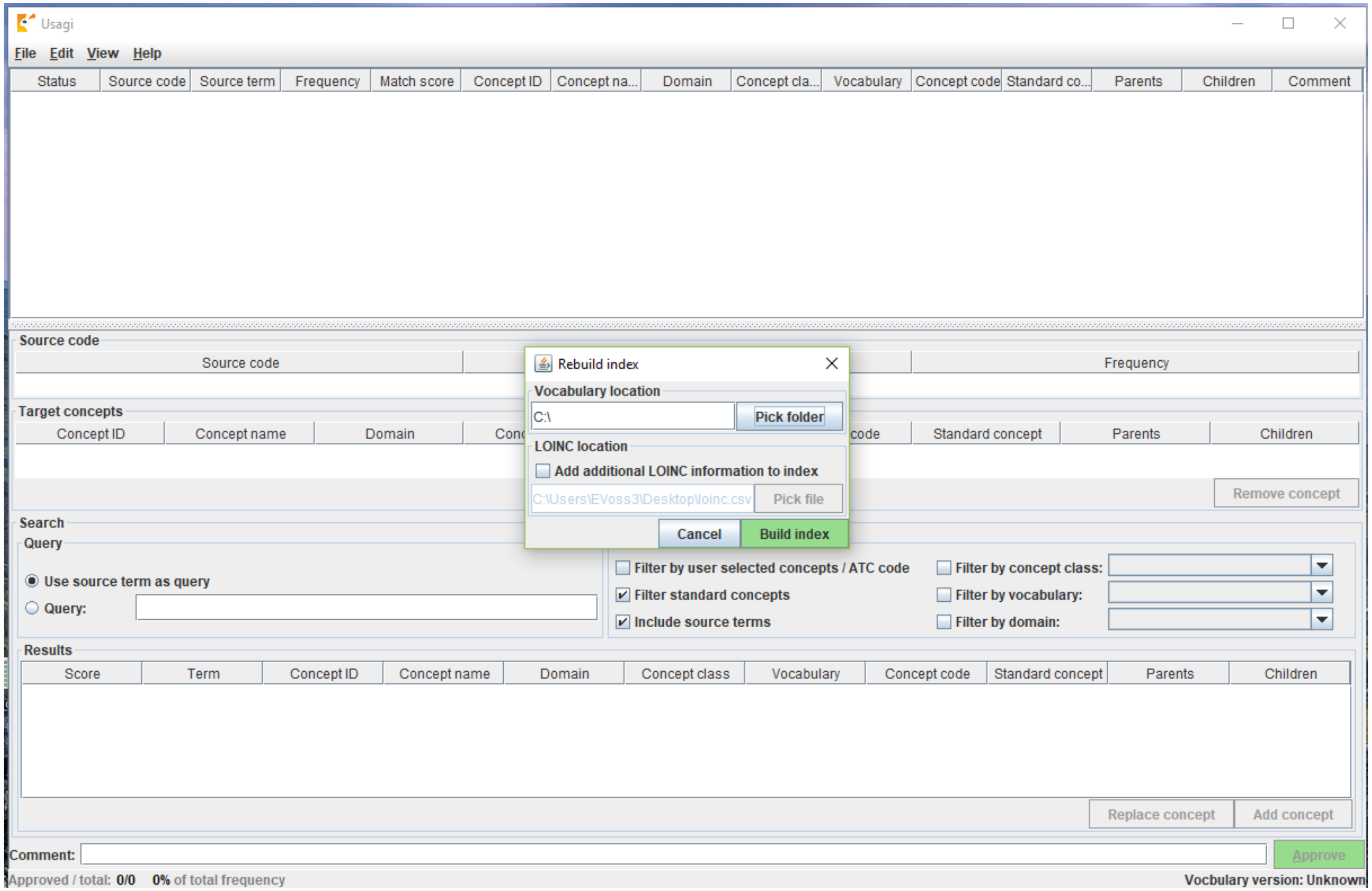
Usagi is an application to help create mappings between coding systems and the Vocabulary standard concepts.

schuemie Update README.md — Latest commit 28f4e76

lib	Reverting to homebrew CSV reading and writing because Commons CSV bre...	2 ye
man	Updated screenshot	2 ye
src/org/ohdsi	Updating copyright year	2 mon
.classpath	Added 'export for review' option in file menu	2 ye
.gitignore	Initial upload	4 ye
.project	Initial upload	4 ye
LICENSE-2.0.html	Initial upload	4 ye
README.md	Update README.md	2 mon

Usagi Process

*3. Have Usagi **build an index** on the Vocabulary*



Vocabulary location: C:\

LOINC location: ☐ Add additional LOINC information to index

C:\Users\EVoss3\Desktop\loinc.csv

● Use source term as query

○ Query:

☐ Filter by user selected concepts / ATC code

☑ Filter standard concepts

☑ Include source terms

☐ Filter by concept class:

☐ Filter by vocabulary:

☐ Filter by domain:

Comment:

Approved / total: 0/0 0% of total frequency

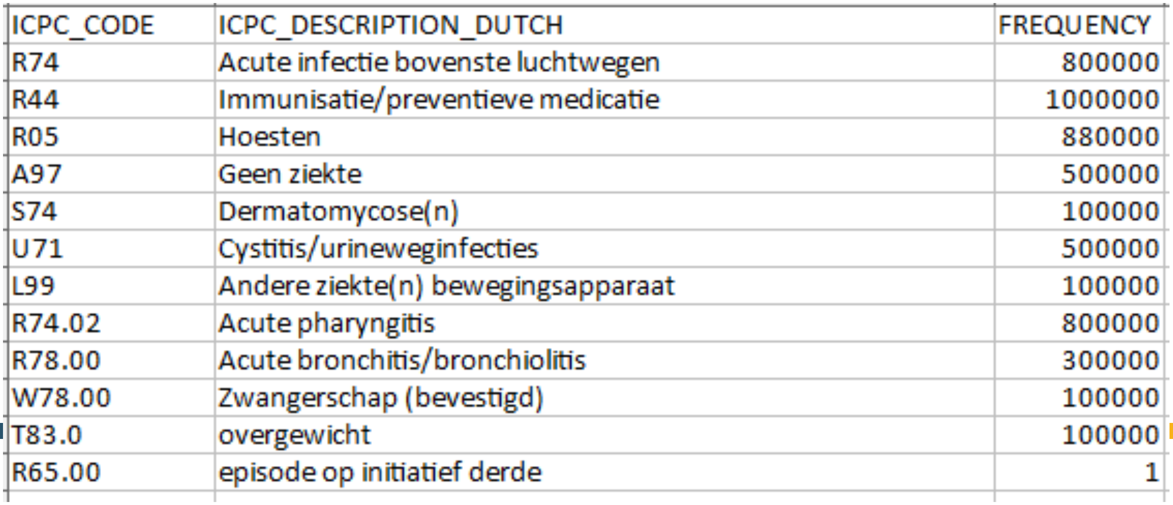
Vocbulary version: Unknown

Usagi Process

4. ***Load your source codes***, *let Usagi process them*

- Generate an XLSX of **distinct codes** from source system with descriptions and frequency

- If the codes are not in English, use Google Translate to convert

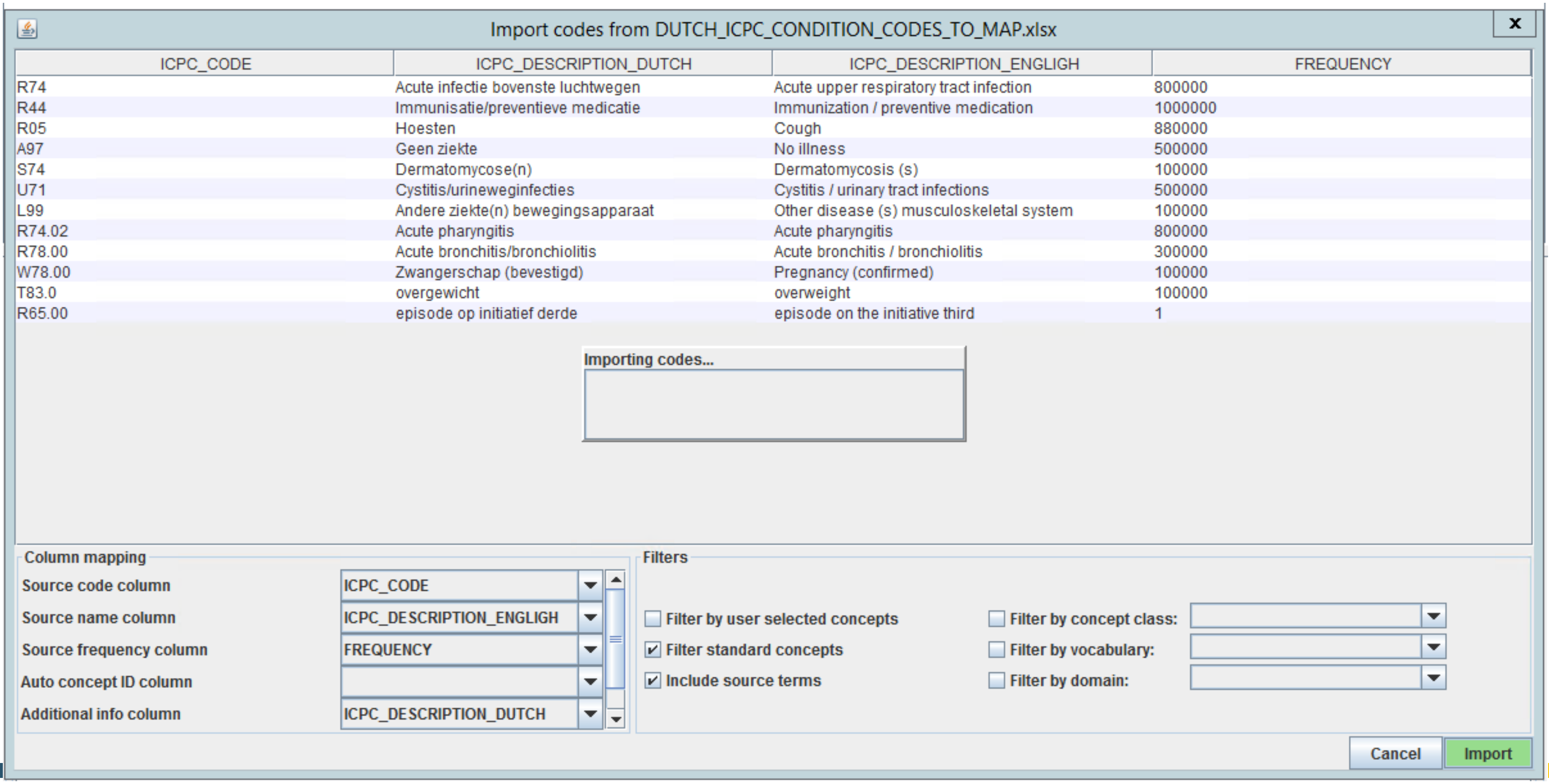


ICPC_CODE	ICPC_DESCRIPTION_DUTCH	FREQUENCY
R74	Acute infectie bovenste luchtwegen	800000
R44	Immunisatie/preventieve medicatie	1000000
R05	Hoesten	880000
A97	Geen ziekte	500000
S74	Dermatomycose(n)	100000
U71	Cystitis/urineweginfecties	500000
L99	Andere ziekte(n) bewegingsapparaat	100000
R74.02	Acute pharyngitis	800000
R78.00	Acute bronchitis/bronchiolitis	300000
W78.00	Zwangerschap (bevestigd)	100000
T83.0	overgewicht	100000
R65.00	episode op initiatief derde	1

Usagi Process

4. ***Load your source codes***, *let Usagi process them*

- Import the codes into Usagi



Import codes from DUTCH_ICPC_CONDITION_CODES_TO_MAP.xlsx

ICPC_CODE	ICPC_DESCRIPTION_DUTCH	ICPC_DESCRIPTION_ENGLIGH	FREQUENCY
R74	Acute infectie bovenste luchtwegen	Acute upper respiratory tract infection	800000
R44	Immunisatie/preventieve medicatie	Immunization / preventive medication	1000000
R05	Hoesten	Cough	880000
A97	Geen ziekte	No illness	500000
S74	Dermatomycose(n)	Dermatomycosis (s)	100000
U71	Cystitis/urineweginfecties	Cystitis / urinary tract infections	500000
L99	Andere ziekte(n) bewegingsapparaat	Other disease (s) musculoskeletal system	100000
R74.02	Acute pharyngitis	Acute pharyngitis	800000
R78.00	Acute bronchitis/bronchiolitis	Acute bronchitis / bronchiolitis	300000
W78.00	Zwangerschap (bevestigd)	Pregnancy (confirmed)	100000
T83.0	overgewicht	overweight	100000
R65.00	episode op initiatief derde	episode on the initiative third	1

Importing codes...

Column mapping

Source code column: ICPC_CODE

Source name column: ICPC_DESCRIPTION_ENGLIGH

Source frequency column: FREQUENCY

Auto concept ID column:

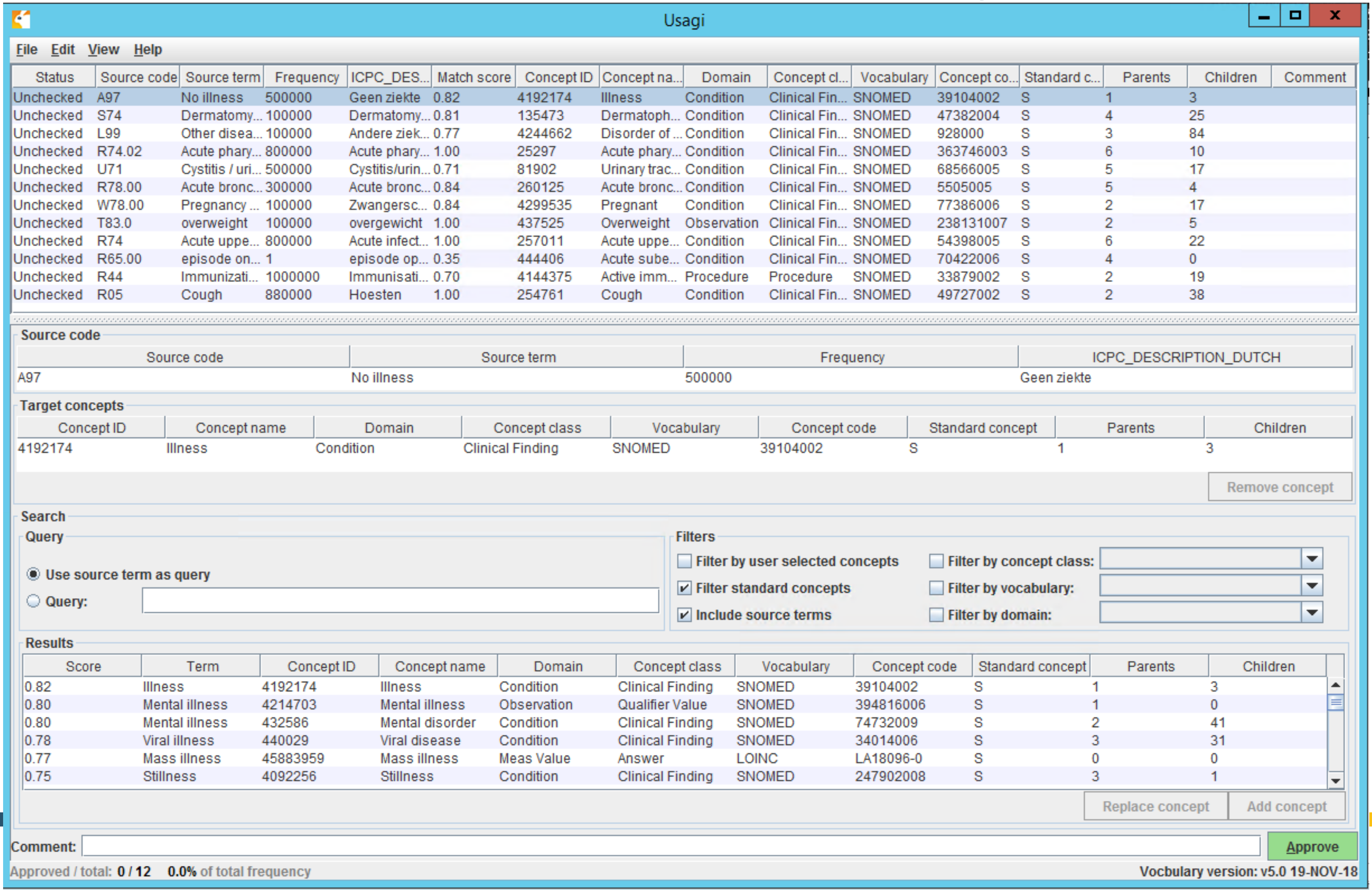
Additional info column: ICPC_DESCRIPTION_DUTCH

Filters

- ☐ Filter by user selected concepts
- ☑ Filter standard concepts
- ☑ Include source terms
- ☐ Filter by concept class:
- ☐ Filter by vocabulary:
- ☐ Filter by domain:

Usagi Process

5. ***Review and update suggest mappings*** *with someone who has medical knowledge*



Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concept co...	Standard c...	Parents	Children	Comment
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104002	S	1	3	
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382004	S	4	25	
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	928000	S	3	84	
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	363746003	S	6	10	
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566005	S	5	17	
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	5505005	S	5	4	
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	2	38	

Source code

Source code	Source term	Frequency	ICPC_DESCRIPTION_DUTCH
A97	No illness	500000	Geen ziekte

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by user selected concepts
☑ Filter standard concepts
☑ Include source terms
□ Filter by concept class:
□ Filter by vocabulary:
□ Filter by domain:

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

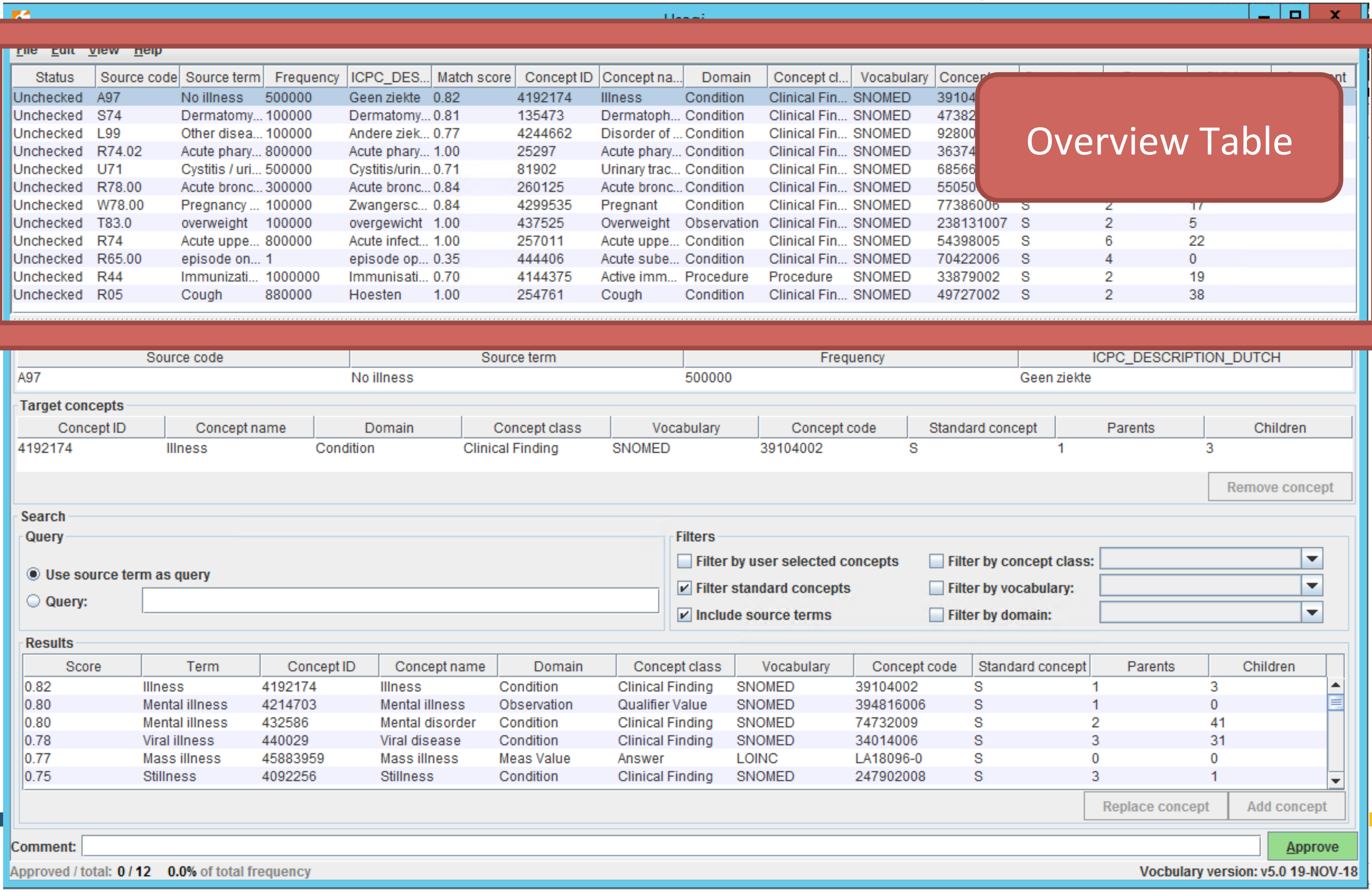
Comment:

Approved / total: 0 / 12 0.0% of total frequency

Vocbulary version: v5.0 19-NOV-18

Usagi Process

5. ***Review and update suggest mappings*** *with someone who has medical knowledge*



Overview Table

Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concep				nt
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104				
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382				
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	92800				
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	36374				
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566				
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	55050				
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	2	38	

Source code	Source term	Frequency	ICPC_DESCRIPTION_DUTCH
A97	No illness	500000	Geen ziekte

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by user selected concepts

☑ Filter standard concepts

☑ Include source terms

□ Filter by concept class:

□ Filter by vocabulary:

□ Filter by domain:

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

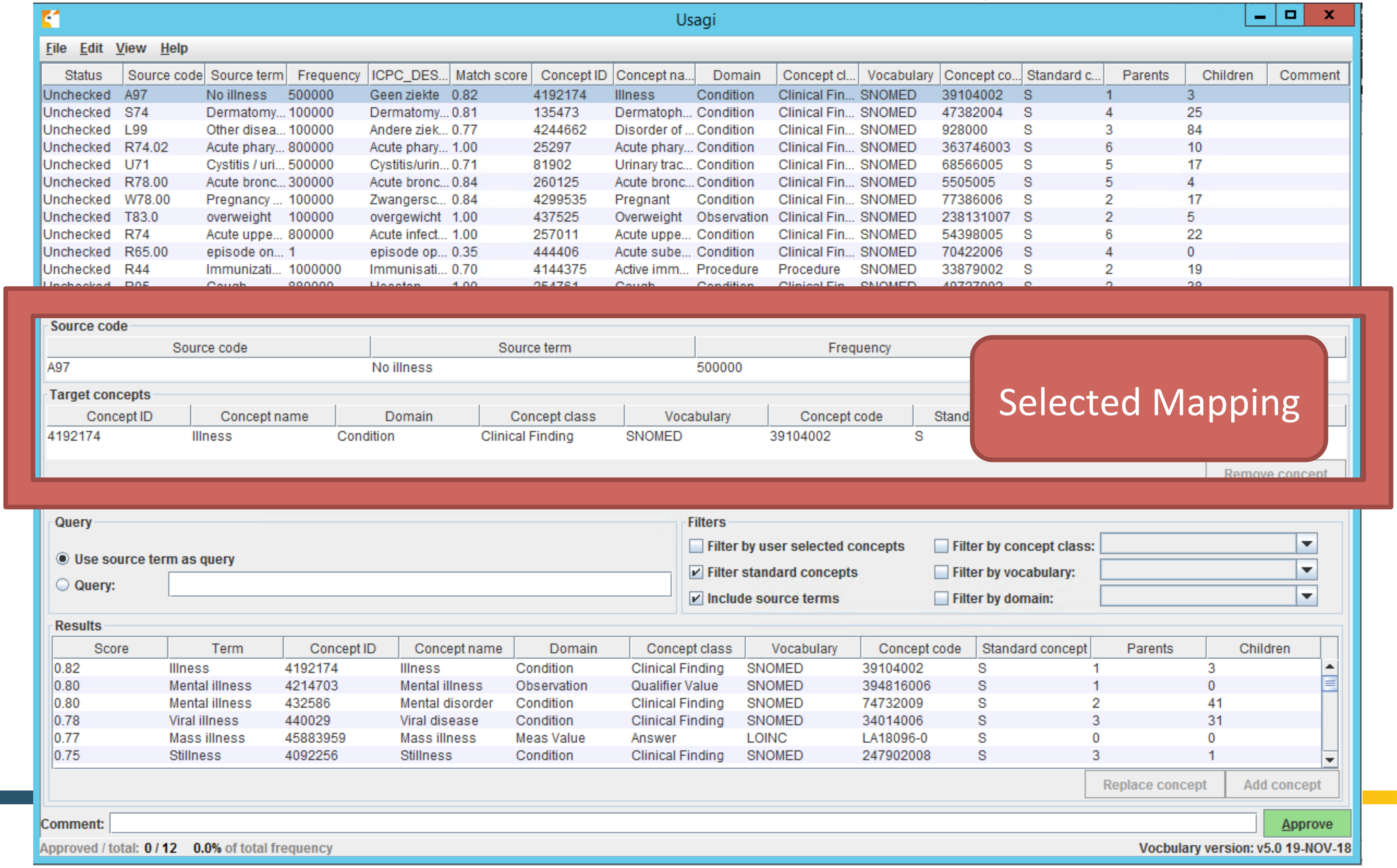
Comment:

Approved / total: 0 / 12 0.0% of total frequency

Vocbulary version: v5.0 19-NOV-18

Usagi Process

5. ***Review and update suggest mappings*** *with someone who has medical knowledge*



Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concept co...	Standard c...	Parents	Children	Comment
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104002	S	1	3	
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382004	S	4	25	
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	928000	S	3	84	
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	363746003	S	6	10	
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566005	S	5	17	
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	5505005	S	5	4	
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	3	38	

Source code

Source code	Source term	Frequency
A97	No illness	500000

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Stand
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S

Selected Mapping

Query

● Use source term as query

○ Query:

Filters

□ Filter by user selected concepts

☑ Filter standard concepts

☑ Include source terms

□ Filter by concept class:

□ Filter by vocabulary:

□ Filter by domain:

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

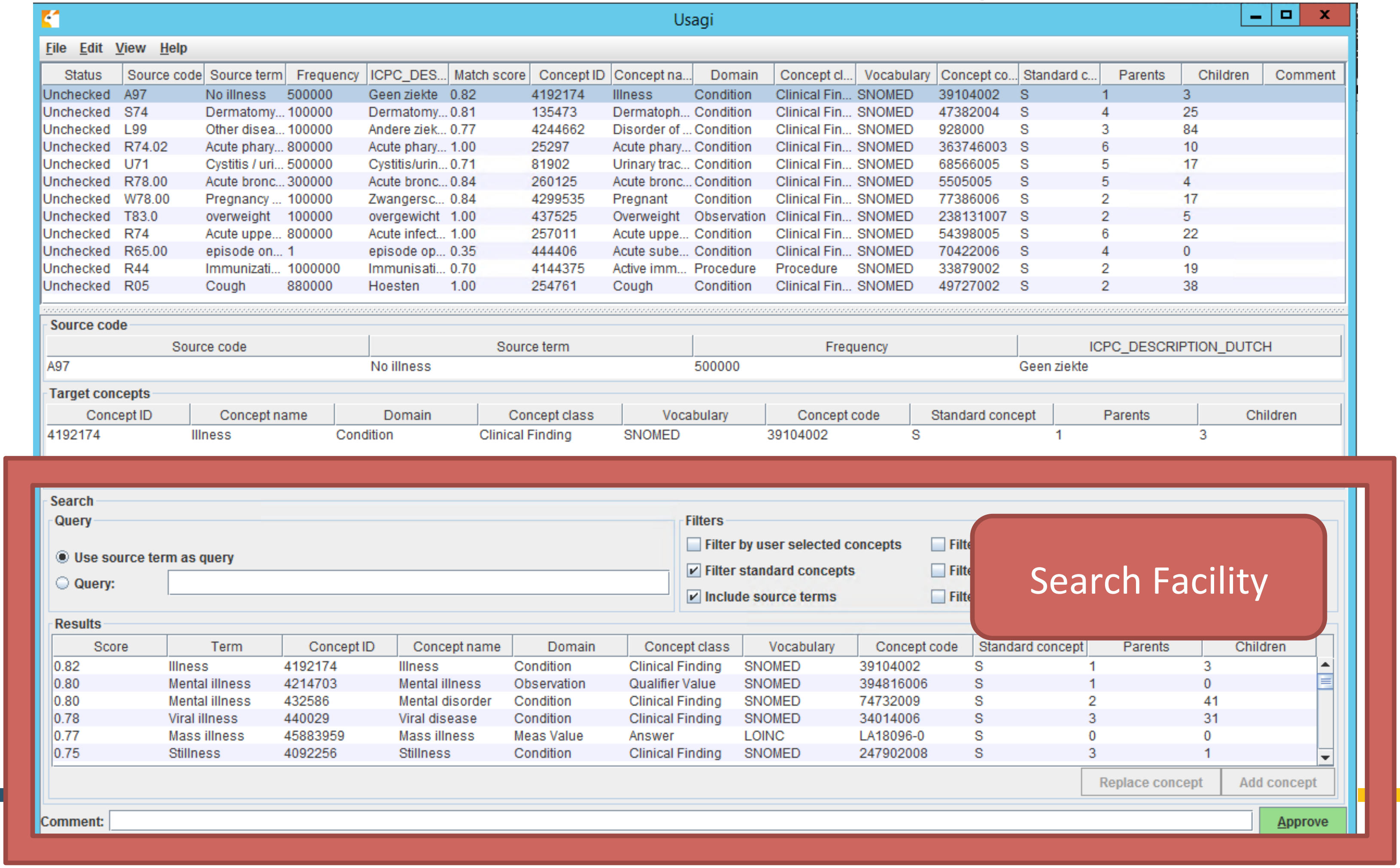
Comment:

Approved / total: 0 / 12 0.0% of total frequency

Vocbulary version: v5.0 19-NOV-18

Usagi Process

5. ***Review and update suggest mappings*** *with someone who has medical knowledge*



Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concept co...	Standard c...	Parents	Children	Comment
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104002	S	1	3	
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382004	S	4	25	
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	928000	S	3	84	
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	363746003	S	6	10	
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566005	S	5	17	
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	5505005	S	5	4	
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	2	38	

Source code

Source code	Source term	Frequency	ICPC_DESCRIPTION_DUTCH
A97	No illness	500000	Geen ziekte

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3

Search

Query

● Use source term as query

○ Query:

Filters

☐ Filter by user selected concepts

☑ Filter standard concepts

☑ Include source terms

☐ Filte

☐ Filte

☐ Filte

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

Search Facility

Usagi Process

*5. **Review and update suggest mappings** with someone who has medical knowledge*

- It may be valuable to sort on "Match Score"; reviewing codes that Usagi is most confident on first may quickly knock out a significant chunk of codes

- Sorting on "Frequency" is valuable, spending more effort on frequent codes versus non-frequent is important

- It is okay to map to zero or 0 – "No matching concept"

- A source code might end up being mapped to two concepts

- You might have what the system considers one domain but the OMOP Vocabulary lumps into another domain

Usagi Process

*6. **Export codes** into the SOURCE_TO_CONCEPT_MAP*

- After you have completed, you will export the relationships

- When exporting you will give a Vocabulary ID, for example JNJ_JMDC_PROVIDERS would signify a Johnson & Johnson map for the database JMDC for provider codes.

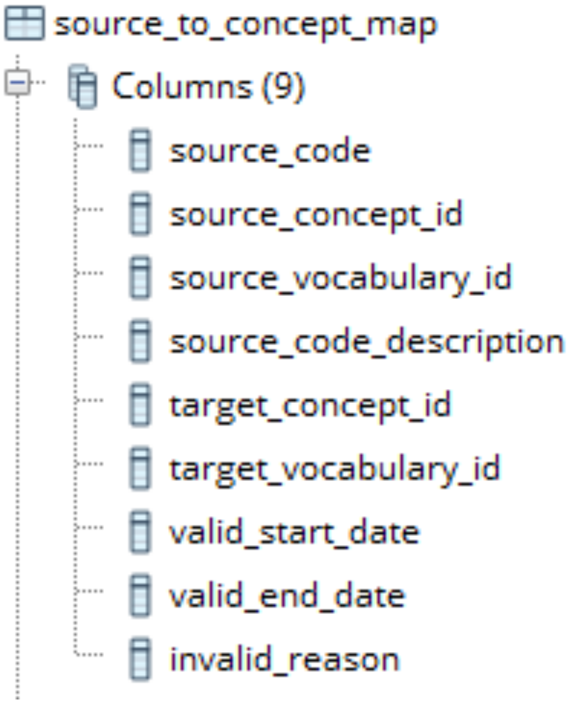
source_code	source_concept_id	source_vocab_id	source_code_description	target_concept_id	target_vocab_id	valid_start_date	valid_end_date	invalid_reason
R74.02	0	TEST_VOCAB	Acute pharyngitis	25297	SNOMED	1/1/1970	12/31/2099	

R74.02 - Acute pharyngitis = 25297 - Acute pharyngitis

Usagi Process

*6. **Export codes** into the SOURCE_TO_CONCEPT_MAP*

- You then load your generated maps into the empty Vocabulary table.



source_to_concept_map

- Columns (9)
  - 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

Usagi – Your Turn

1. ✓ Get a copy of the **Vocabulary** from ATHENA

2. ✓ Download **Usagi**

3. ✓ Have Usagi **build an index** on the Vocabulary

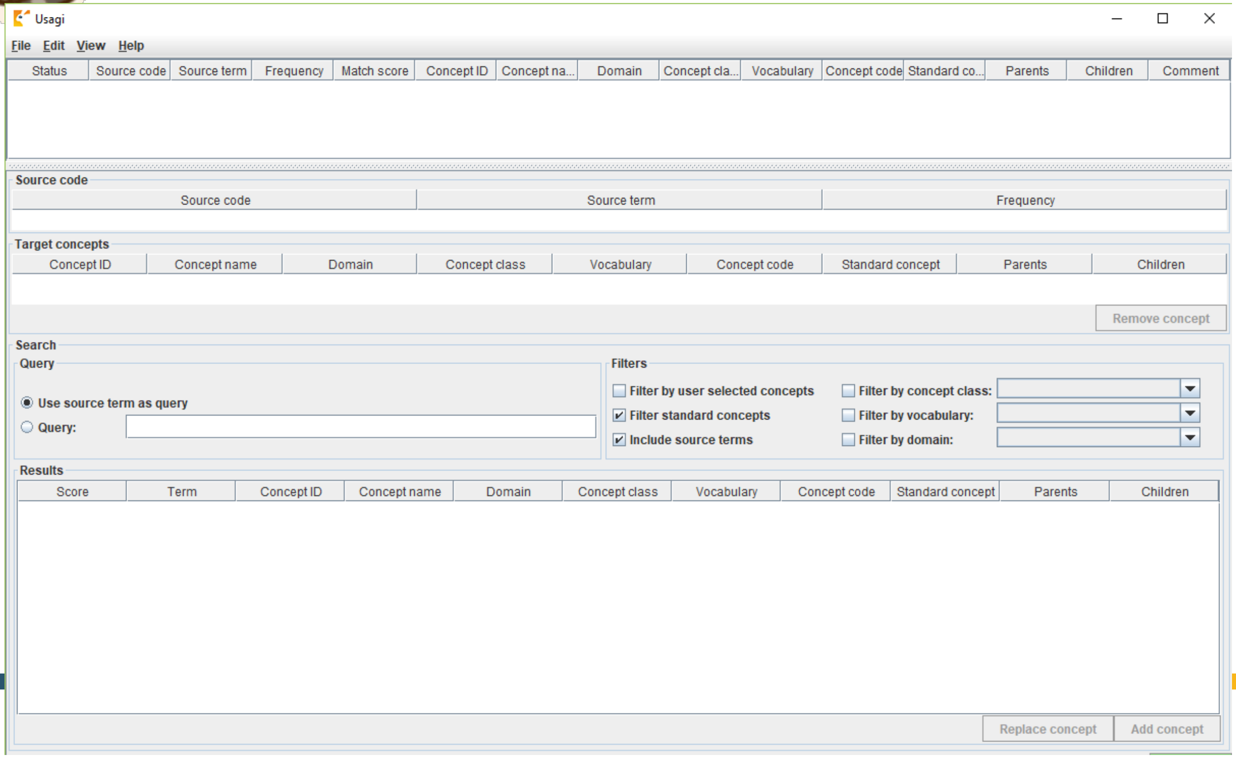
4. **Load your source codes** and let Usagi process them

5. **Review and update suggest mappings** with someone who has medical knowledge

6. **Export codes** into the SOURCE_TO_CONCEPT_MAP

Now Your Turn:
Open Usagi

- Click on Usagi shortcut
- Go into the Usagi-1.1.6 folder
- Open Usagi.jar



Usagi

File Edit View Help

Status	Source code	Source term	Frequency	Match score	Concept ID	Concept na...	Domain	Concept cla...	Vocabulary	Concept code	Standard co...	Parents	Children	Comment

Source code

Source code	Source term	Frequency

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by user selected concepts

☑ Filter standard concepts

☑ Include source terms

□ Filter by concept class:

□ Filter by vocabulary:

□ Filter by domain:

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children

Usagi – Your Turn

- We have provided a small subset of codes to try to map

https://github.com/OHDSI/Tutorial-ETL/

-> Materials -> Usagi -> DUTCH_ICPC_CONDITION_CODES_TO_MAP.xlsx

- These condition codes are in Dutch ICPC codes and need to be mapped to standard concepts

Usagi – Your Turn

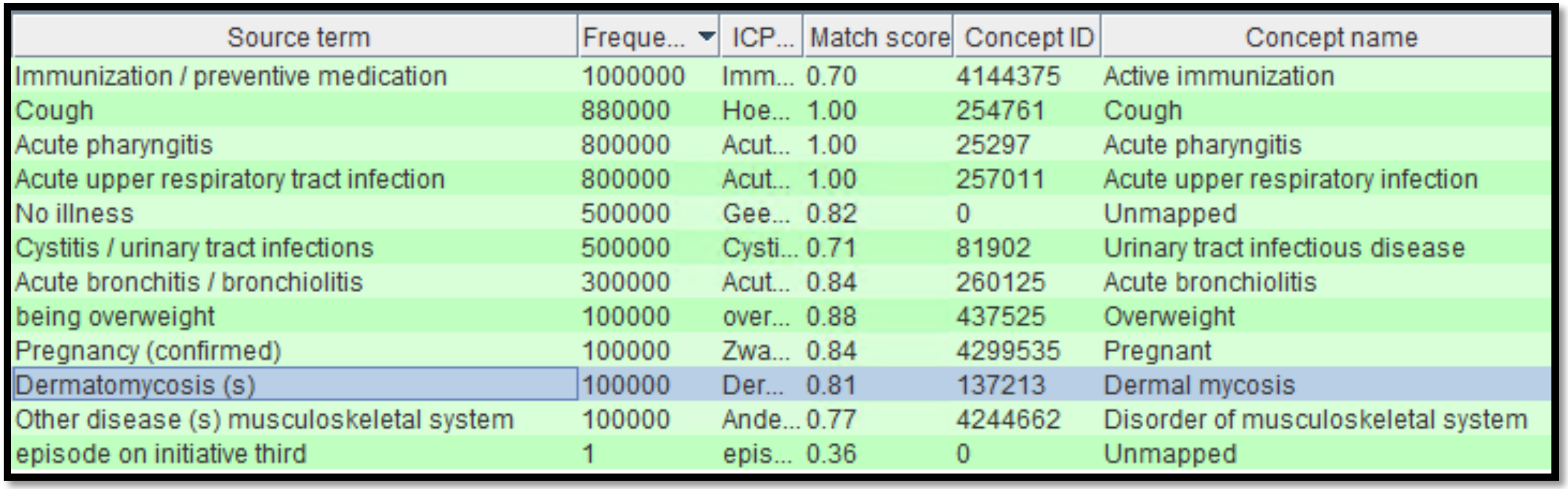
- Your mission:
  - Download the codes to map
  - Translate codes to English
  - Import codes into Usagi
  - Map to standard concepts
  - Export SOURCE_TO_CONCEPT_MAP table

- For help review the User Guide:
  - https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html#usagi

Usagi – Your Turn

- What CONCEPT_ID do you map "Dermatomycosis (s)" to?

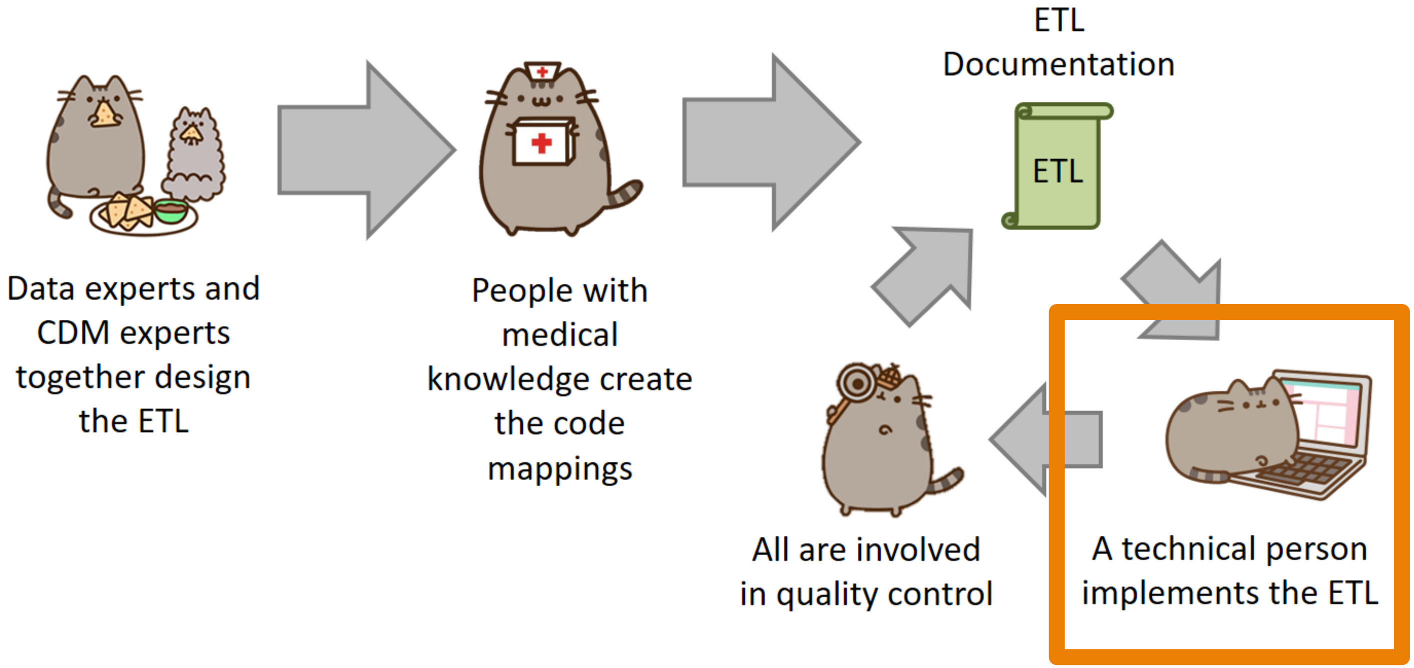
Usagi – Your Turn



Source term	Freque...	ICP...	Match score	Concept ID	Concept name
Immunization / preventive medication	1000000	Imm...	0.70	4144375	Active immunization
Cough	880000	Hoe...	1.00	254761	Cough
Acute pharyngitis	800000	Acut...	1.00	25297	Acute pharyngitis
Acute upper respiratory tract infection	800000	Acut...	1.00	257011	Acute upper respiratory infection
No illness	500000	Gee...	0.82	0	Unmapped
Cystitis / urinary tract infections	500000	Cysti...	0.71	81902	Urinary tract infectious disease
Acute bronchitis / bronchiolitis	300000	Acut...	0.84	260125	Acute bronchiolitis
being overweight	100000	over...	0.88	437525	Overweight
Pregnancy (confirmed)	100000	Zwa...	0.84	4299535	Pregnant
Dermatomycosis (s)	100000	Der...	0.81	137213	Dermal mycosis
Other disease (s) musculoskeletal system	100000	Ande...	0.77	4244662	Disorder of musculoskeletal system
episode on initiative third	1	epis...	0.36	0	Unmapped

OHDSI

Observational Health Data Sciences and Informatics



Data experts and CDM experts together design the ETL

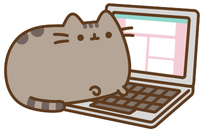
People with medical knowledge create the code mappings

ETL Documentation

A technical person implements the ETL

All are involved in quality control

ETL Implementation

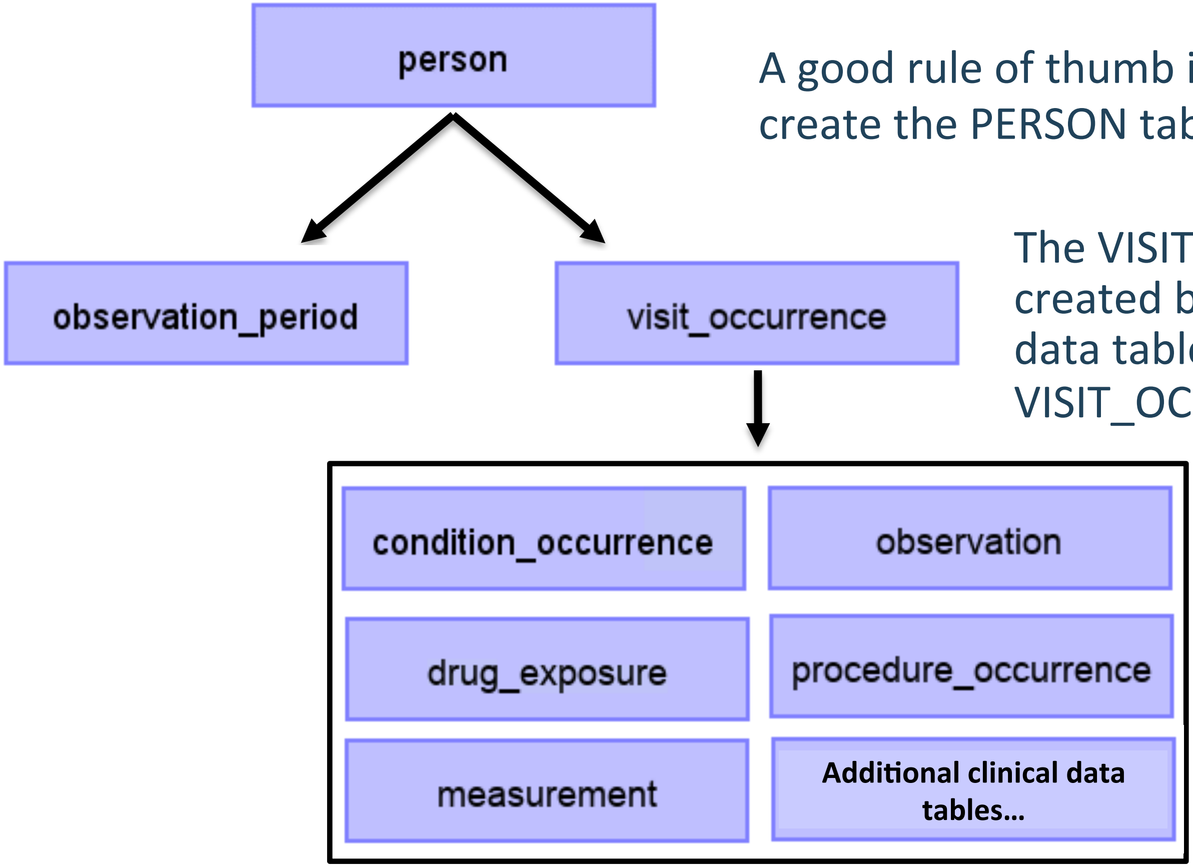


There are multiple tools available to implement your ETL

In this example we created a builder using SQL and R, though your choice will largely depend on the size and complexity of the ETL design

ETL Implementation

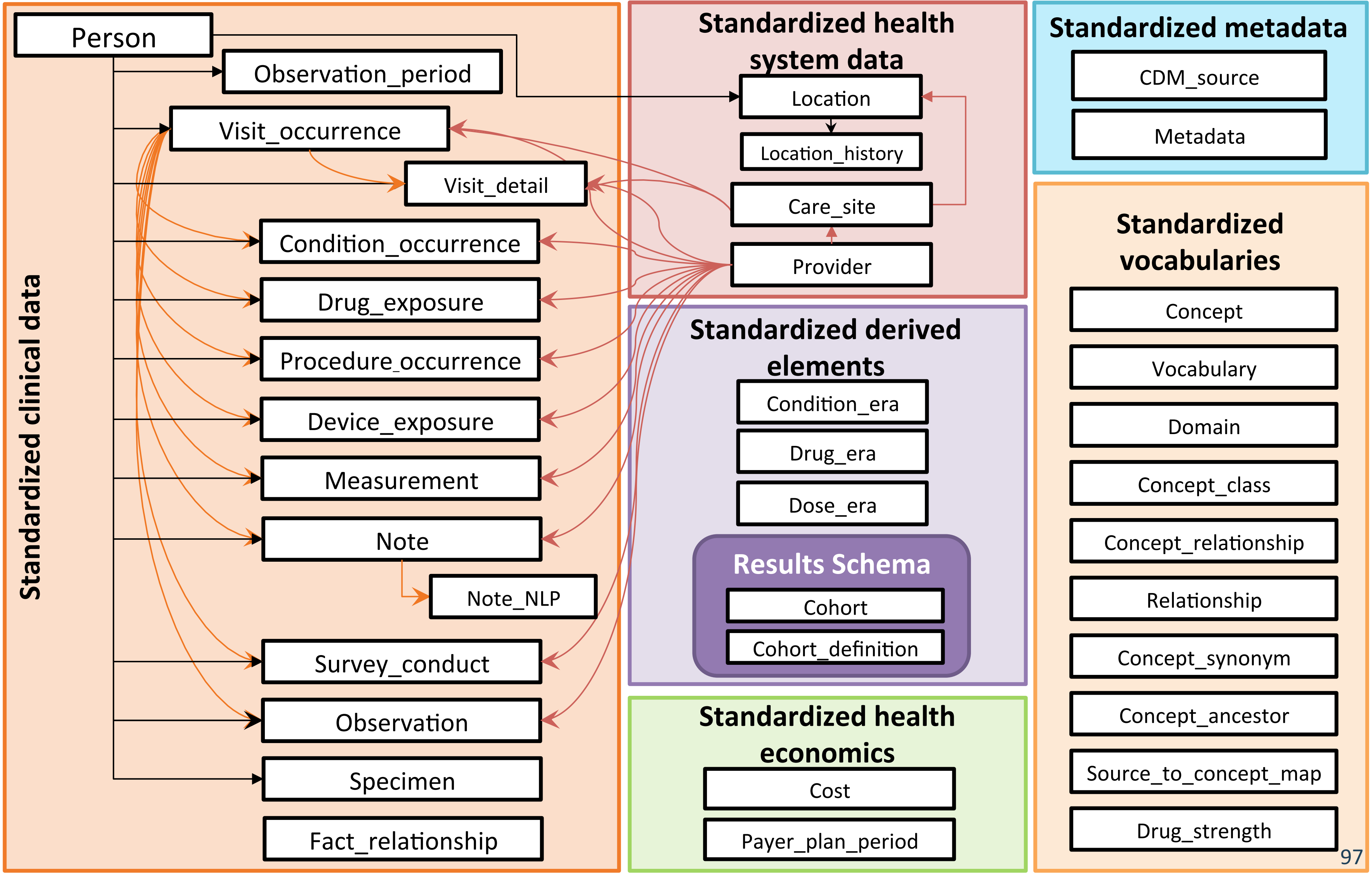
General Flow of Implementation



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

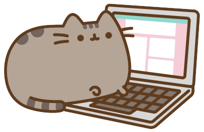
The VISIT_OCCURRENCE table must be created before the standardized clinical data tables as they all refer to the VISIT_OCCURRENCE_ID

CDM Version 6 Key Domains



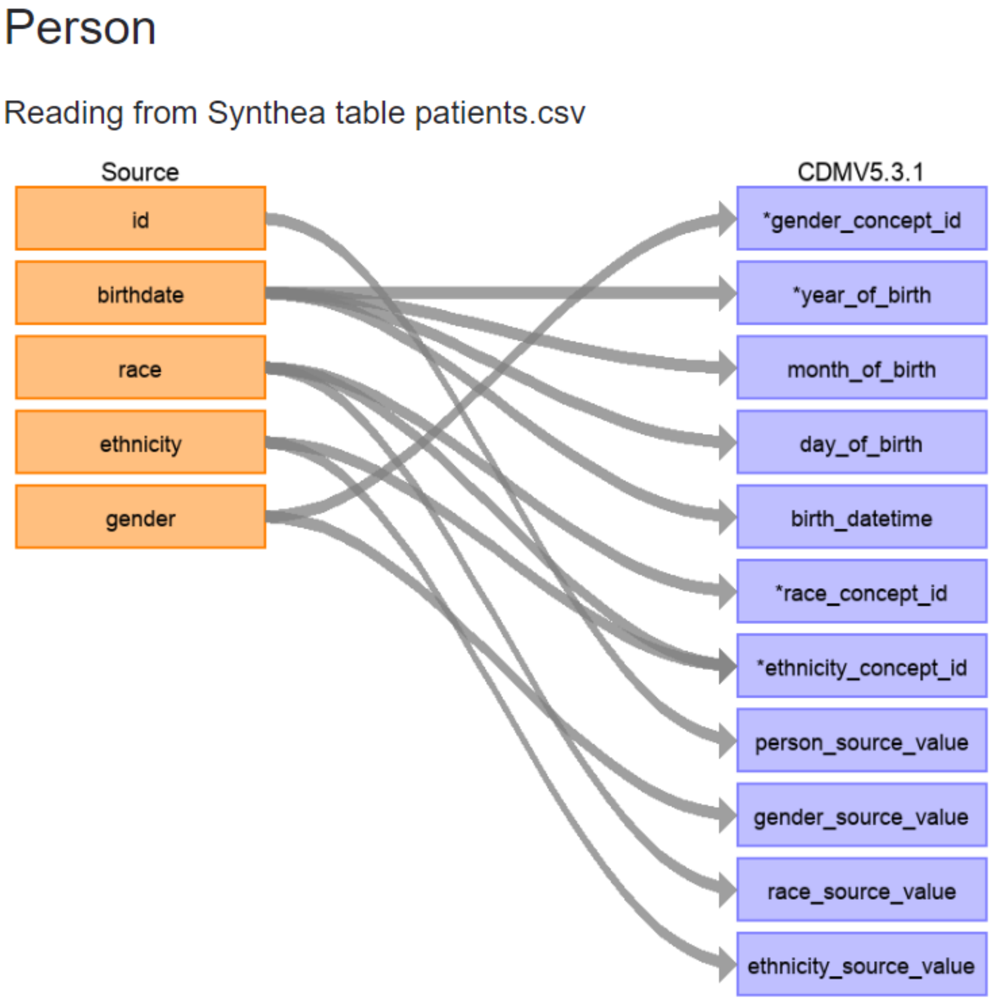
ETL Implementation

person



First, let us review the logic we decided on for how the PERSON table should be created.

Navigate in your browser to: https://ohdsi.github.io/ETL-Synthea/Person.html

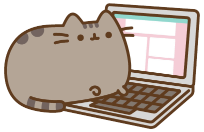


Person

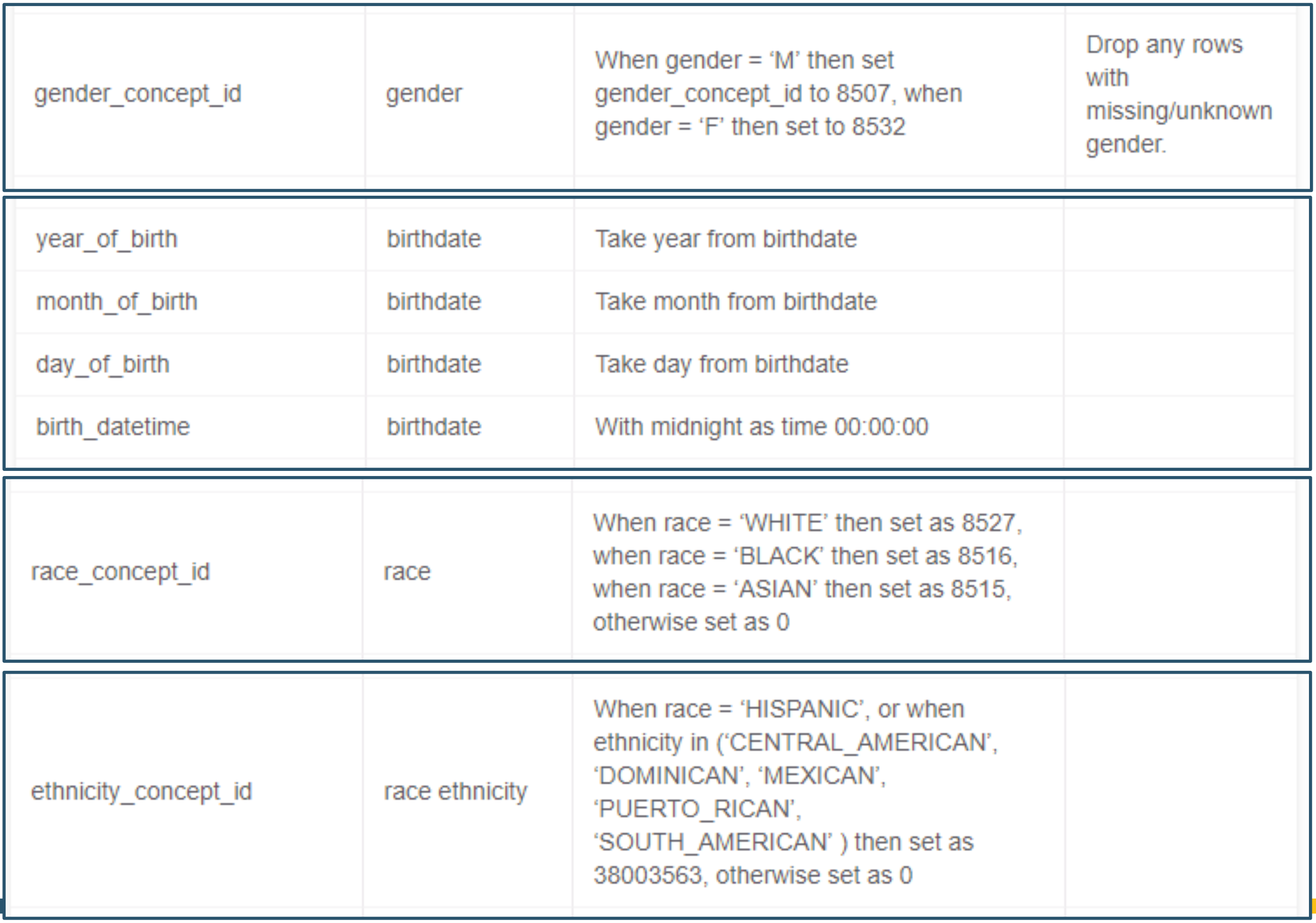
Reading from Synthea table patients.csv

ETL Implementation

person



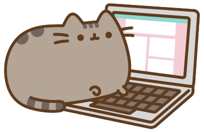
First, let's review the logic we decided on for how the PERSON table should be created.



Gender:	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.
Birthdate:	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:	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:	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	

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?

To open the query while we review:

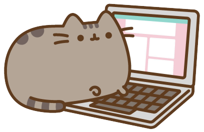
https://github.com/OHDSI/Tutorial-ETL

Materials → Implementation→ Insert_Person_Lauren.sql

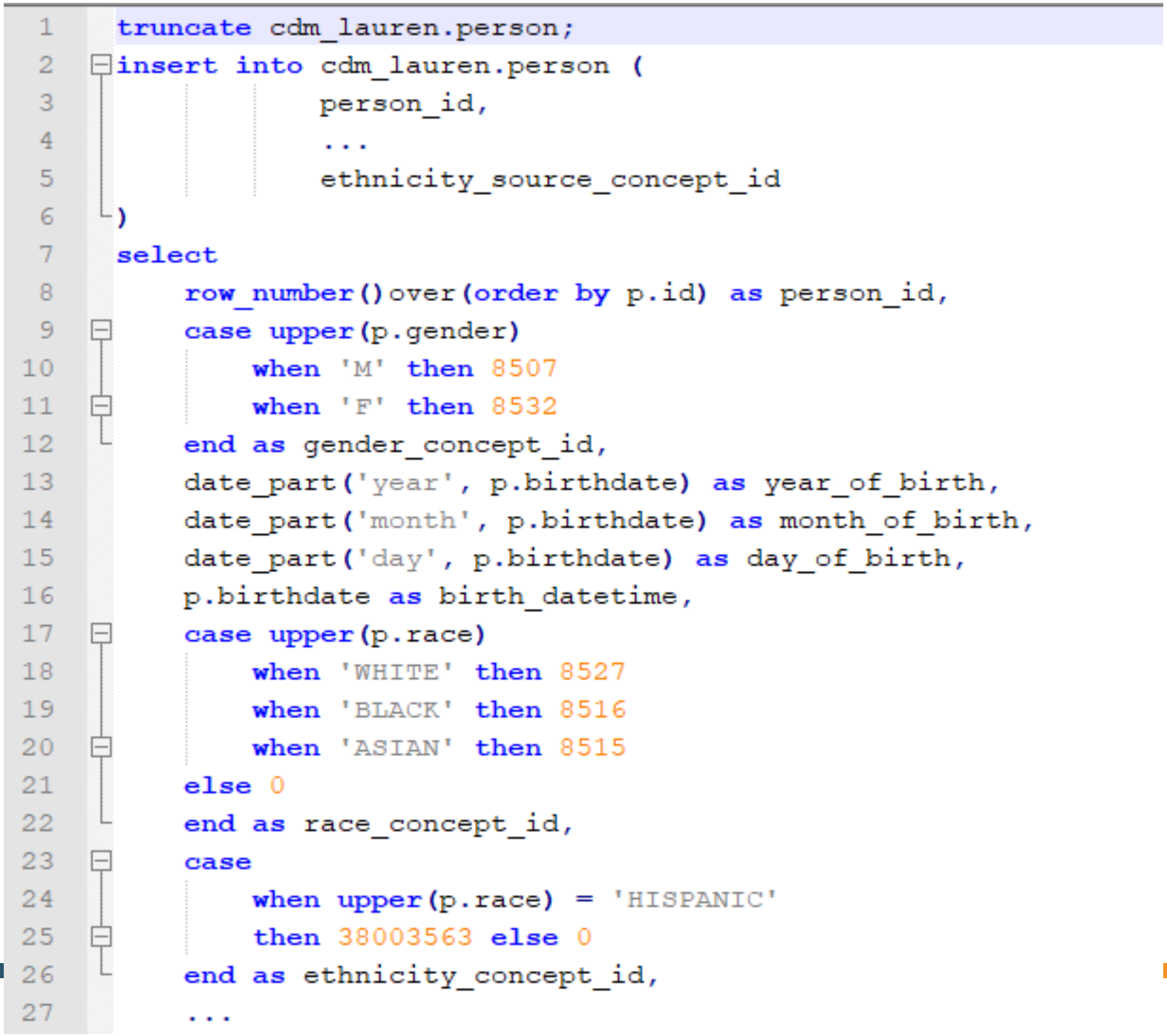
You can either view it directly in GitHub or download it and open it in pgAdmin4

ETL Implementation

person



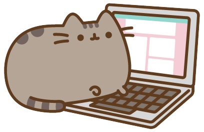
How should the PERSON table logic be implemented in SQL?



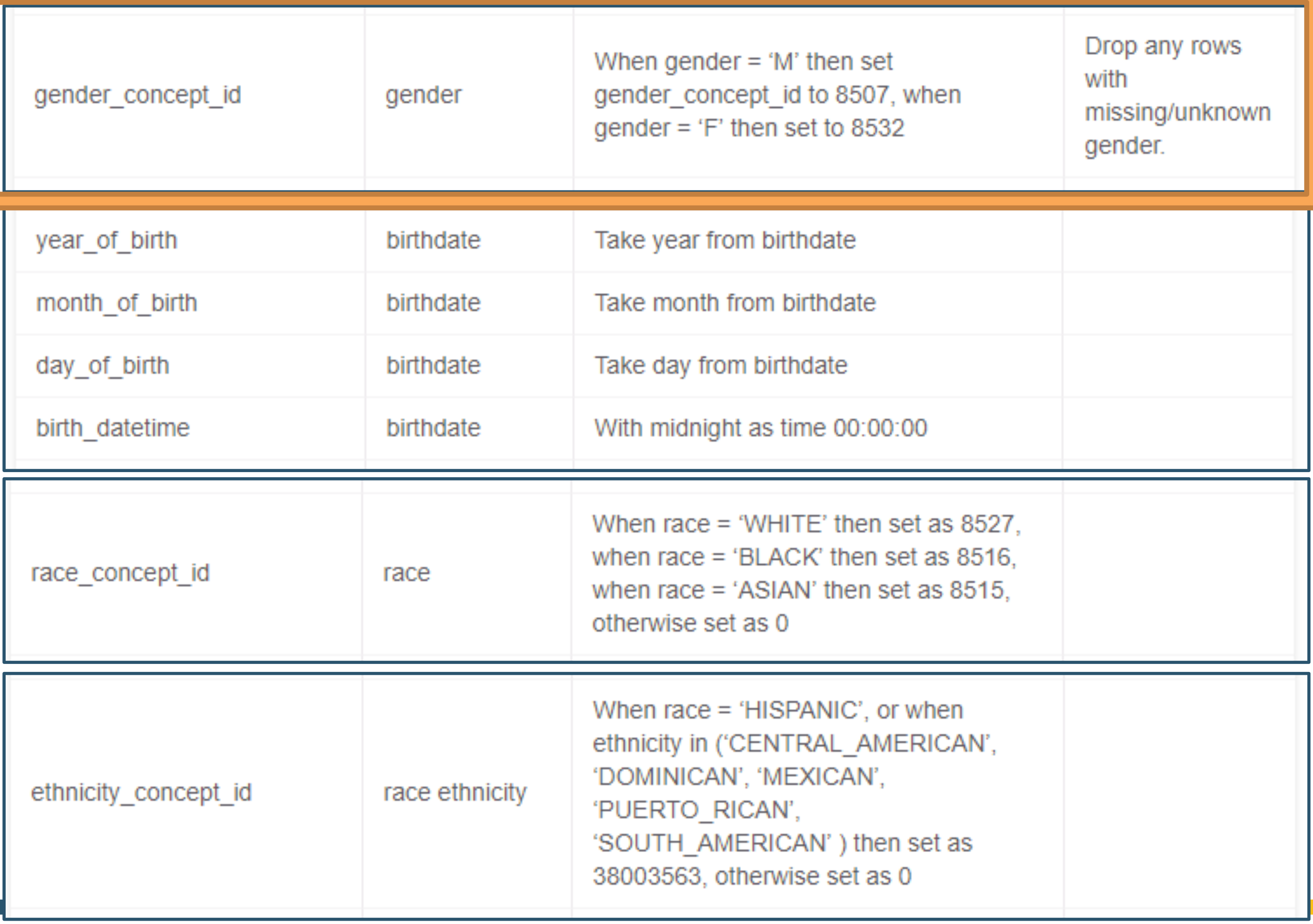
```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
    date_part('day', p.birthdate) as day_of_birth,
    p.birthdate as birth_datetime,
    case upper(p.race)
        when 'WHITE' then 8527
        when 'BLACK' then 8516
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

ETL Implementation

person



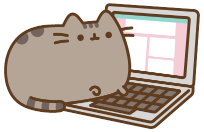
Let’s review the logic we decided on for how the PERSON table should be created.



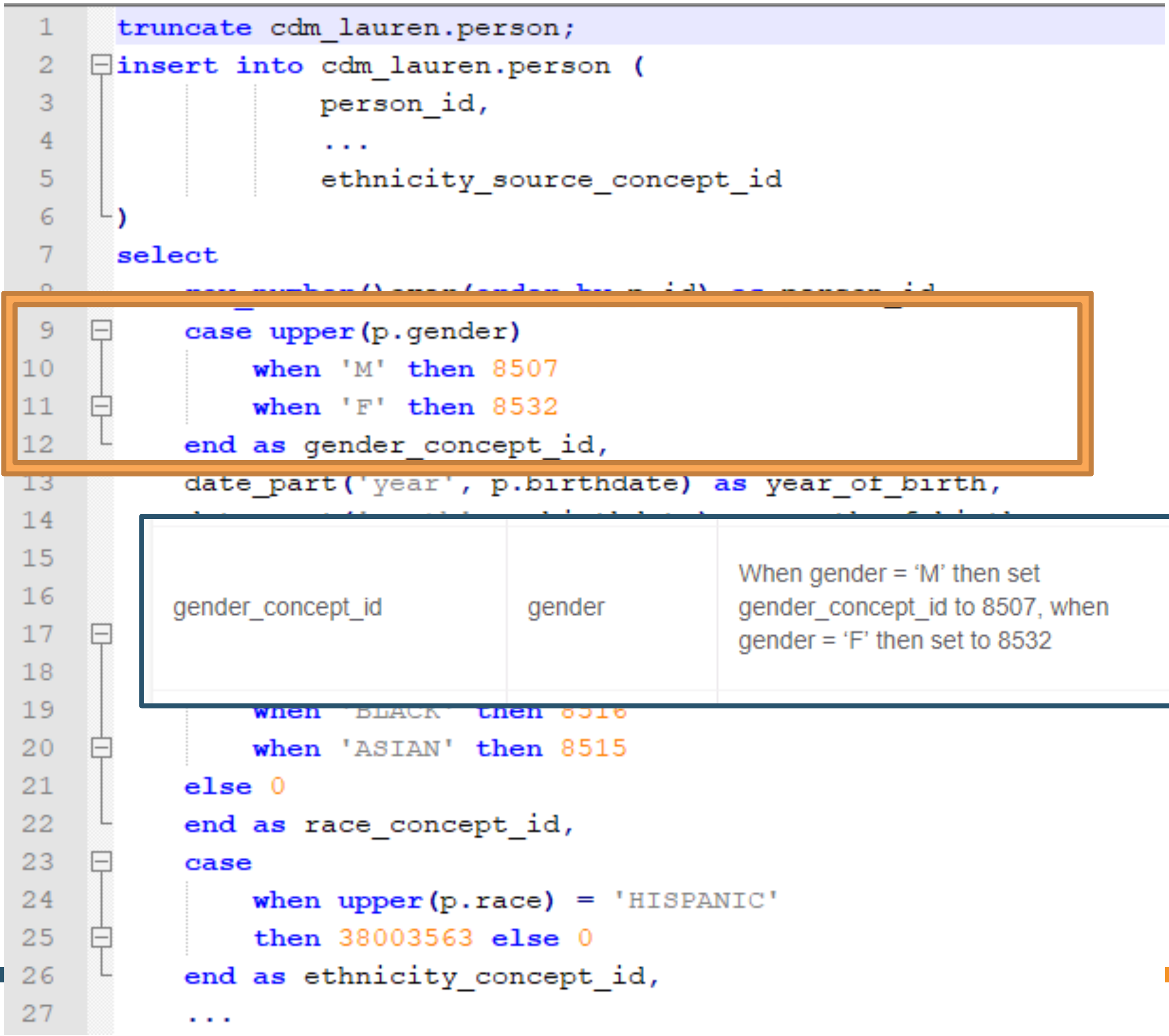
Gender:	gender_concept_id	gender	When gender = ‘M’ then set gender_concept_id to 8507, when gender = ‘F’ then set to 8532	Drop any rows with missing/unknown gender.
Birthdate:	year_of_birth	birthdate	Take year from birthdate	
	month_of_birth	birthdate	Take month from birthdate	
	day_of_birth	birthdate	Take day from birthdate	
	birth_datetime	birthdate	With midnight as time 00:00:00	
Race:	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:	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	

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?



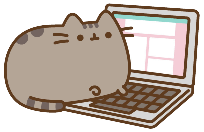
```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
        when 'BLACK' then 8516
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

Gender

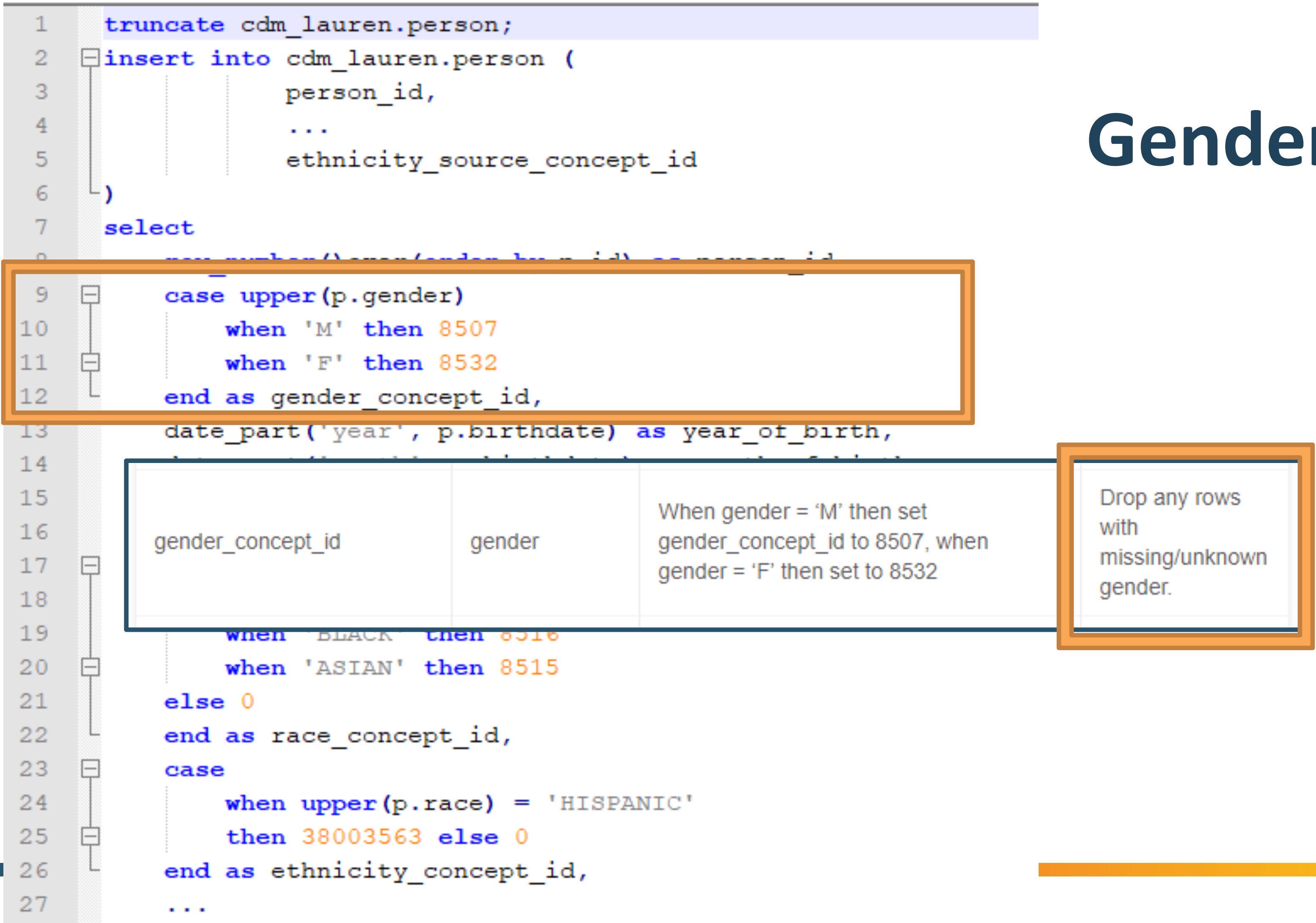
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?



```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
        when 'BLACK' then 8516
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

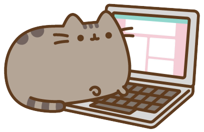
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.

Gender

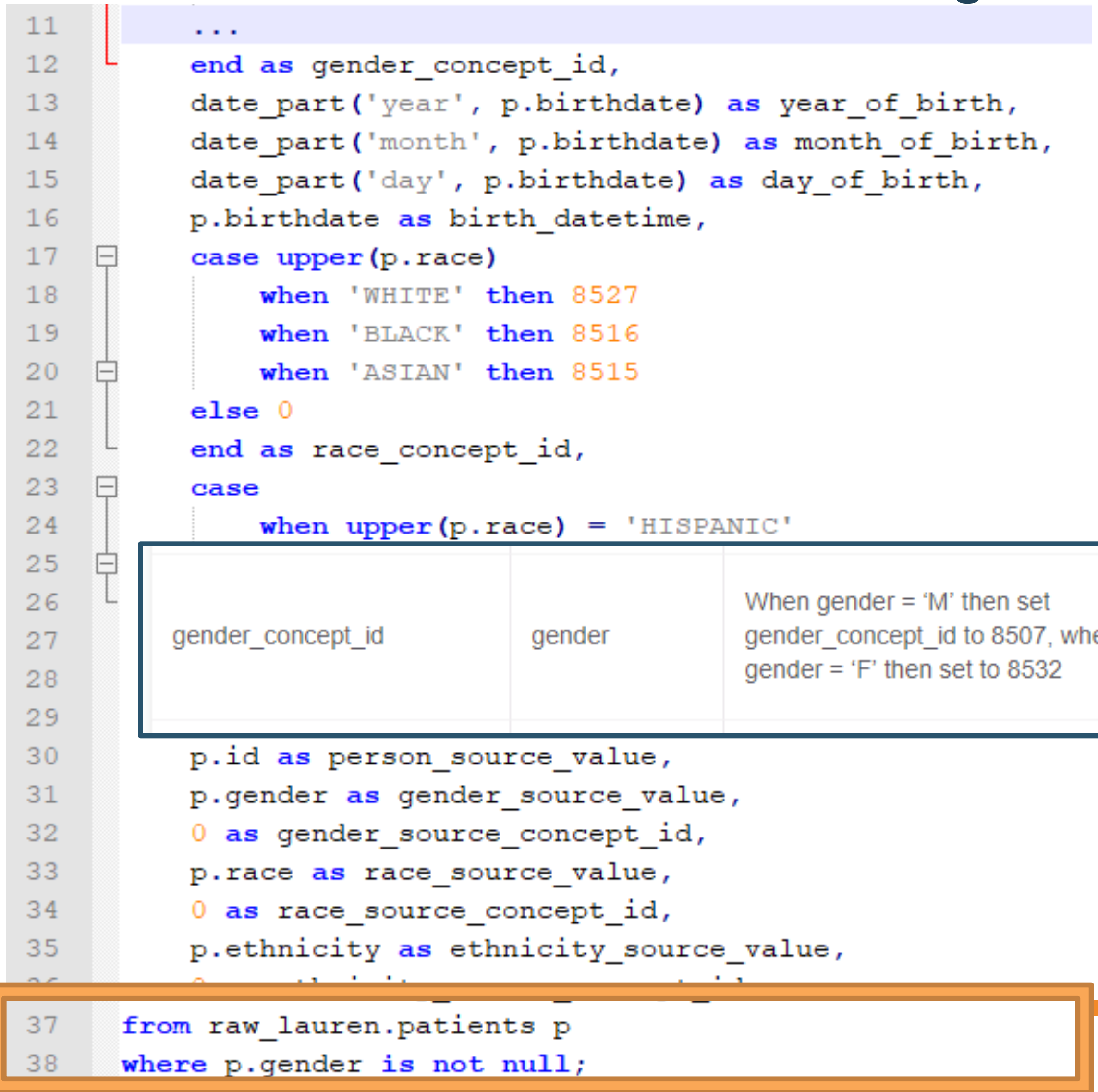
??

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?



```
    ...
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
    date_part('day', p.birthdate) as day_of_birth,
    p.birthdate as birth_datetime,
    case upper(p.race)
        when 'WHITE' then 8527
        when 'BLACK' then 8516
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
    p.id as person_source_value,
    p.gender as gender_source_value,
    0 as gender_source_concept_id,
    p.race as race_source_value,
    0 as race_source_concept_id,
    p.ethnicity as ethnicity_source_value,
    0 as ethnicity_source_concept_id,
from raw_lauren.patients p
where p.gender is not null;
```

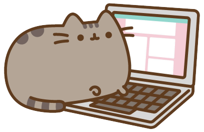
Gender

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.

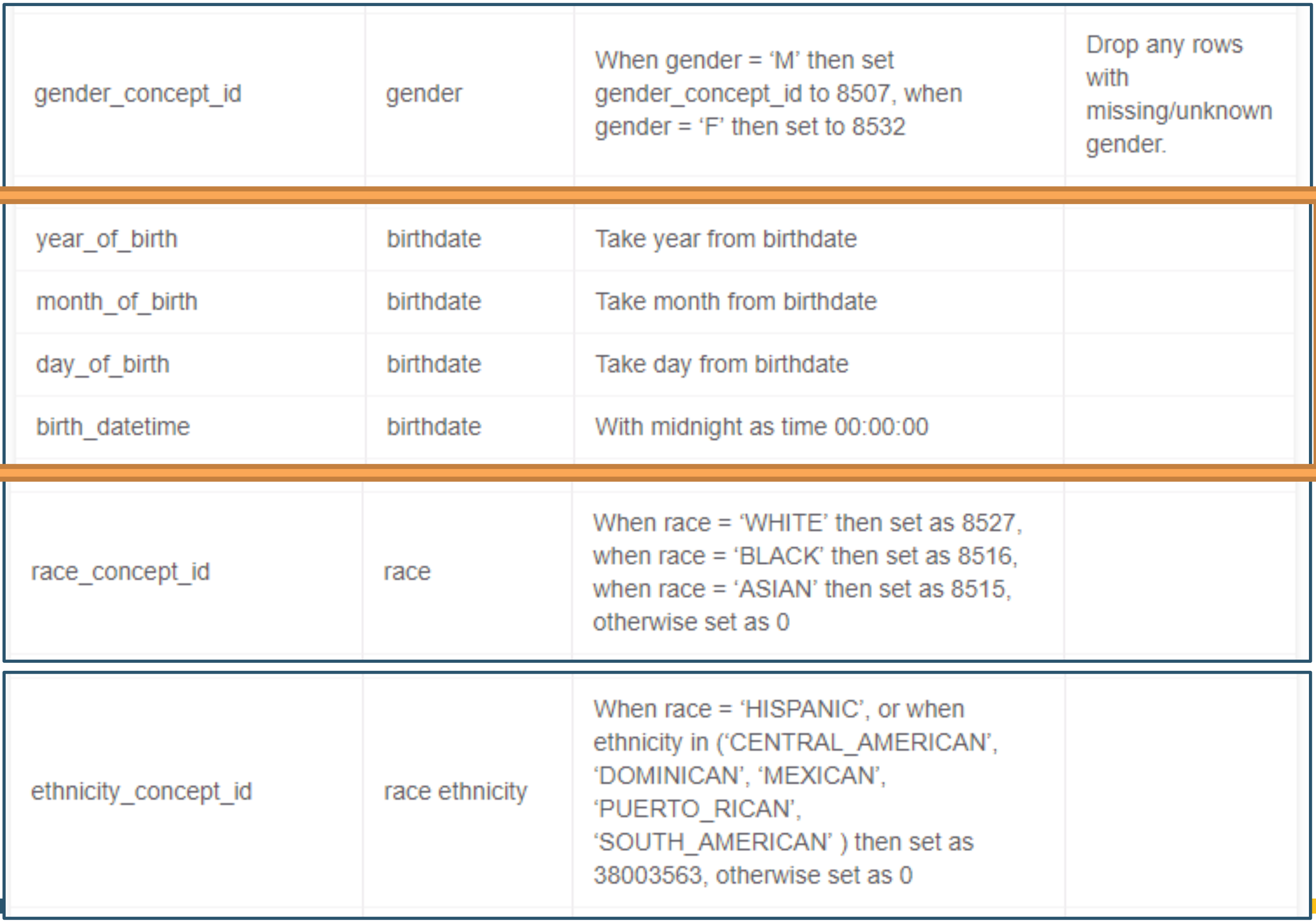
??

ETL Implementation

person



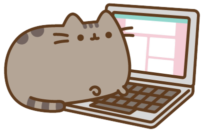
Let's review the logic we decided on for how the PERSON table should be created.



Gender:	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.
Birthdate:	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:	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:	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	

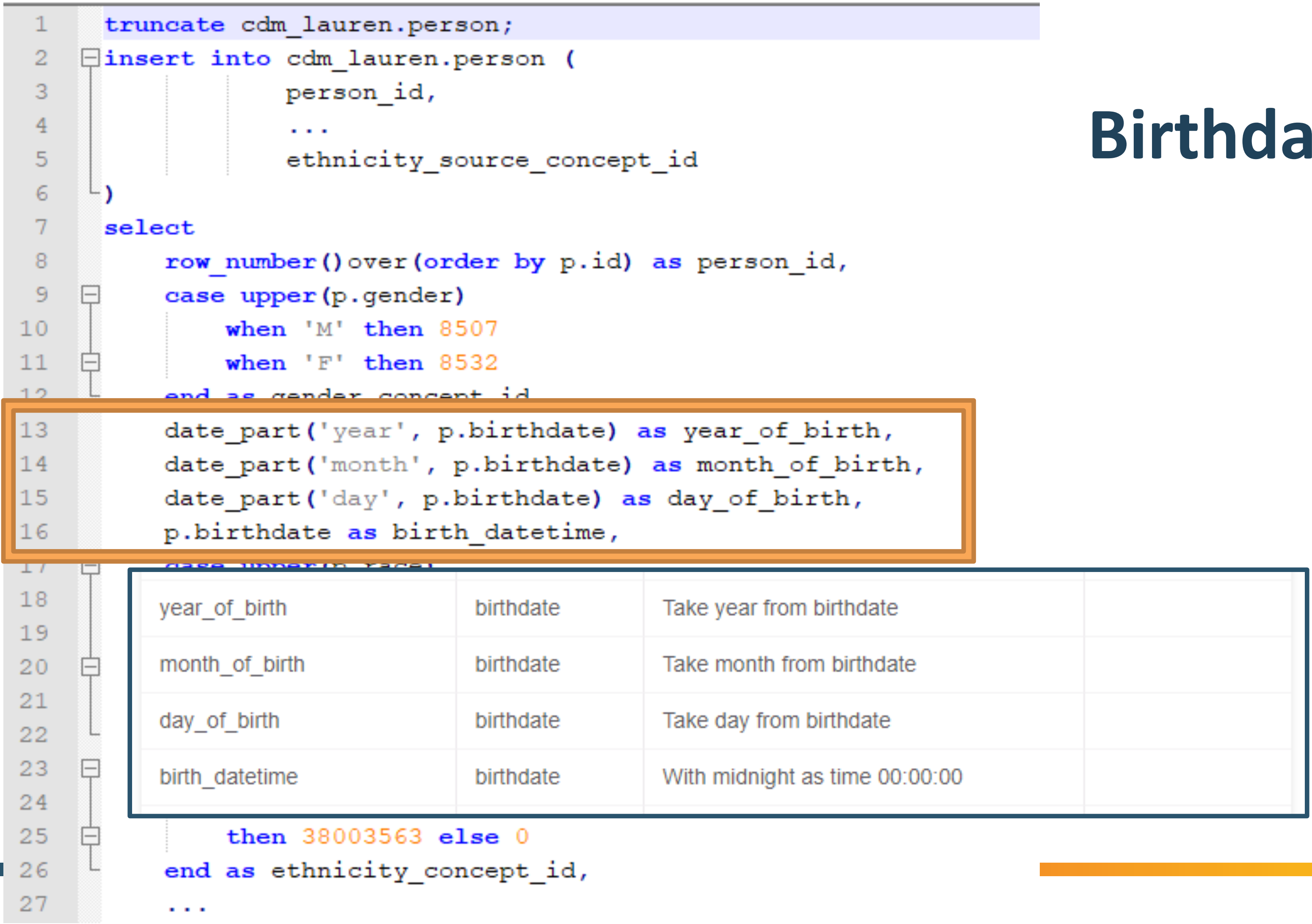
ETL Implementation

person



How should the PERSON table logic be implemented in SQL?

Birthdate



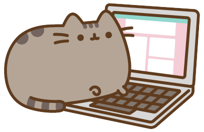
```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
    date_part('day', p.birthdate) as day_of_birth,
    p.birthdate as birth_datetime,
    case upper(p.race)
```

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	

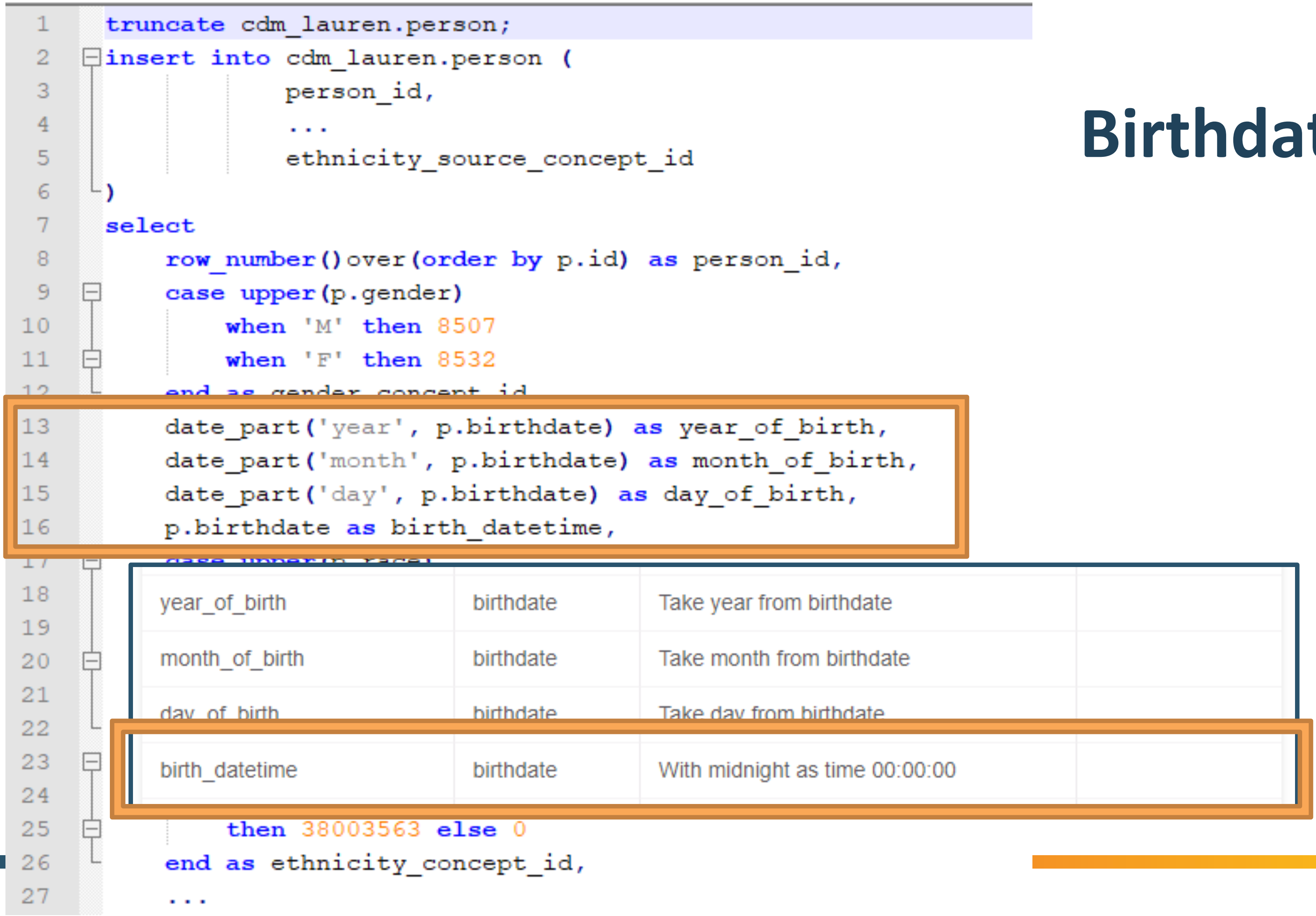
```
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?



```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
    date_part('day', p.birthdate) as day_of_birth,
    p.birthdate as birth_datetime,
    case upper(p.race)
```

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	

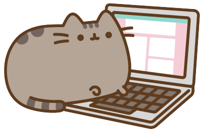
```
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

Birthdate

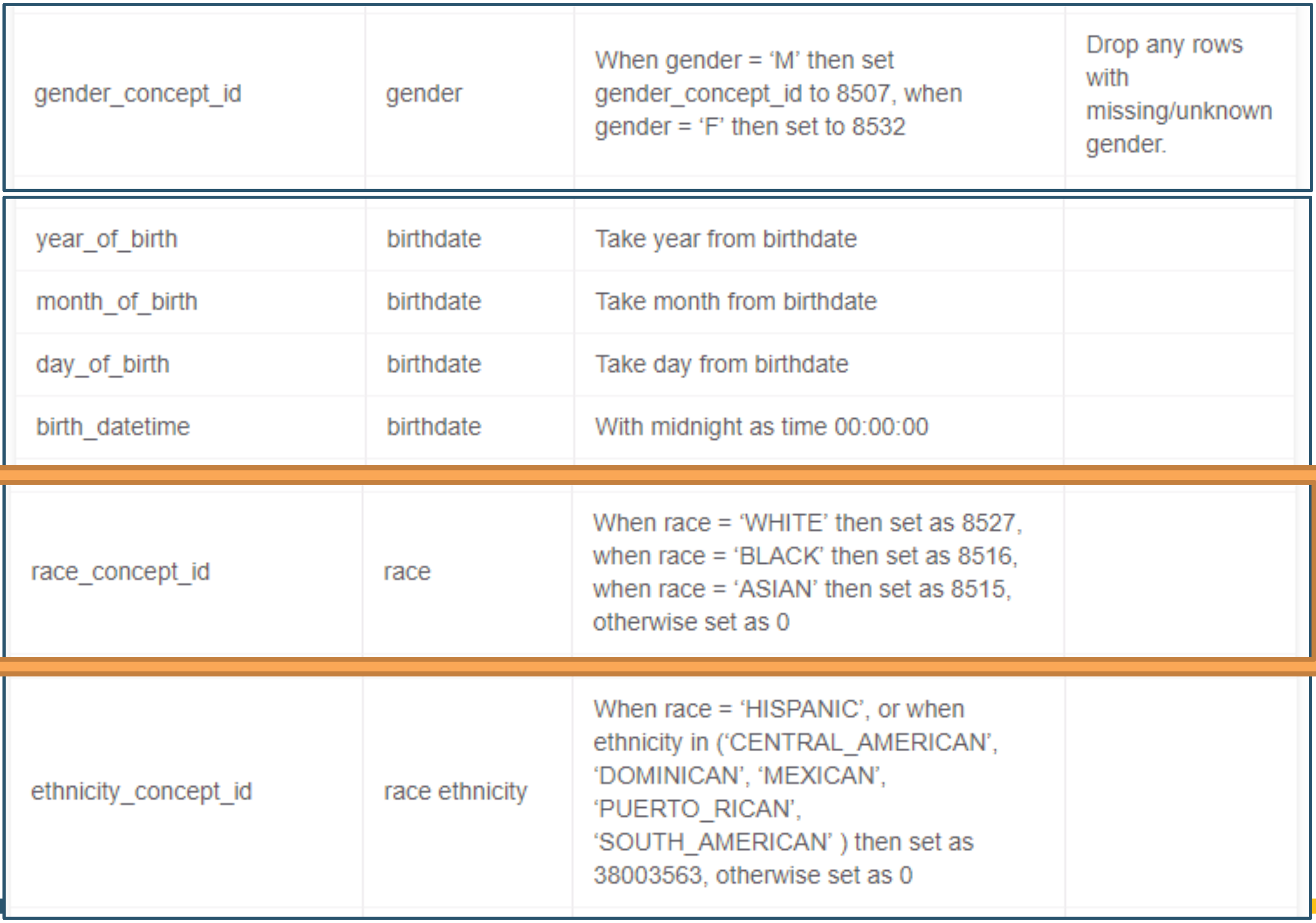
??

ETL Implementation

person



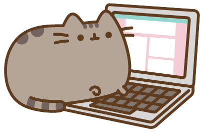
Let’s review the logic we decided on for how the PERSON table should be created.



Gender:	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.
Birthdate:	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:	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:	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	

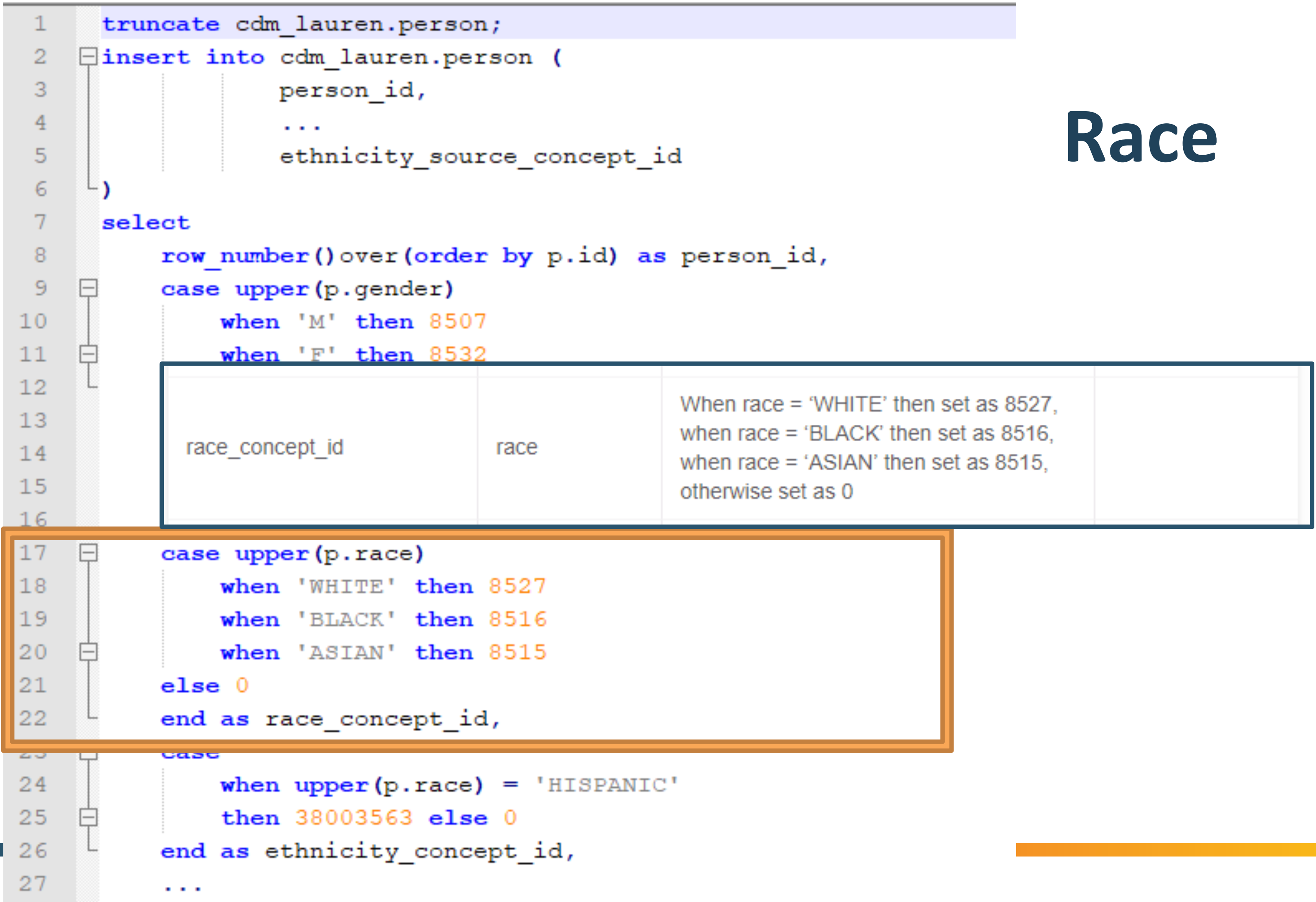
ETL Implementation

person



How should the PERSON table logic be implemented in SQL?

Race



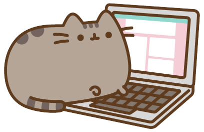
```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
```

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	

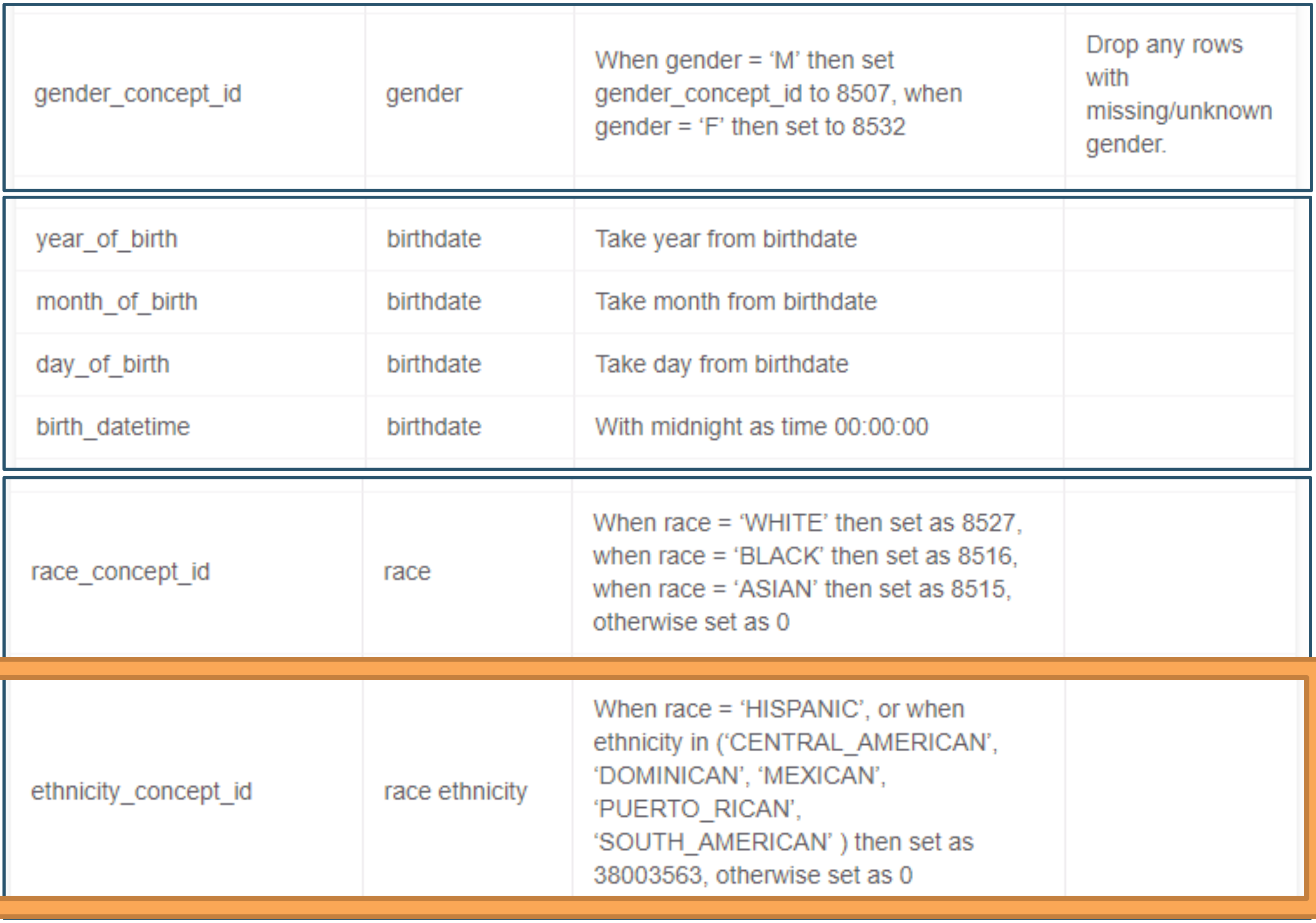
```
    case upper(p.race)
        when 'WHITE' then 8527
        when 'BLACK' then 8516
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

ETL Implementation

person



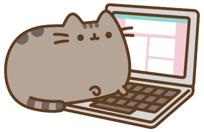
Let’s review the logic we decided on for how the PERSON table should be created.



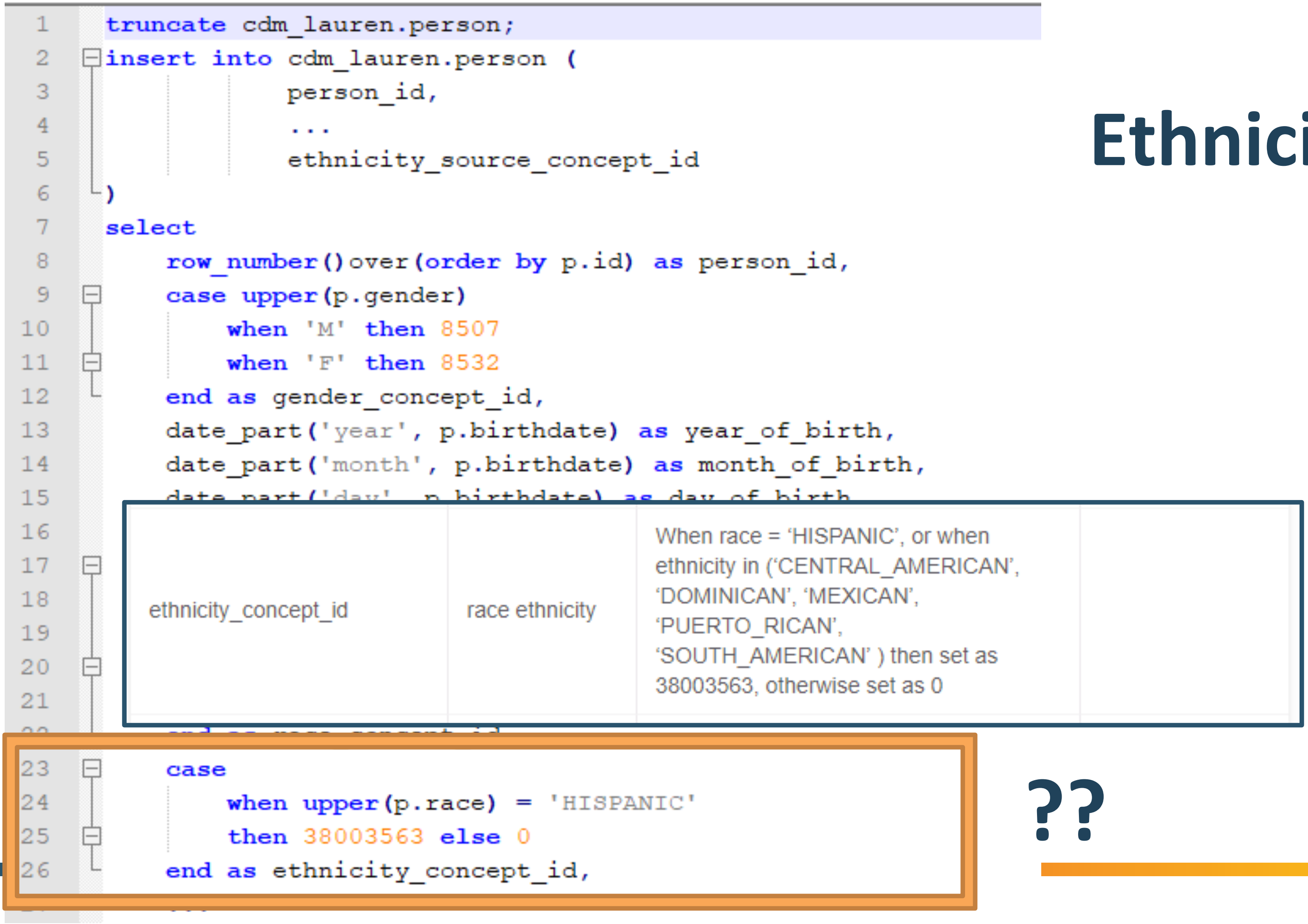
Gender:	gender_concept_id	gender	When gender = ‘M’ then set gender_concept_id to 8507, when gender = ‘F’ then set to 8532	Drop any rows with missing/unknown gender.
Birthdate:	year_of_birth	birthdate	Take year from birthdate	
	month_of_birth	birthdate	Take month from birthdate	
	day_of_birth	birthdate	Take day from birthdate	
	birth_datetime	birthdate	With midnight as time 00:00:00	
Race:	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:	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	

ETL Implementation

person



How should the PERSON table logic be implemented in SQL?



```
truncate cdm_lauren.person;
insert into cdm_lauren.person (
            person_id,
            ...
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
    date_part('month', p.birthdate) as month_of_birth,
    date_part('day', p.birthdate) as day_of_birth
```

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	

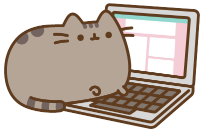
```
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else 0
    end as ethnicity_concept_id,
    ...
```

Ethnicity

??

ETL Implementation

person



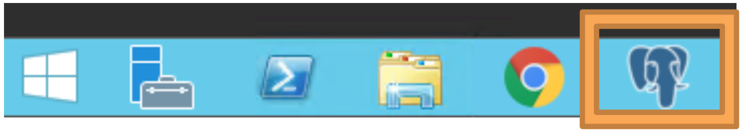
Now let us run the code and create the PERSON table in the cdm_lauren schema

1. Download the query from:

   https://github.com/OHDSI/Tutorial-ETL

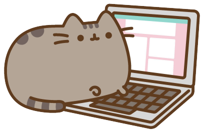
   Materials → Implementation→ Insert_Person_Lauren.sql

2. Open up pgAdmin4 using the icon on the task bar

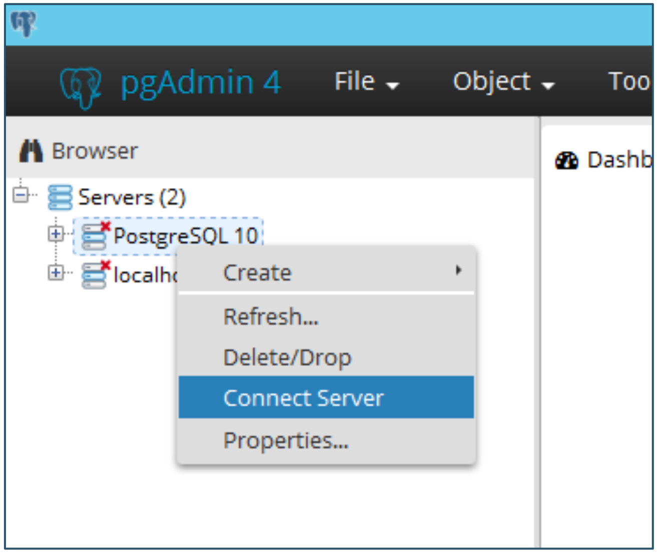


ETL Implementation

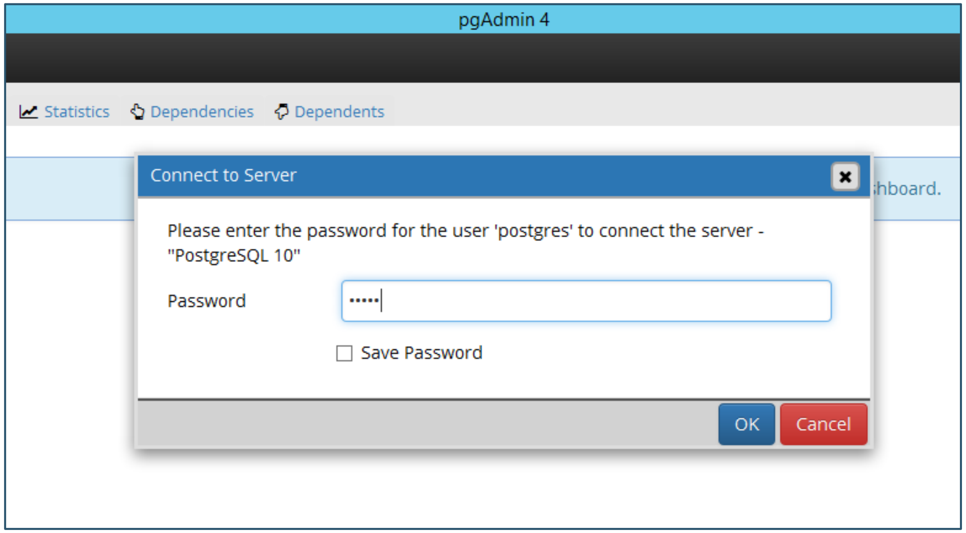
person



3. Expand the server list and right-click on PostgreSQL 10 and choose Connect Server from the drop-down menu



4. When it asks for a password, type in ohdsi



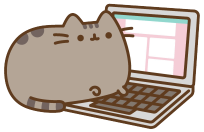
Please enter the password for the user 'postgres' to connect the server - "PostgreSQL 10"

Password •••••

□ Save Password

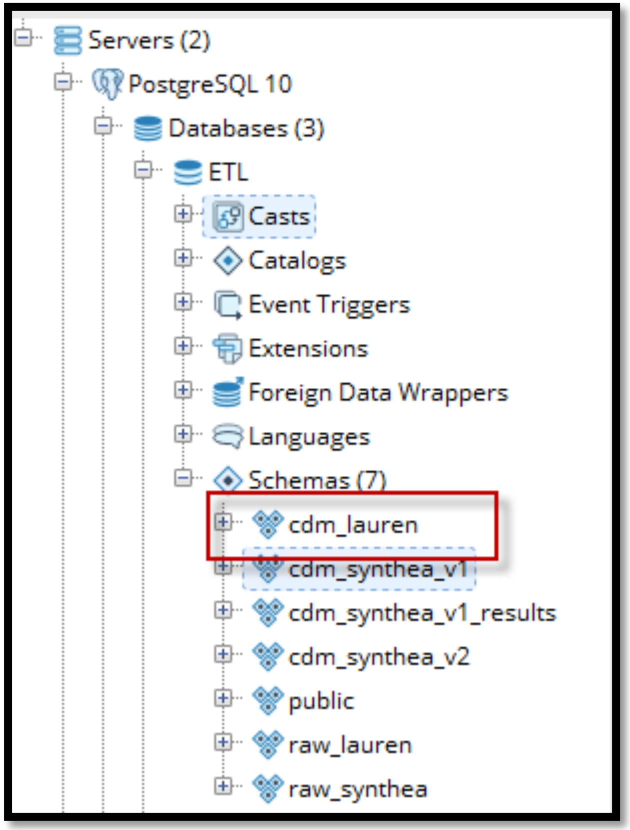
ETL Implementation

person



- Open up to and select the CDM (which has a copy of the vocab)
- Tools → Query Tool
- Type the following and hit F5 to run:

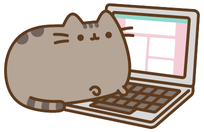
```
SET SEARCH_PATH TO
CDM_LAUREN;
```

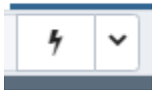


- Servers (2)
  - PostgreSQL 10
    - Databases (3)
      - ETL
        - Casts
        - Catalogs
        - Event Triggers
        - Extensions
        - Foreign Data Wrappers
        - Languages
        - Schemas (7)
          - cdm_lauren
          - cdm_synthea_v1
          - cdm_synthea_v1_results
          - cdm_synthea_v2
          - public
          - raw_lauren
          - raw_synthea

ETL Implementation

person



7. Paste the sql code to create the PERSON table into the query window and press F5 or 

NOTE:

The ‘truncate’ statement at the beginning deletes anything that is in the table already without deleting the table itself (helpful if you make a mistake)

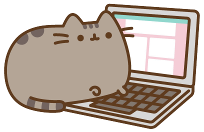
QUESTIONS:

How would you check that your PERSON table was created?

How could you fix the ethnicity mapping?

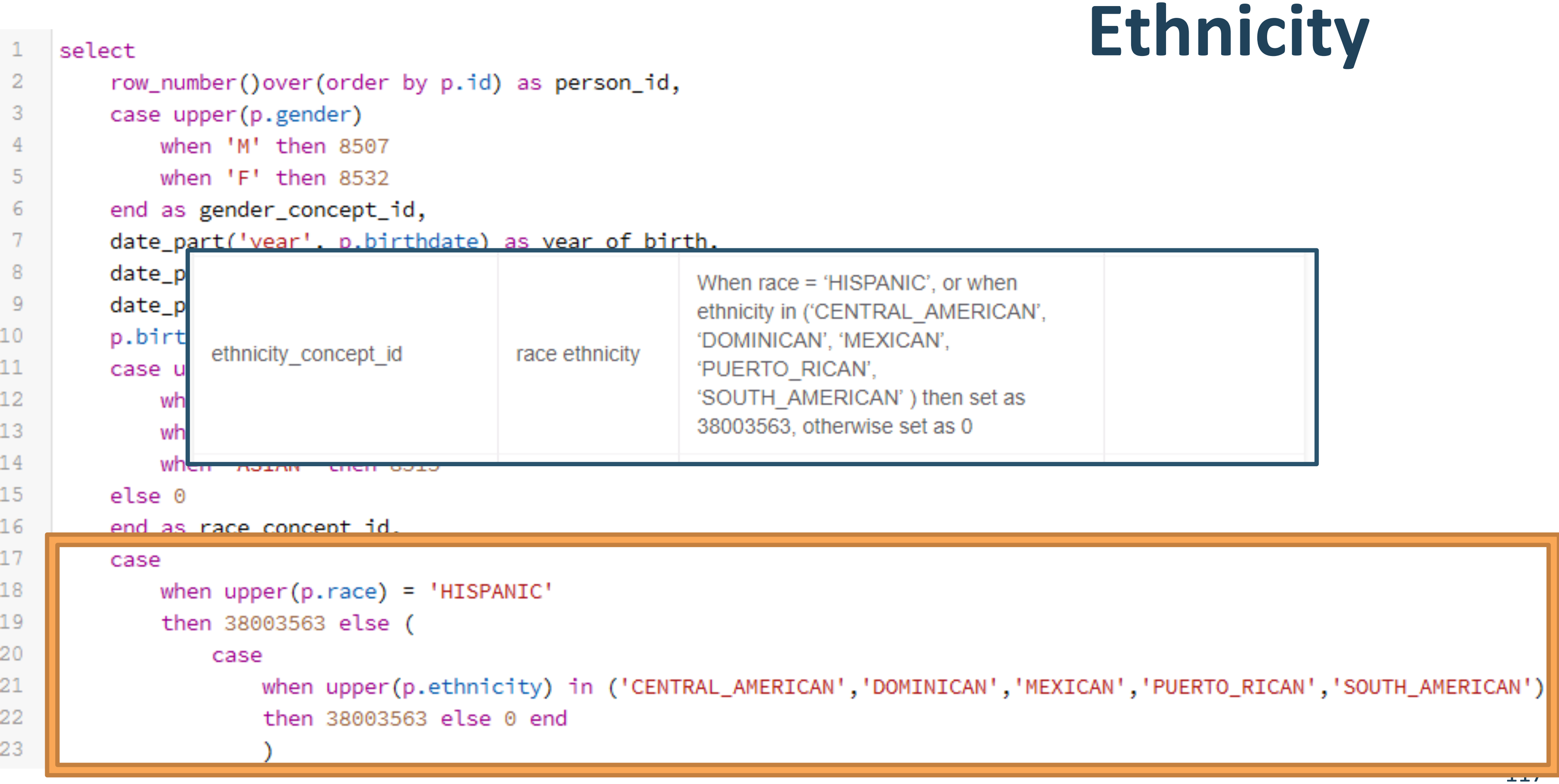
ETL Implementation

person



Data Quality at implantation – ethnicity correction

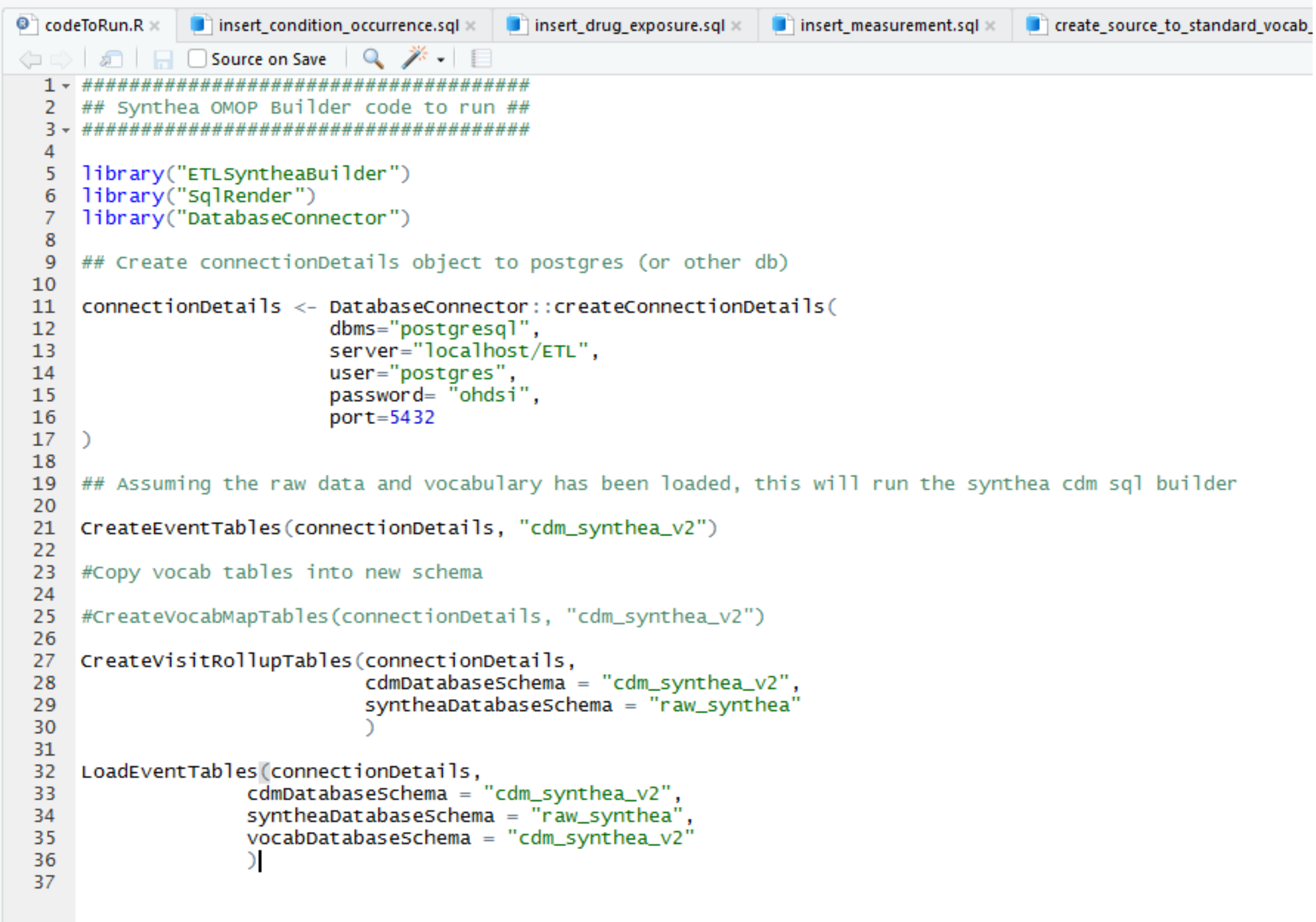
Ethnicity



```
select
    row_number()over(order by p.id) as person_id,
    case upper(p.gender)
        when 'M' then 8507
        when 'F' then 8532
    end as gender_concept_id,
    date_part('year', p.birthdate) as year_of_birth,
    date_p
    date_p
    p.birt
    case u
        wh
        wh
        when 'ASIAN' then 8515
    else 0
    end as race_concept_id,
    case
        when upper(p.race) = 'HISPANIC'
        then 38003563 else (
            case
                when upper(p.ethnicity) in ('CENTRAL_AMERICAN','DOMINICAN','MEXICAN','PUERTO_RICAN','SOUTH_AMERICAN')
                then 38003563 else 0 end
                )
```

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	

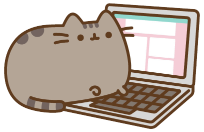
Build the rest of the tables . . .



```
######################################
## Synthea OMOP Builder code to run ##
######################################

library("ETLSyntheaBuilder")
library("SqlRender")
library("DatabaseConnector")

## Create connectionDetails object to postgres (or other db)

connectionDetails <- DatabaseConnector::createConnectionDetails(
                     dbms="postgresql",
                     server="localhost/ETL",
                     user="postgres",
                     password= "ohdsi",
                     port=5432
)

## Assuming the raw data and vocabulary has been loaded, this will run the synthea cdm sql builder

CreateEventTables(connectionDetails, "cdm_synthea_v2")

#Copy vocab tables into new schema

#CreateVocabMapTables(connectionDetails, "cdm_synthea_v2")

CreateVisitRollupTables(connectionDetails,
                        cdmDatabaseSchema = "cdm_synthea_v2",
                        syntheaDatabaseSchema = "raw_synthea"
                        )

LoadEventTables(connectionDetails,
              cdmDatabaseSchema = "cdm_synthea_v2",
              syntheaDatabaseSchema = "raw_synthea",
              vocabDatabaseSchema = "cdm_synthea_v2"
              )
```

github.com/ohdsi/ETL-synthea

Resources



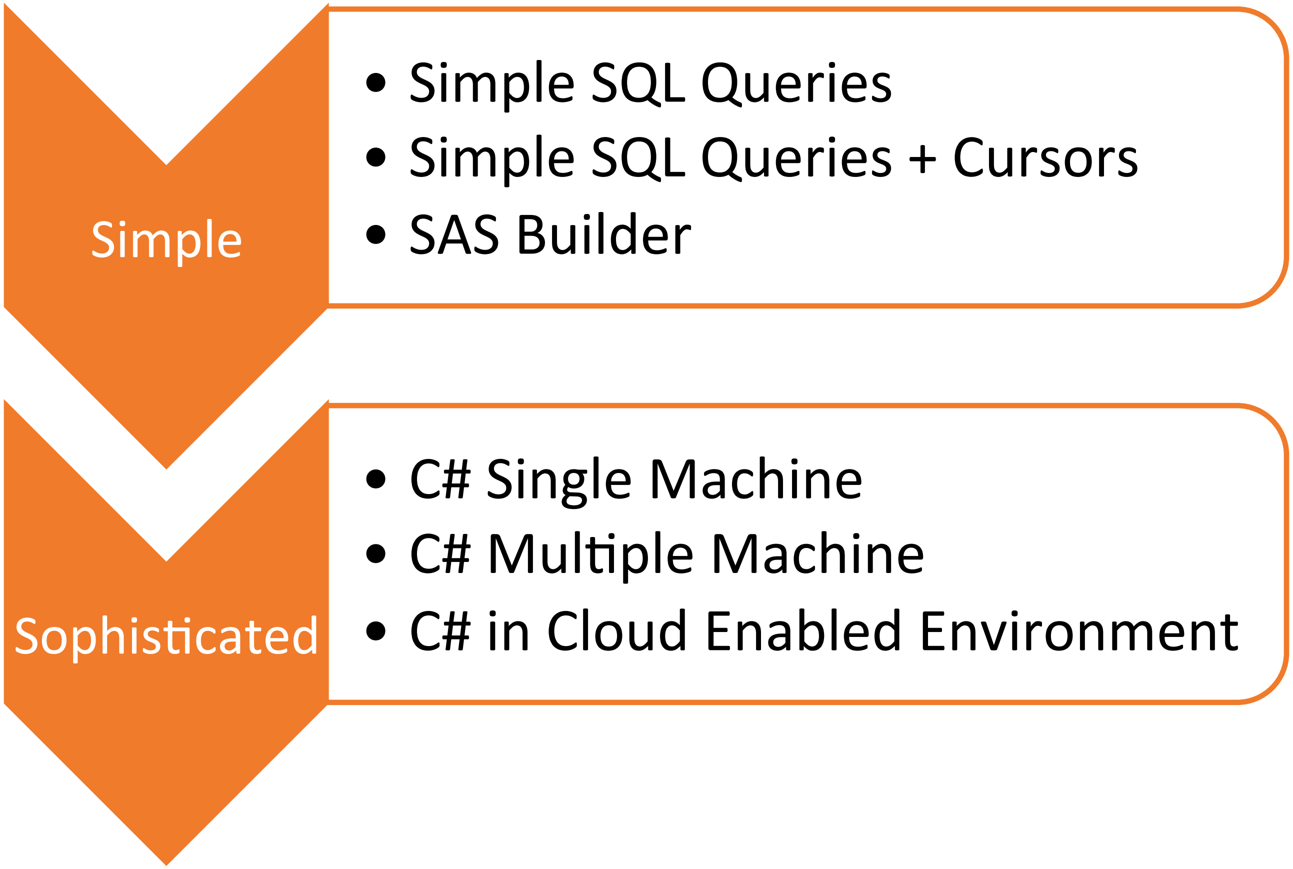
- The full Synthea builder can be found here:
  https://github.com/OHDSI/ETL-Synthea

- Another example of a R/SQL builder for a much larger database:
  https://github.com/OHDSI/ETL-HealthVerityBuilder

- A builder created using .NET:
  https://github.com/OHDSI/ETL-CDMBuilder

- A builder created using the AWS lambda functionality:
  https://github.com/OHDSI/ETL-lambdabuilder
  (in development)

Example Builder 1:
Janssen CDM Builder Over Time



Simple

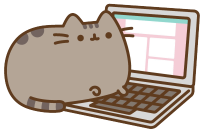
- Simple SQL Queries
- Simple SQL Queries + Cursors
- SAS Builder

Sophisticated

- C# Single Machine
- C# Multiple Machine
- C# in Cloud Enabled Environment

Data Experts
& CDM Experts

Data Experts
& CDM Experts



Technical
Experts

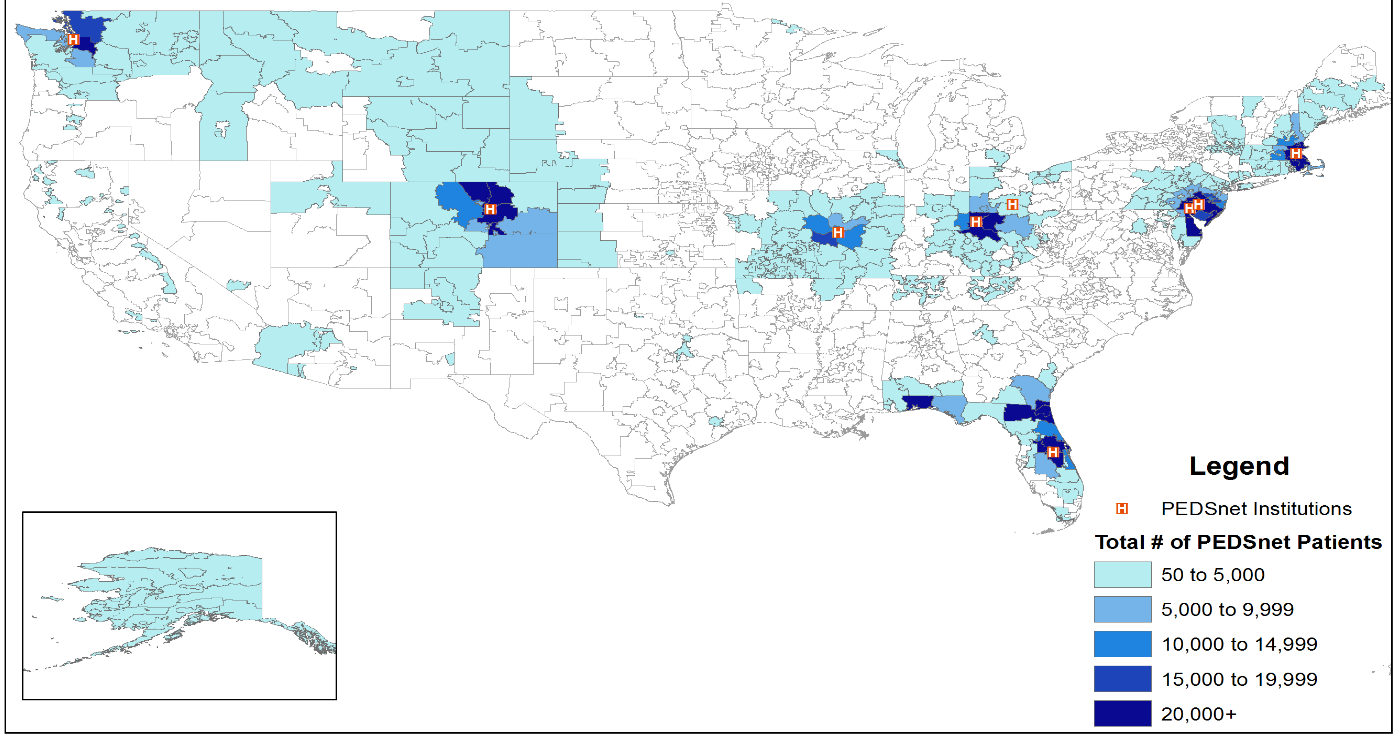
https://github.com/OHDSI/ETL-CDMBuilder

Example Builder 2: PEDSnet

PEDSnet (n=6.2 million patients)
8 contributing sites

Data Coordinating Center:
Children's Hospital of
Philadelphia (CHOP)

January 2009 to September 2014



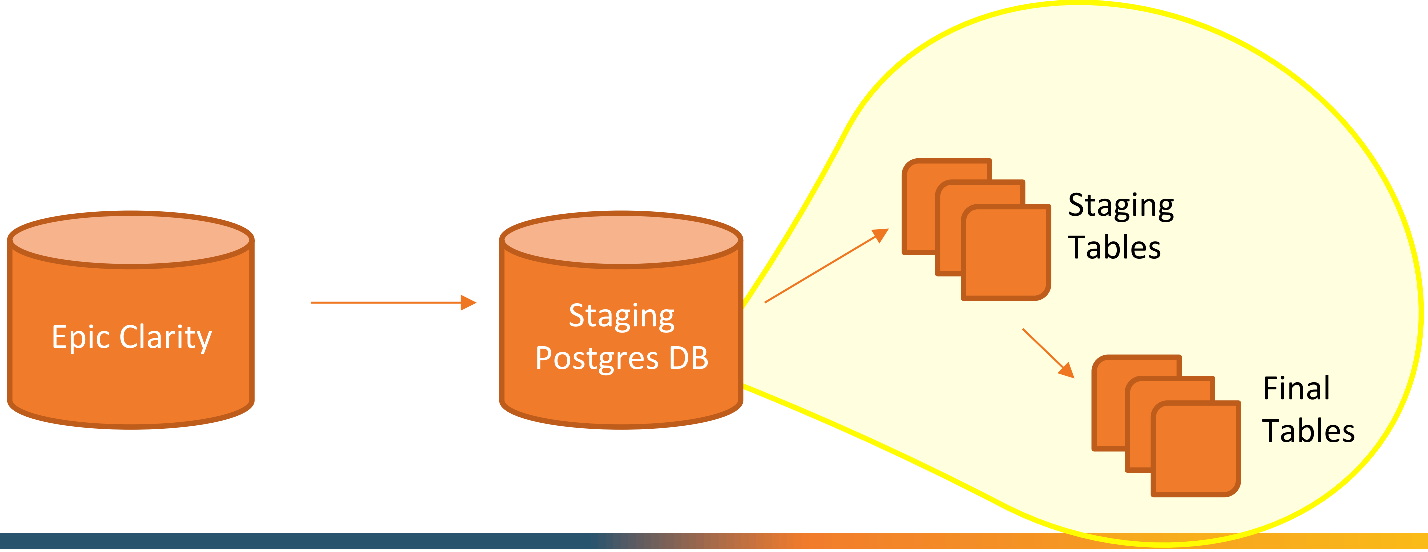
CHOP

Children's Hospital of Philadelphia

- Children's Hospital of Philadelphia
  - – Data Coordinating Center (quarterly submissions)
    - PEDSnet DDL
    - ETL Conventions
    - Data Quality
    - Data Science
  - – Also, a submitting site:
    - ~ 1.2 million patients
    - ~ 55 million visits
    - ~ 700 million clinical facts

CHOP ETL Flow –More like LTE

L	• Load (very little re-organization of data)
T	• Transform (Mapping of concepts, remapping ETL)
E	• Extract to final PEDSnet (OMOP-like) tables



Challenges/Lessons Learned

- We ultimately have to make decisions about our data:
  - What do we include?
    - Cancelled visits with associated information, reflects known workflow for research visits
  - What data do we exclude?
    - Cancelled Labs, Procedures
    - Test patients
    - Patients with lab only data (Adults lab/blood work, genetics)
  - Who makes these decisions?
    - Data Committee/Data Modeling Working Group
    - Local Informaticist and Analyst team

Challenges/Lessons Learned

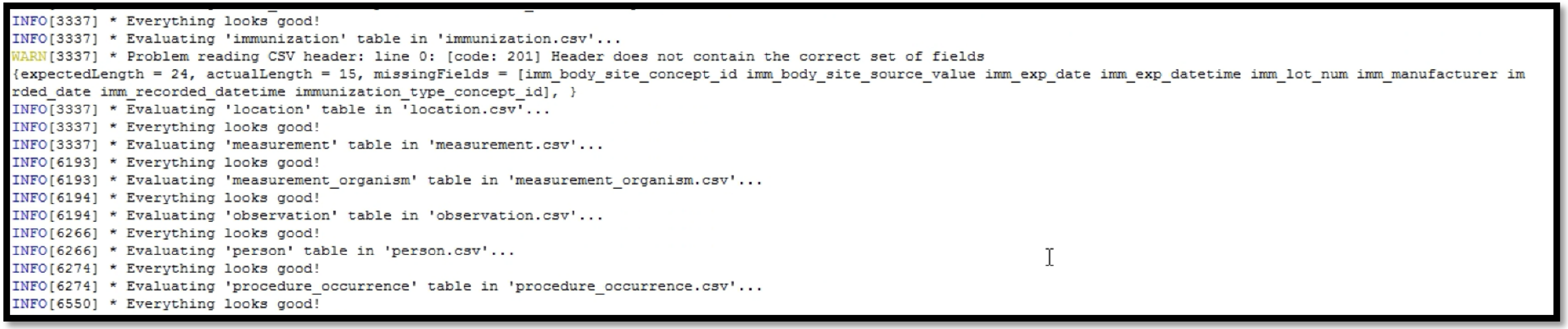
- Our ETL is time-constraint due to clinical system ETL
  - Structured program to take into account midnight system wide shutdown for ETL

- Clinical data does not always fit OMOP rules
  - Multivitamin prescriptions with 2055 `end_date`
  - Fetal Procedures `procedure_start_date` before `birth_date`
  - Autopsies procedures `procedure_start_date` after `death_date`
  - Multiple “encounters” associated with one visit

- Intermediate/Temporary tables are crucial for debugging
  - Tables containing source identification numbers (IDS such as MRNS, patient ids, source system ids) alongside OMOP data before “final version”

Data Validation: Data Model Validator

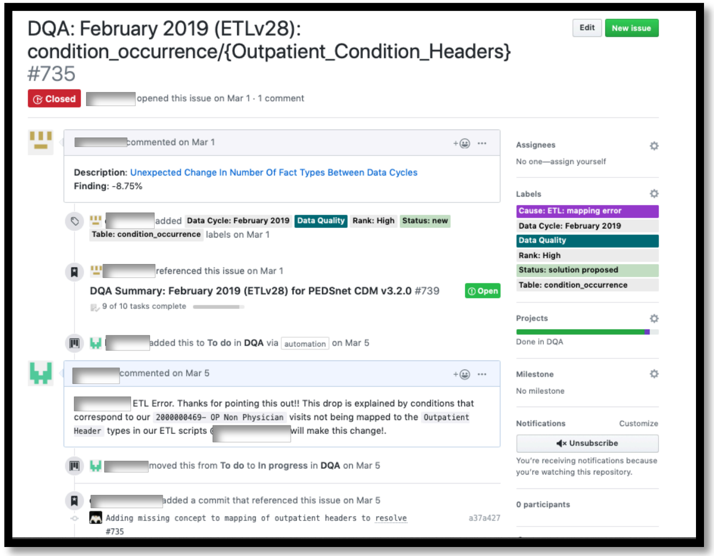
- Validates table structures and data types
- Prompts user to specify the model and version number
- Alerts if there are any unexpected columns and/or tables
- https://github.com/infomodels/infomodels (OMOP model supported)



```
INFO[3337] * Everything looks good!
INFO[3337] * Evaluating 'immunization' table in 'immunization.csv'...
WARN[3337] * Problem reading CSV header: line 0: [code: 201] Header does not contain the correct set of fields
{expectedLength = 24, actualLength = 15, missingFields = [imm_body_site_concept_id imm_body_site_source_value imm_exp_date imm_exp_datetime imm_lot_num imm_manufacturer im
rded_date imm_recorded_datetime immunization_type_concept_id], }
INFO[3337] * Evaluating 'location' table in 'location.csv'...
INFO[3337] * Everything looks good!
INFO[3337] * Evaluating 'measurement' table in 'measurement.csv'...
INFO[6193] * Everything looks good!
INFO[6193] * Evaluating 'measurement_organism' table in 'measurement_organism.csv'...
INFO[6194] * Everything looks good!
INFO[6194] * Evaluating 'observation' table in 'observation.csv'...
INFO[6266] * Everything looks good!
INFO[6266] * Evaluating 'person' table in 'person.csv'...
INFO[6274] * Everything looks good!
INFO[6274] * Evaluating 'procedure_occurrence' table in 'procedure_occurrence.csv'...
INFO[6550] * Everything looks good!
```

Data Validation: Data Quality Framework

- Automated Program where issues are flagged as GitHub issues categorized by table, domain and priority (High, Medium, Low)
- Checks fall into the following categories:
  - **Fidelity/Reliability:** Is this data correct? Is it being coded/mapped correctly?
  - **Consistency/Internal Validity:** Are there any drops/inconsistencies between submissions?
  - **Accuracy:** Does the data correctly reflect clinical characteristics of patients?
  - **Completeness :** Is there data that is missing?
  - https://pedsnet.org/data/data-quality/



DQA: February 2019 (ETLv28): condition_occurrence/{Outpatient_Condition_Headers} #735

[redacted] opened this issue on Mar 1 · 1 comment

[redacted] commented on Mar 1

Description: Unexpected Change In Number Of Fact Types Between Data Cycles
Finding: -8.75%

[redacted] added **Data Cycle: February 2019** **Data Quality** **Rank: High** **Status: new** **Table: condition_occurrence** labels on Mar 1

[redacted] referenced this issue on Mar 1

DQA Summary: February 2019 (ETLv28) for PEDSnet CDM v3.2.0 #739

9 of 10 tasks complete

[redacted] added this to **To do** in **DQA** via automation on Mar 5

[redacted] commented on Mar 5

[redacted] ETL Error. Thanks for pointing this out!! This drop is explained by conditions that correspond to our `2000000469- OP Non Physician` visits not being mapped to the `Outpatient Header` types in our ETL scripts [redacted] will make this change!.

[redacted] moved this from **To do** to **In progress** in **DQA** on Mar 5

[redacted] added a commit that referenced this issue on Mar 5

`Adding missing concept to mapping of outpatient headers to resolve #735` a37a427

Assignees
No one—assign yourself

Labels
Cause: ETL: mapping error
Data Cycle: February 2019
Data Quality
Rank: High
Status: solution proposed
Table: condition_occurrence

Projects
Done in DQA

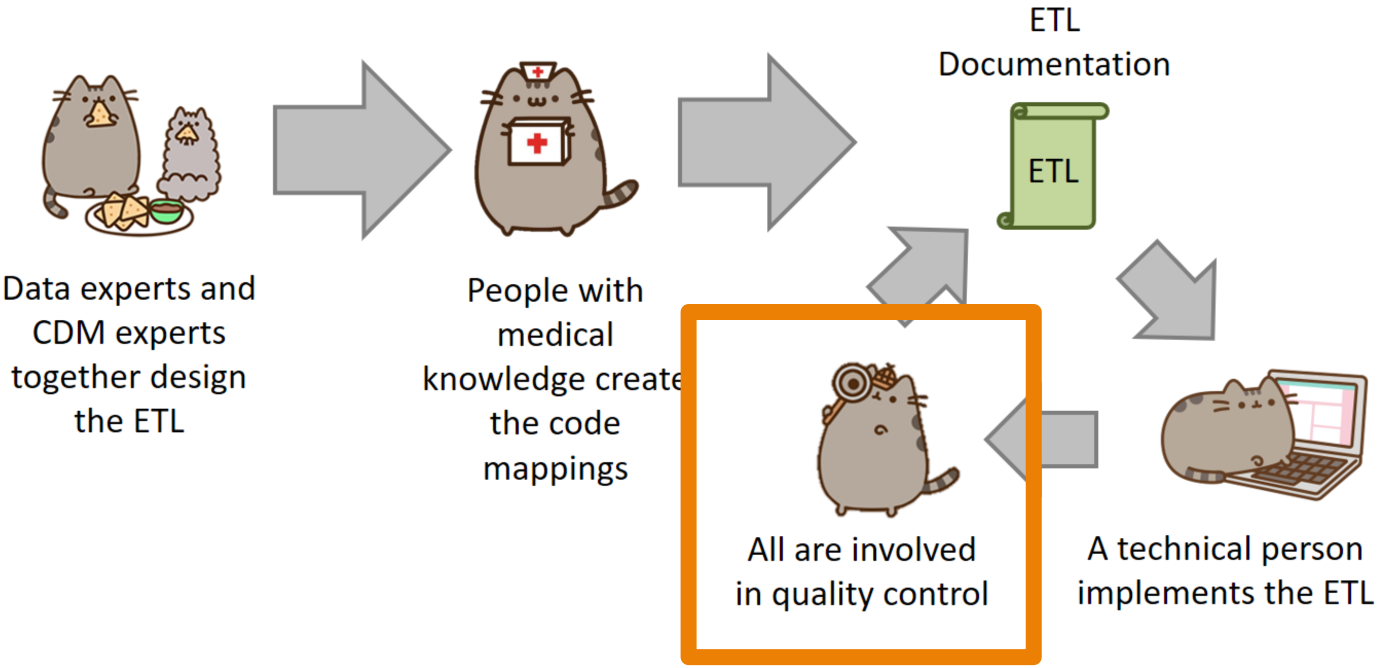
Milestone
No milestone

Notifications Customize
You're receiving notifications because you're watching this repository.

0 participants

OHDSI

Observational Health Data Sciences and Informatics



Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

ETL

A technical person implements the ETL

All are involved in quality control

Quality

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

Rabbit-in-a-Hat Test Case Framework

Achilles

DataQualityDashboard (DQD)

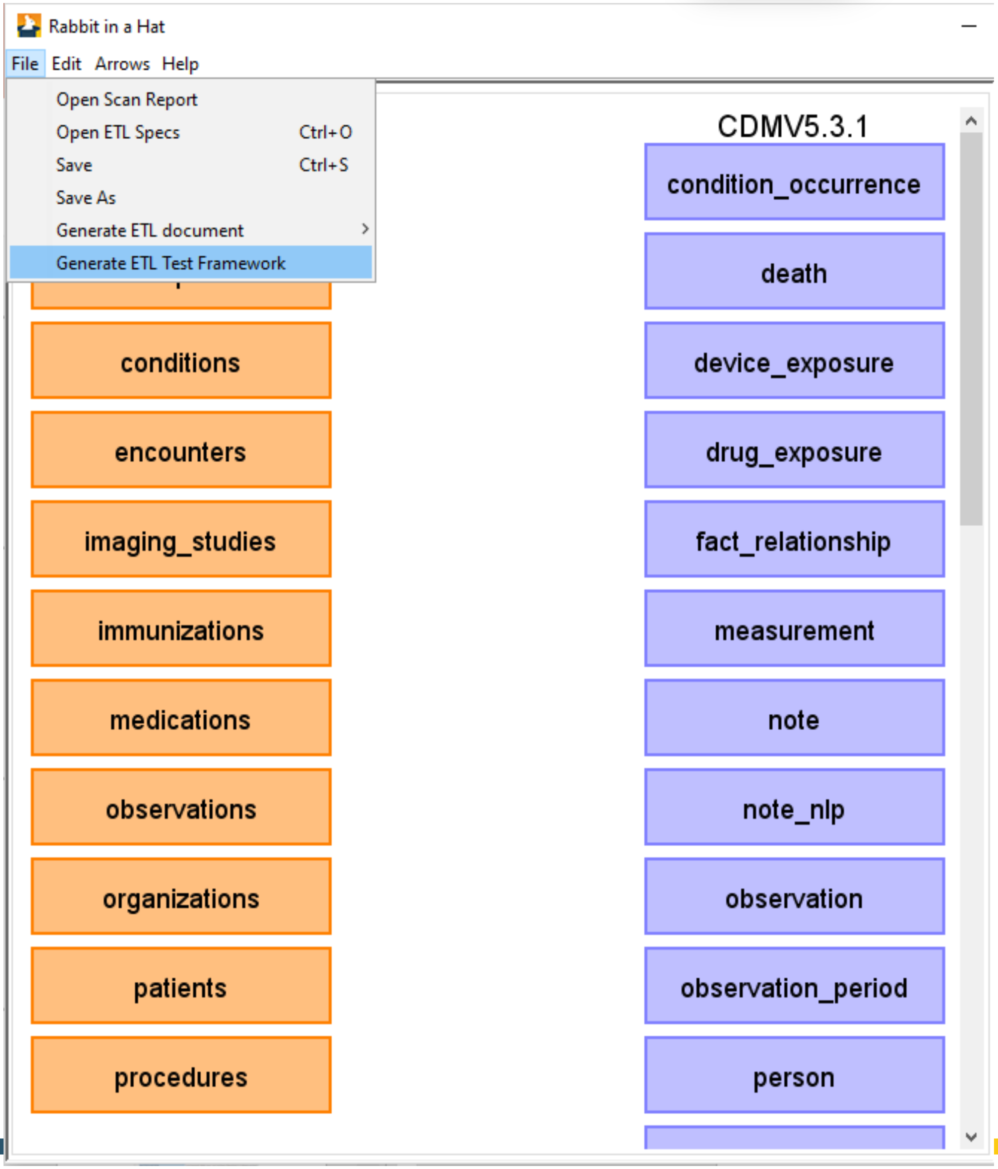
Unit Test Cases

- Testing your CDM builder is important:
  - ETL 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 test, or small pieces of code that can automatically check single aspects of the ETL design

Unit Test Cases

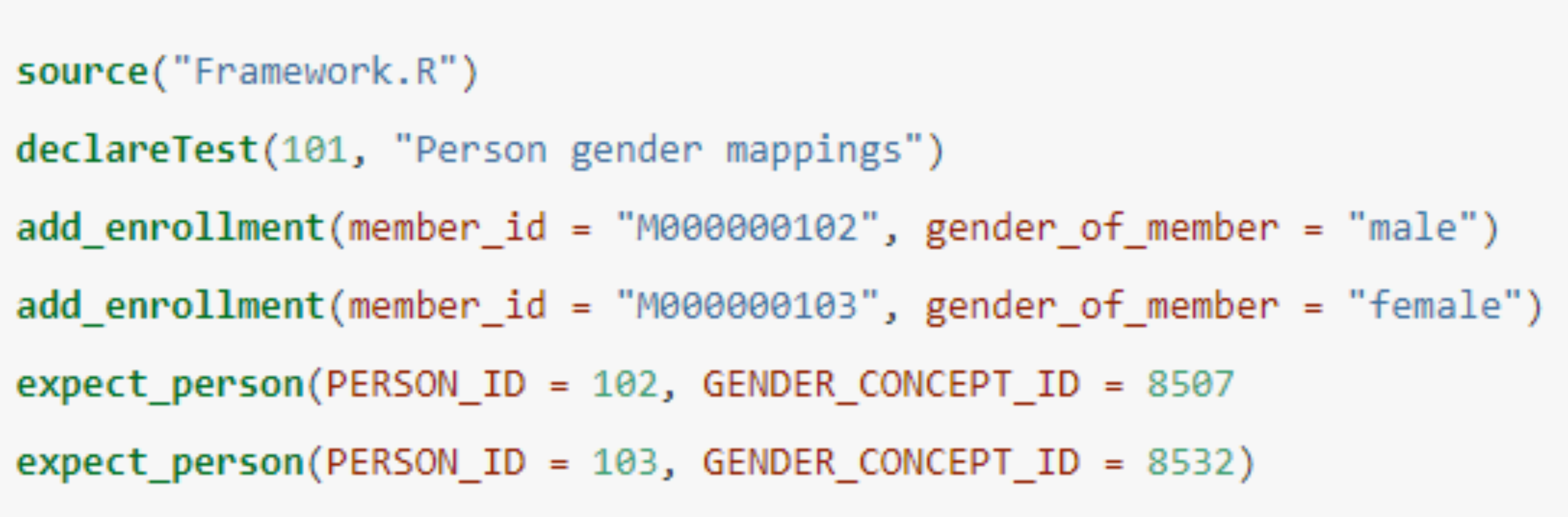
Rabbit-in-a-Hat

The application has a feature called **‘Generate ETL Test Framework’.** This feature allows you to create ‘fake’ people as a way to test your ETL logic.



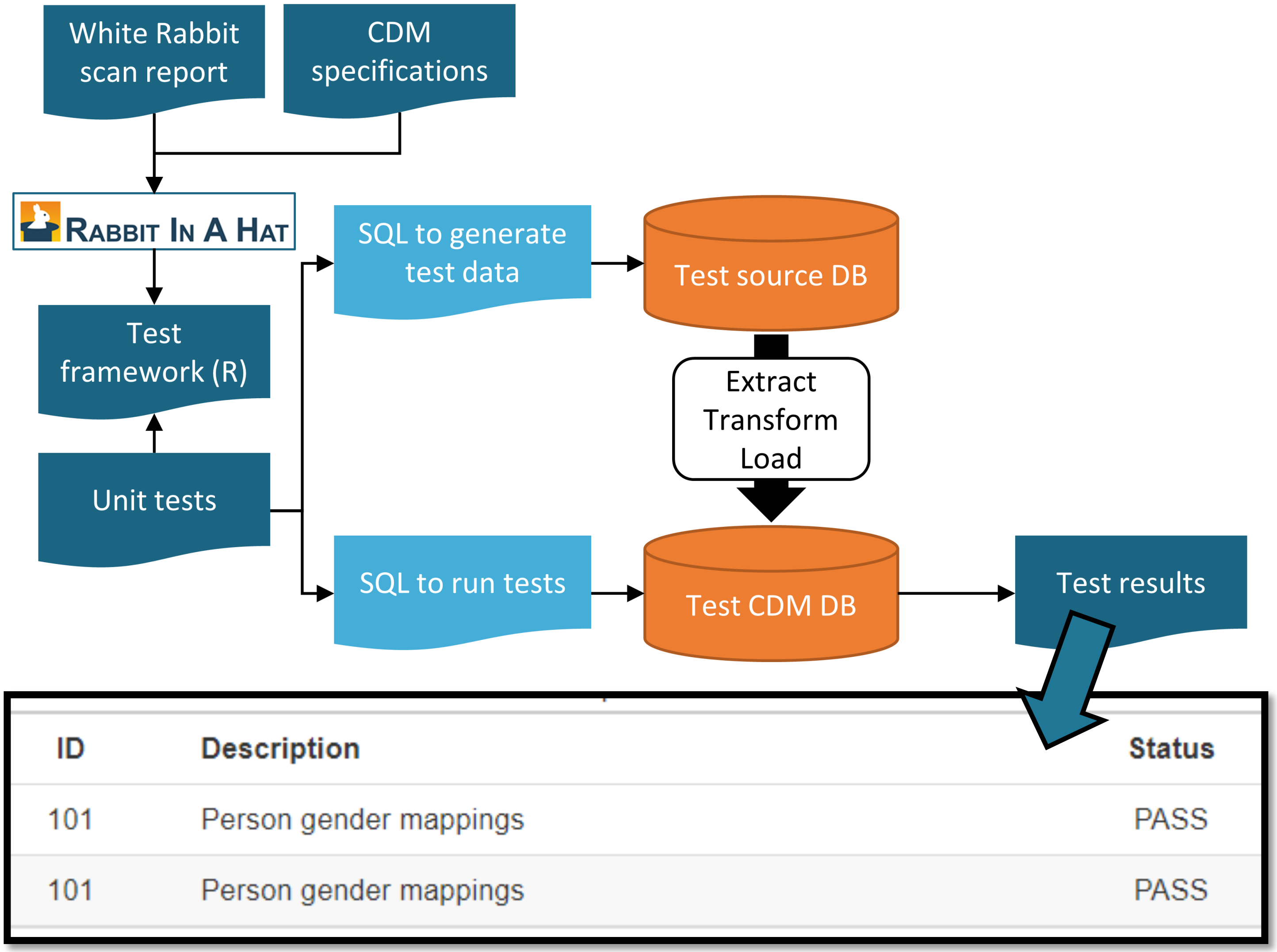
Unit Test Cases

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

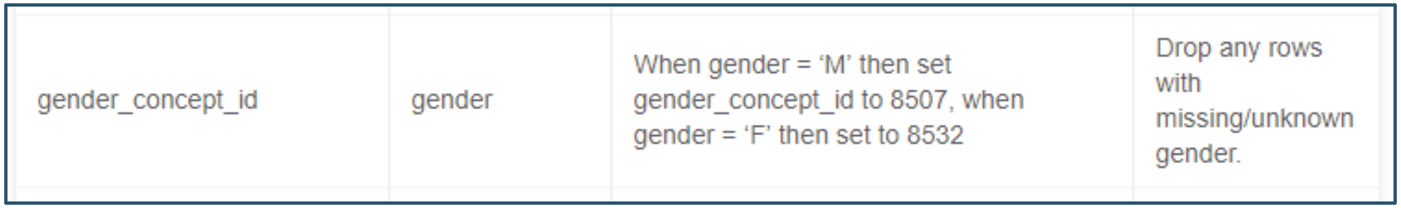
Unit Test Cases

- An example of how this was done for the Synthea data is available from: https://github.com/OHDSI/Tutorial-ETL/tree/master/materials/Unit%20Tests

- The file that creates the test cases as a series of insert statement is **RunSyntheaTestCases.r**

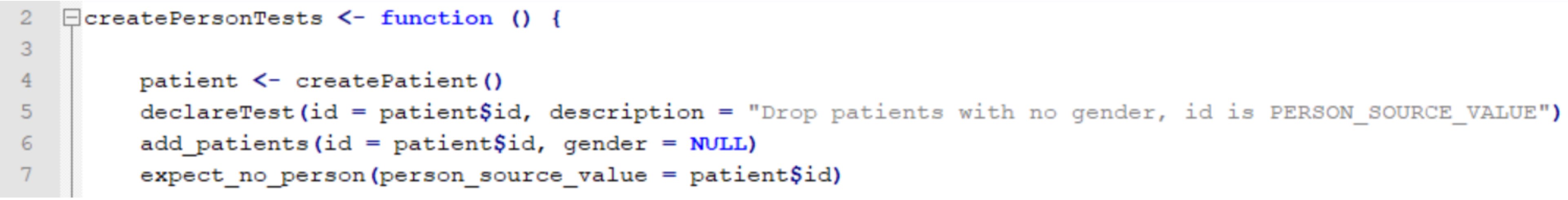
Unit Test Cases

Let us revisit the PERSON table logic:

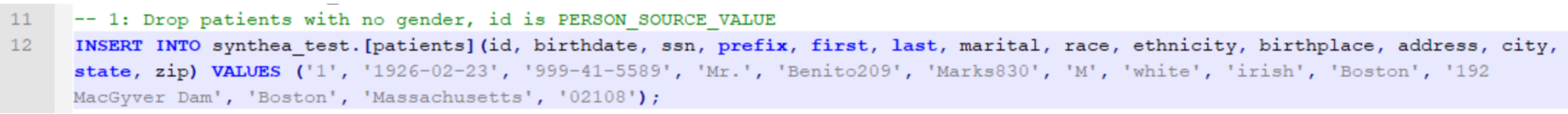


gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.

How could we create a test case for this?



```
createPersonTests <- function () {

    patient <- createPatient()
    declareTest(id = patient$id, description = "Drop patients with no gender, id is PERSON_SOURCE_VALUE")
    add_patients(id = patient$id, gender = NULL)
    expect_no_person(person_source_value = patient$id)
```



```
-- 1: Drop patients with no gender, id is PERSON_SOURCE_VALUE
INSERT INTO synthea_test.[patients](id, birthdate, ssn, prefix, first, last, marital, race, ethnicity, birthplace, address, city,
state, zip) VALUES ('1', '1926-02-23', '999-41-5589', 'Mr.', 'Benito209', 'Marks830', 'M', 'white', 'irish', 'Boston', '192
MacGyver Dam', 'Boston', 'Massachusetts', '02108');
```

Achilles

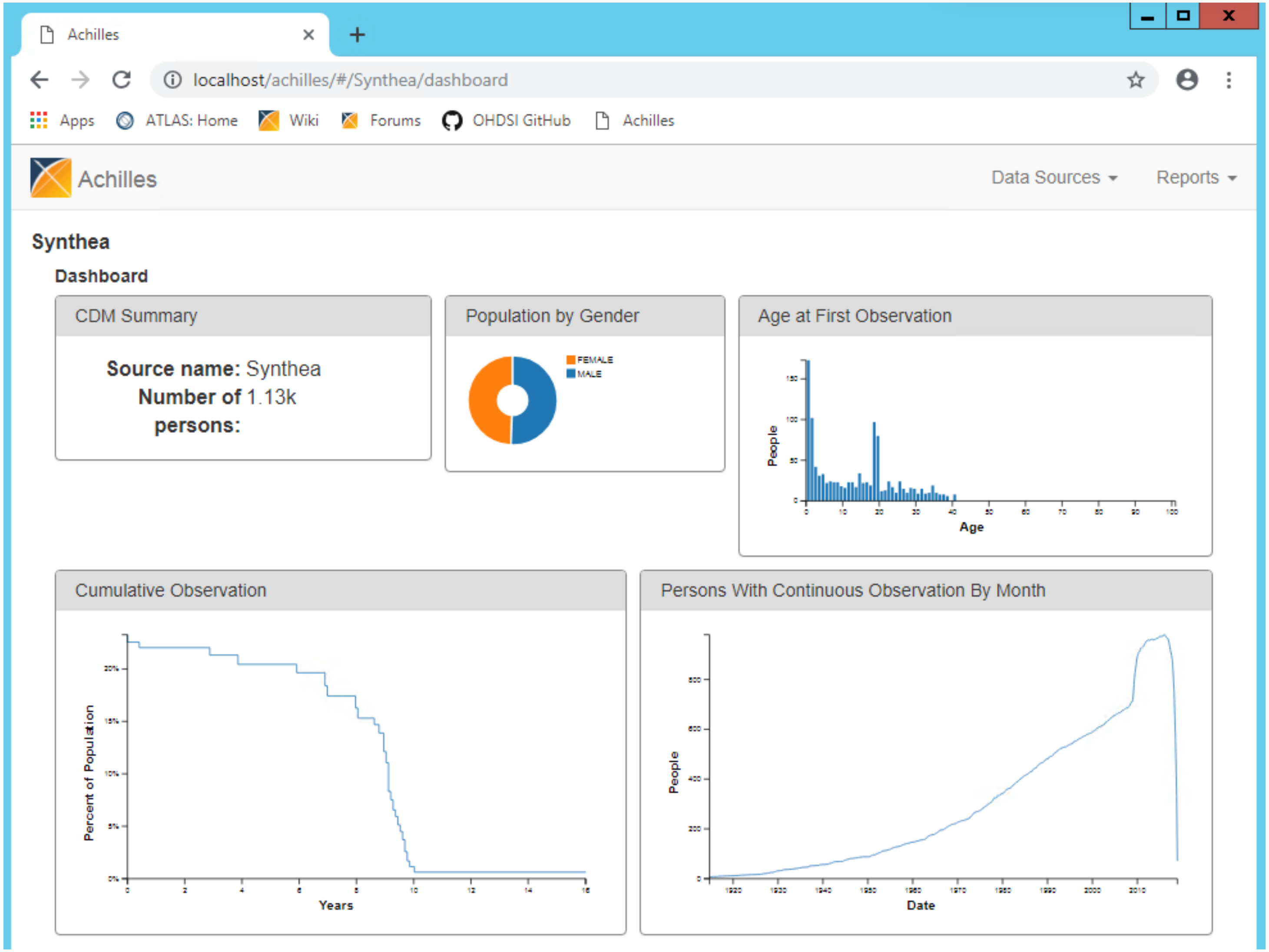
Achilles is a data characterization and quality tool available for download here:

https://github.com/OHDSI/Achilles

For an example of how it was run for our sample data, that R script is located here:

https://github.com/OHDSI/Tutorial-ETL/blob/master/materials/Achilles/achillesRun.R

Achilles



Synthea

Dashboard

CDM Summary

Source name: Synthea
Number of persons: 1.13k

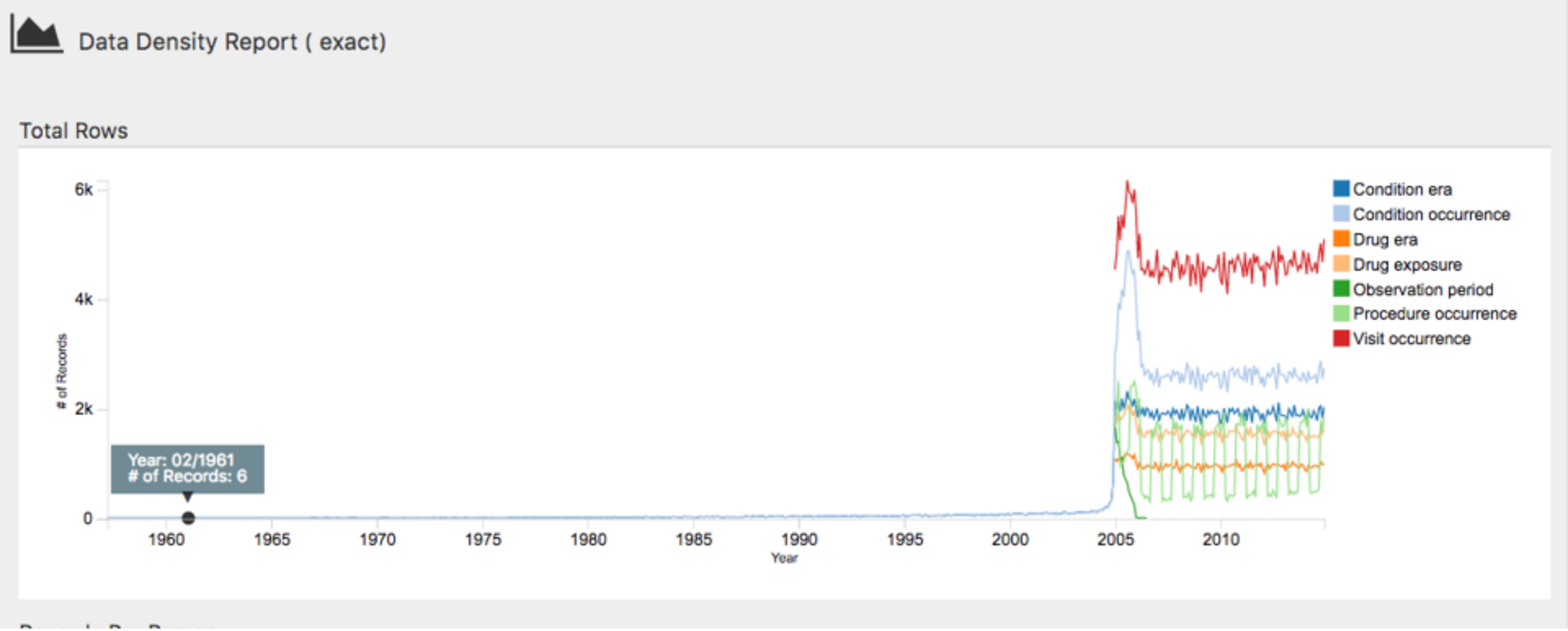
Population by Gender

Age at First Observation

Cumulative Observation

Persons With Continuous Observation By Month

Achilles

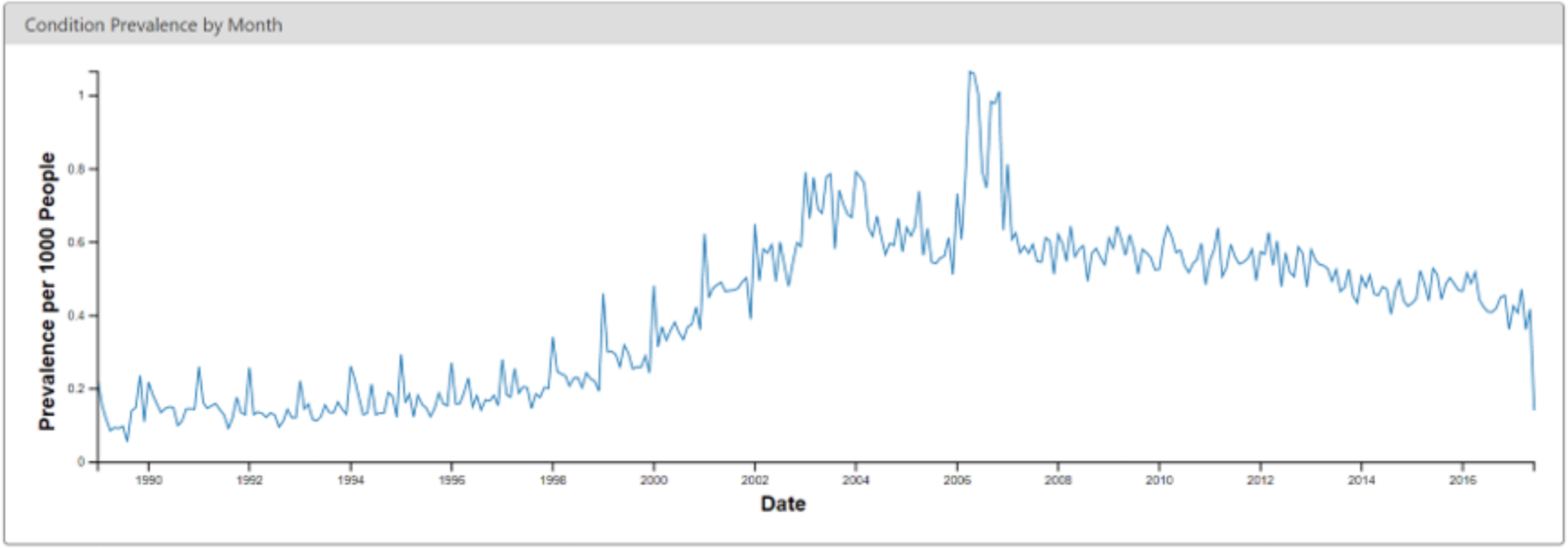


Data Density Report (exact)

Total Rows

This plot shows that the bulk of the data starts in 2005. However, there also appear to be a few records from around 1961, which is likely an error in the data.

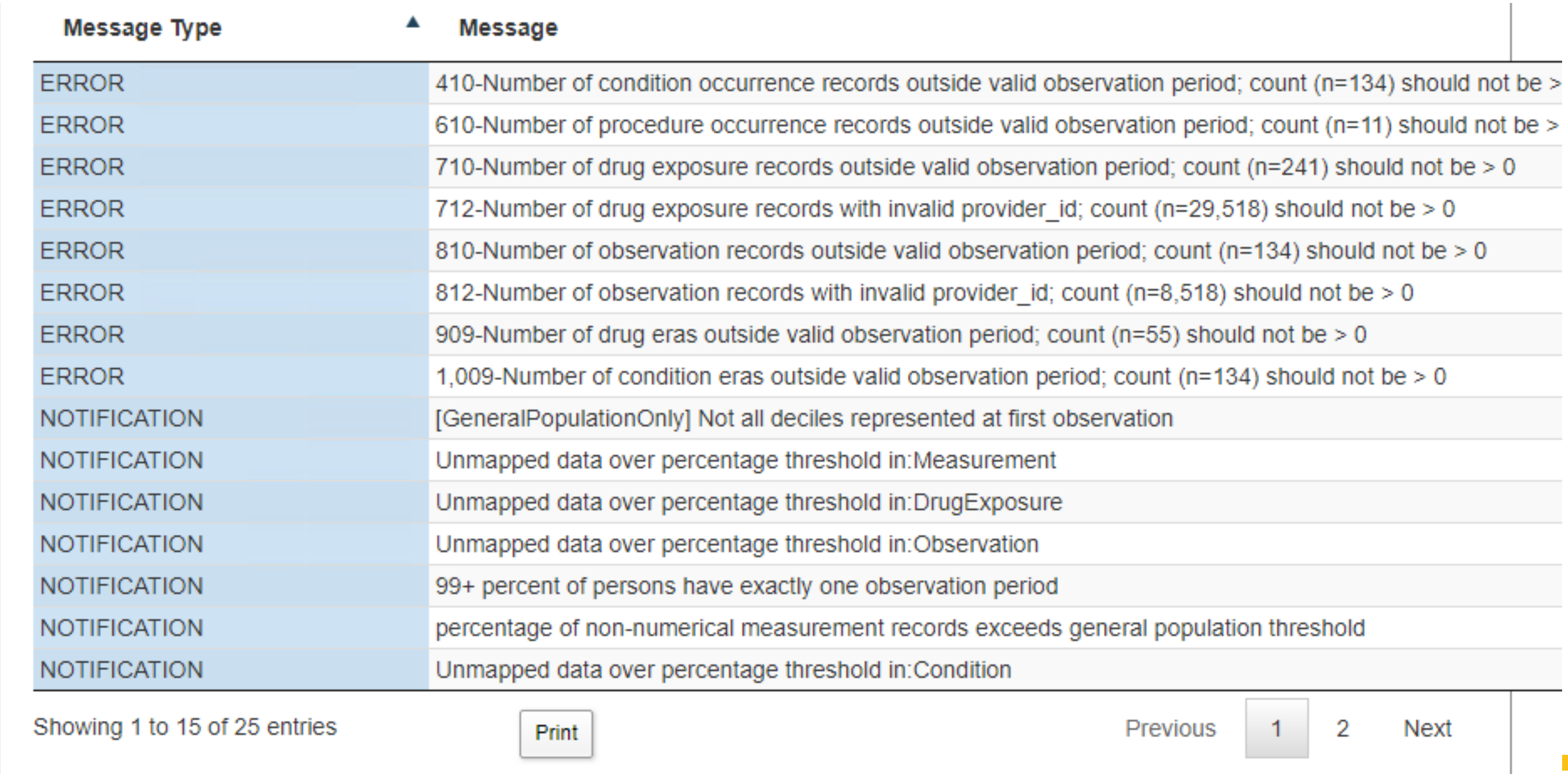
Achilles



This change coincides with changes in the reimbursement rules in this specific country, leading to more diagnoses but probably not a true increase in prevalence in the underlying population.

Achilles Heel

Achilles heel is a report generated by the Achilles application that will run a series of data quality checks on the CDM using the Achilles data

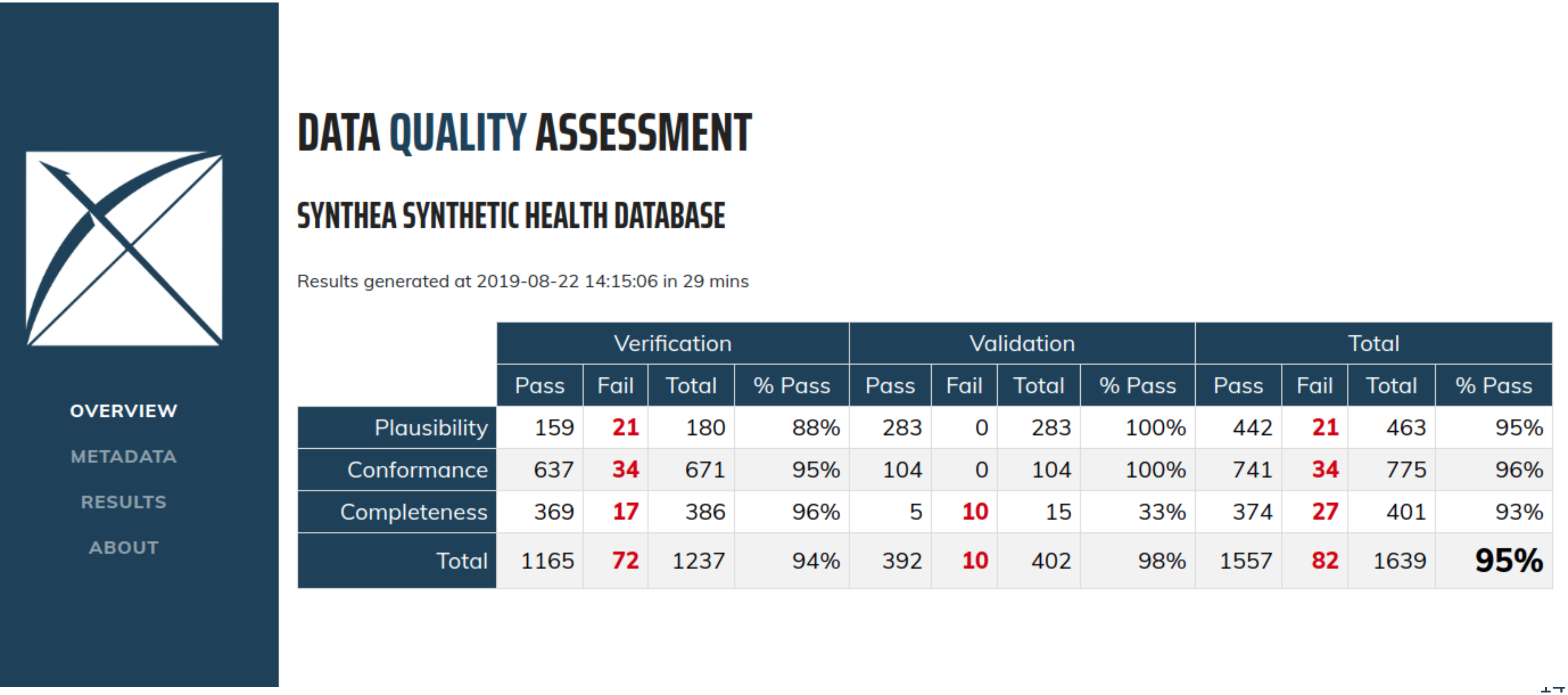


Message Type ▲	Message
ERROR	410-Number of condition occurrence records outside valid observation period; count (n=134) should not be >
ERROR	610-Number of procedure occurrence records outside valid observation period; count (n=11) should not be >
ERROR	710-Number of drug exposure records outside valid observation period; count (n=241) should not be > 0
ERROR	712-Number of drug exposure records with invalid provider_id; count (n=29,518) should not be > 0
ERROR	810-Number of observation records outside valid observation period; count (n=134) should not be > 0
ERROR	812-Number of observation records with invalid provider_id; count (n=8,518) should not be > 0
ERROR	909-Number of drug eras outside valid observation period; count (n=55) should not be > 0
ERROR	1,009-Number of condition eras outside valid observation period; count (n=134) should not be > 0
NOTIFICATION	[GeneralPopulationOnly] Not all deciles represented at first observation
NOTIFICATION	Unmapped data over percentage threshold in:Measurement
NOTIFICATION	Unmapped data over percentage threshold in:DrugExposure
NOTIFICATION	Unmapped data over percentage threshold in:Observation
NOTIFICATION	99+ percent of persons have exactly one observation period
NOTIFICATION	percentage of non-numerical measurement records exceeds general population threshold
NOTIFICATION	Unmapped data over percentage threshold in:Condition

Showing 1 to 15 of 25 entries

DataQualityDashboard (DQD)

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



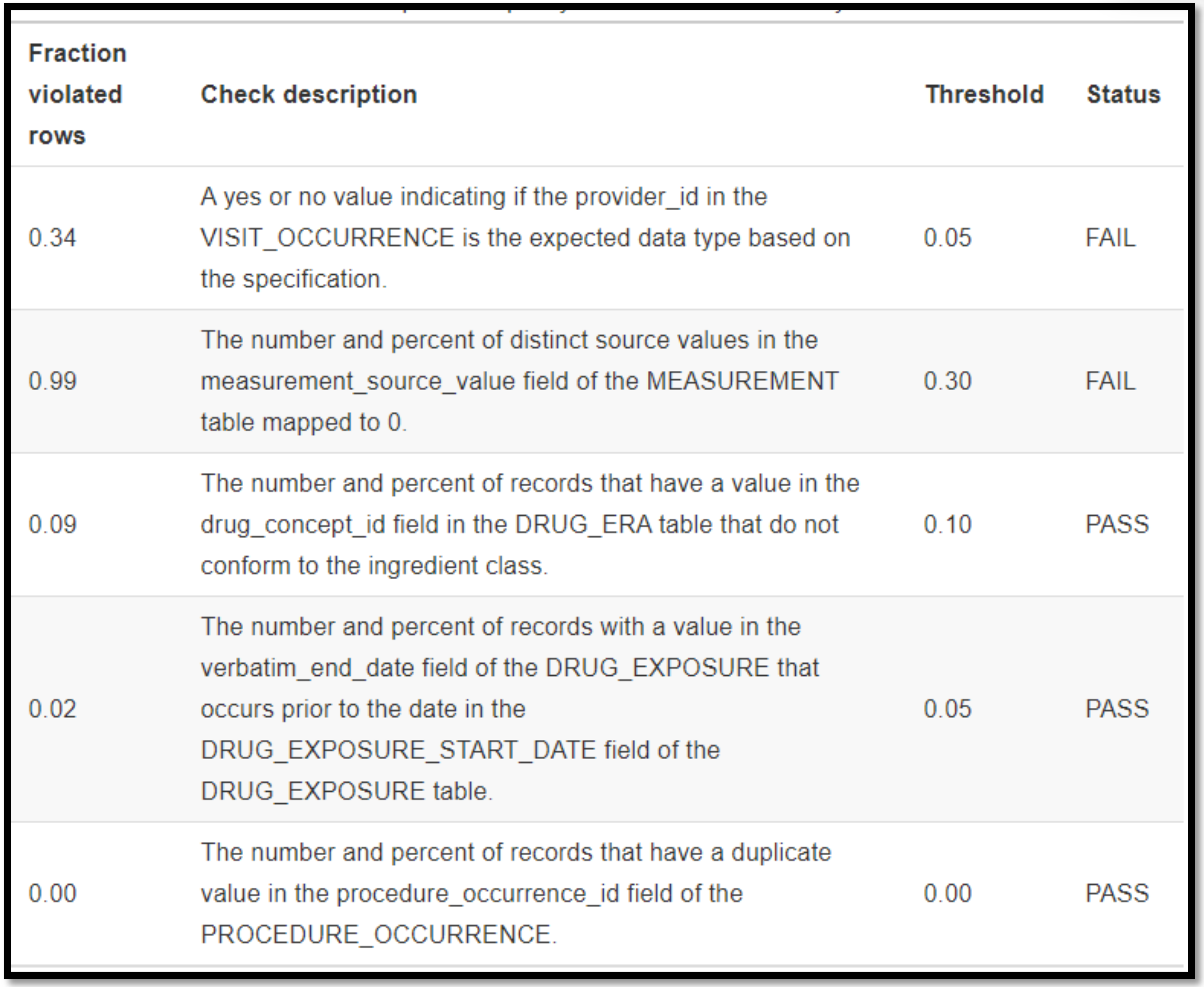
DATA QUALITY ASSESSMENT

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%**

DQD Example Rules

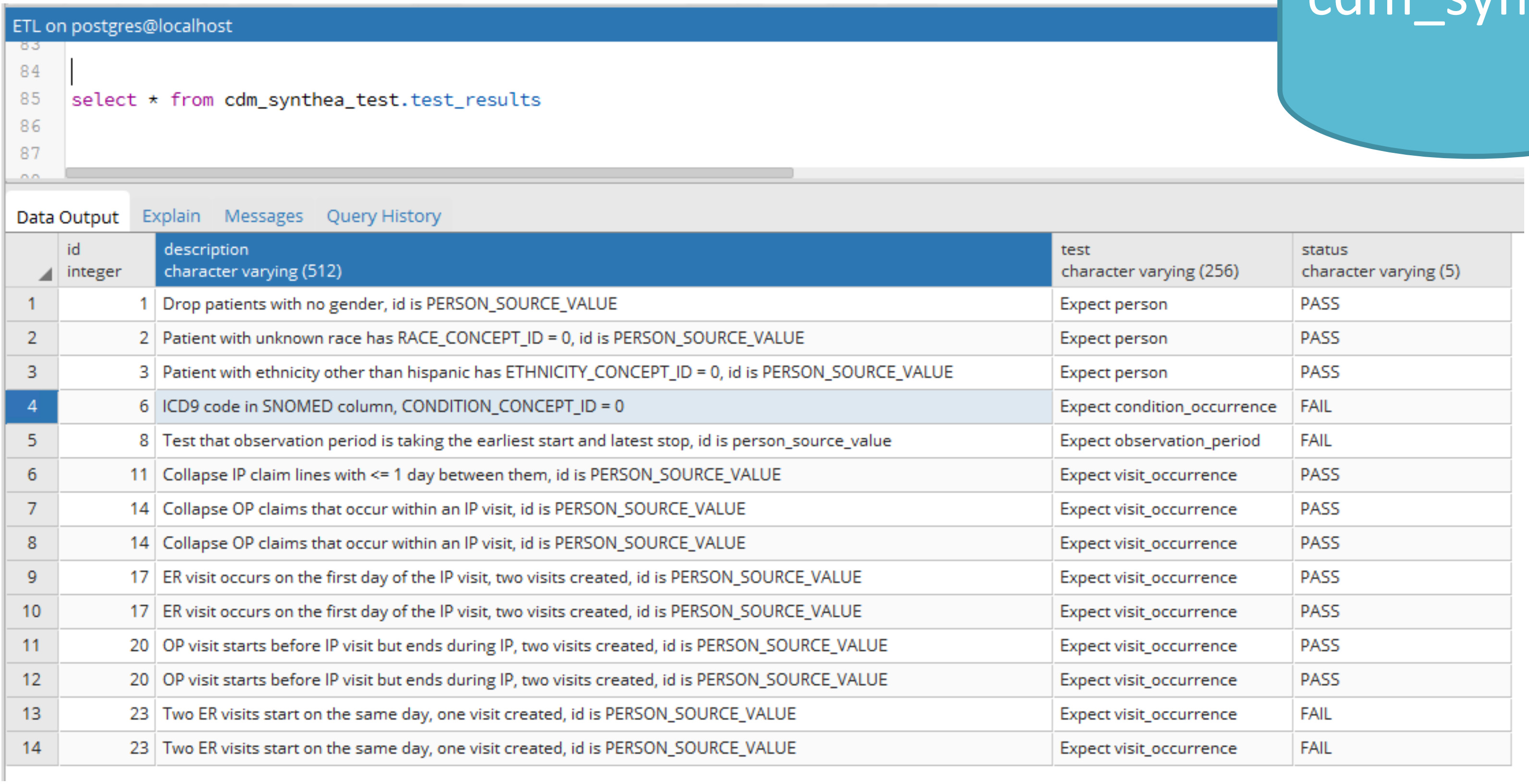


Fraction violated rows	Check description	Threshold	Status
0.34	A yes or no value indicating if the provider_id in the VISIT_OCCURRENCE is the expected data type based on the specification.	0.05	FAIL
0.99	The number and percent of distinct source values in the measurement_source_value field of the MEASUREMENT 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

Issues in our synthetic data?

- Did our test cases run?

cdm_synthea



ETL on postgres@localhost

```
select * from cdm_synthea_test.test_results
```

Data Output | Explain | Messages | Query History

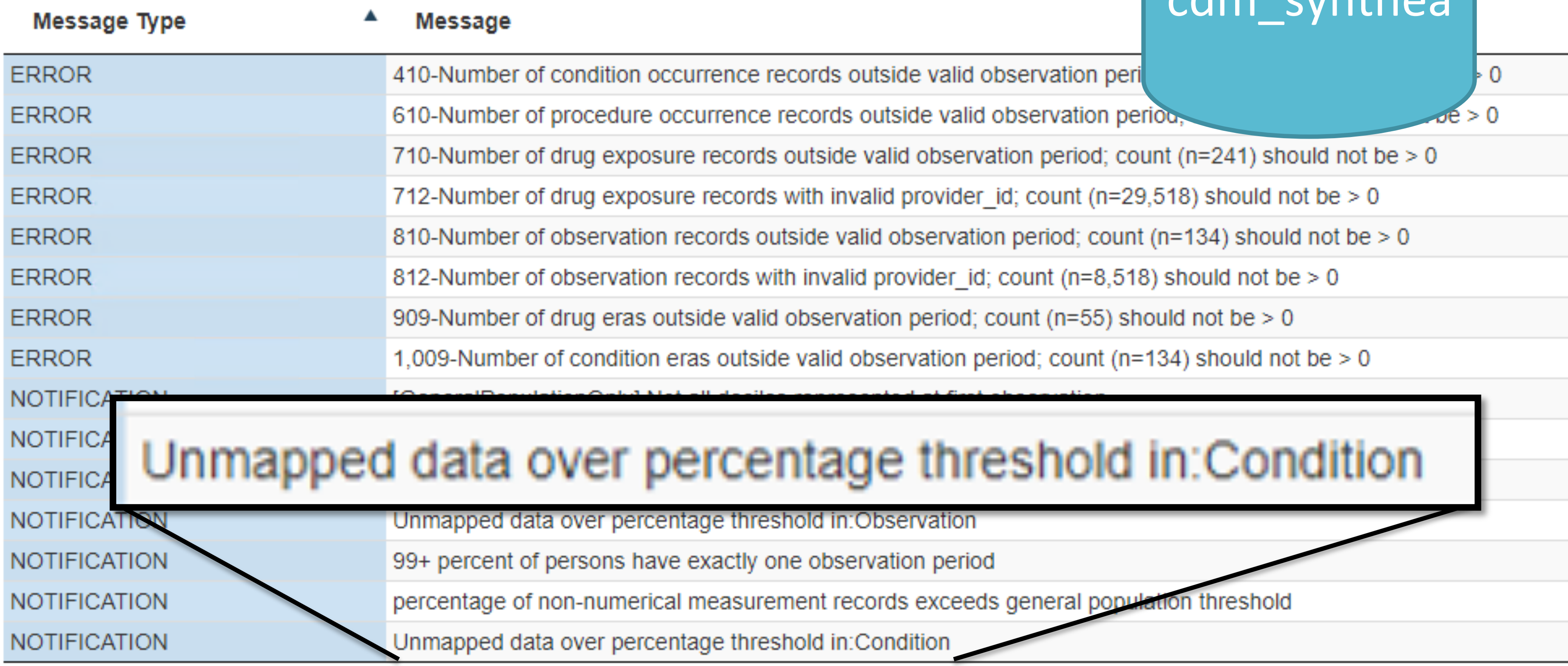
	id integer	description character varying (512)	test character varying (256)	status character varying (5)
1	1	Drop patients with no gender, id is PERSON_SOURCE_VALUE	Expect person	PASS
2	2	Patient with unknown race has RACE_CONCEPT_ID = 0, id is PERSON_SOURCE_VALUE	Expect person	PASS
3	3	Patient with ethnicity other than hispanic has ETHNICITY_CONCEPT_ID = 0, id is PERSON_SOURCE_VALUE	Expect person	PASS
4	6	ICD9 code in SNOMED column, CONDITION_CONCEPT_ID = 0	Expect condition_occurrence	FAIL
5	8	Test that observation period is taking the earliest start and latest stop, id is person_source_value	Expect observation_period	FAIL
6	11	Collapse IP claim lines with <= 1 day between them, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
7	14	Collapse OP claims that occur within an IP visit, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
8	14	Collapse OP claims that occur within an IP visit, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
9	17	ER visit occurs on the first day of the IP visit, two visits created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
10	17	ER visit occurs on the first day of the IP visit, two visits created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
11	20	OP visit starts before IP visit but ends during IP, two visits created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
12	20	OP visit starts before IP visit but ends during IP, two visits created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	PASS
13	23	Two ER visits start on the same day, one visit created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	FAIL
14	23	Two ER visits start on the same day, one visit created, id is PERSON_SOURCE_VALUE	Expect visit_occurrence	FAIL

Issues in our synthetic data?

Achilles

- Did Achilles notice anything?

cdm_synthea



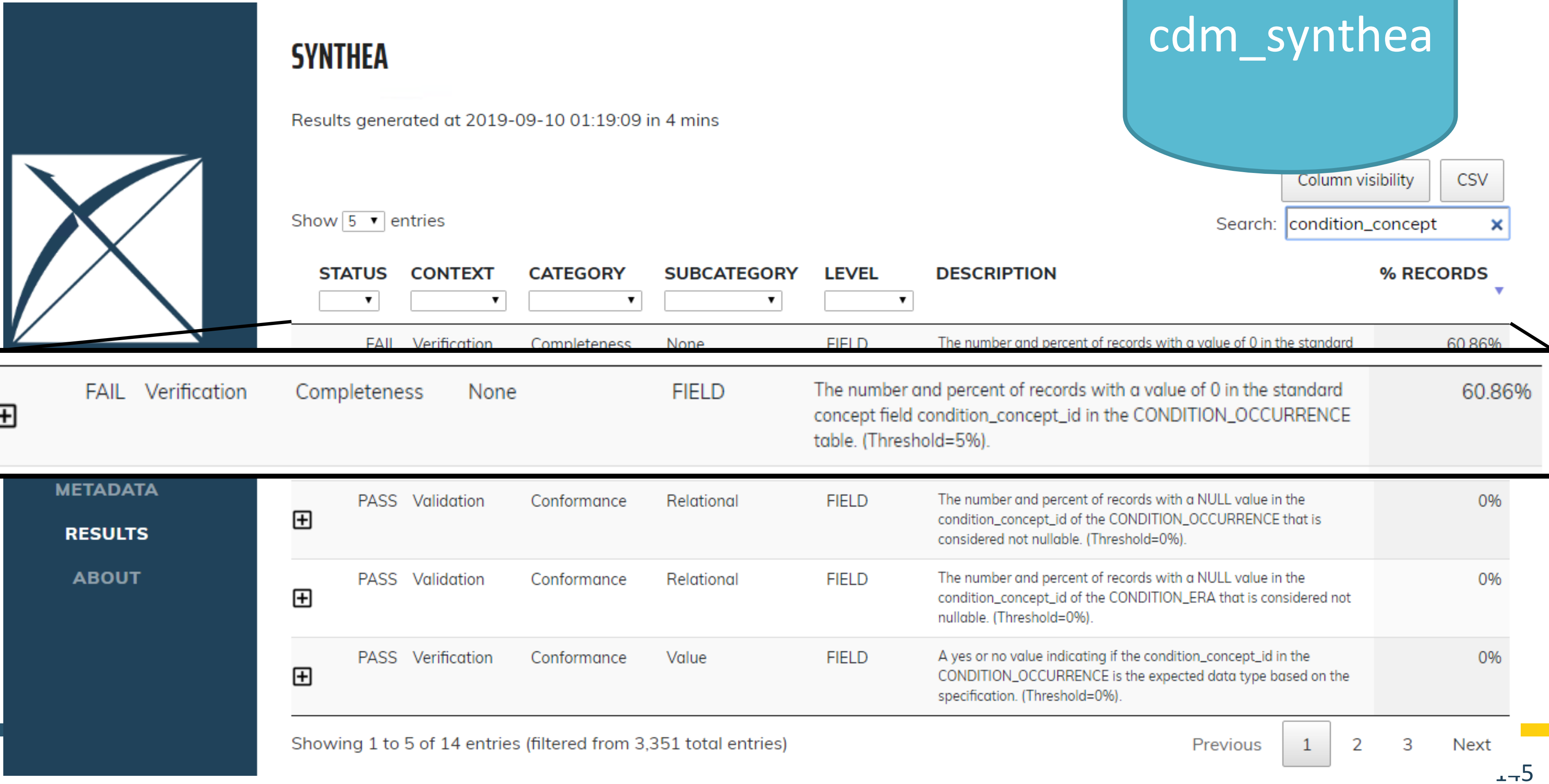
Message Type	Message
ERROR	410-Number of condition occurrence records outside valid observation peri > 0
ERROR	610-Number of procedure occurrence records outside valid observation period, be > 0
ERROR	710-Number of drug exposure records outside valid observation period; count (n=241) should not be > 0
ERROR	712-Number of drug exposure records with invalid provider_id; count (n=29,518) should not be > 0
ERROR	810-Number of observation records outside valid observation period; count (n=134) should not be > 0
ERROR	812-Number of observation records with invalid provider_id; count (n=8,518) should not be > 0
ERROR	909-Number of drug eras outside valid observation period; count (n=55) should not be > 0
ERROR	1,009-Number of condition eras outside valid observation period; count (n=134) should not be > 0
NOTIFICATION	[GeneralPopulationOnly] Not all deciles represented at first observation
NOTIFICA	
NOTIFICA	
NOTIFICATION	Unmapped data over percentage threshold in:Observation
NOTIFICATION	99+ percent of persons have exactly one observation period
NOTIFICATION	percentage of non-numerical measurement records exceeds general population threshold
NOTIFICATION	Unmapped data over percentage threshold in:Condition

Unmapped data over percentage threshold in:Condition

Issues in our synthetic data?

- Did DQD notice anything?

cdm_synthea



SYNTHEA

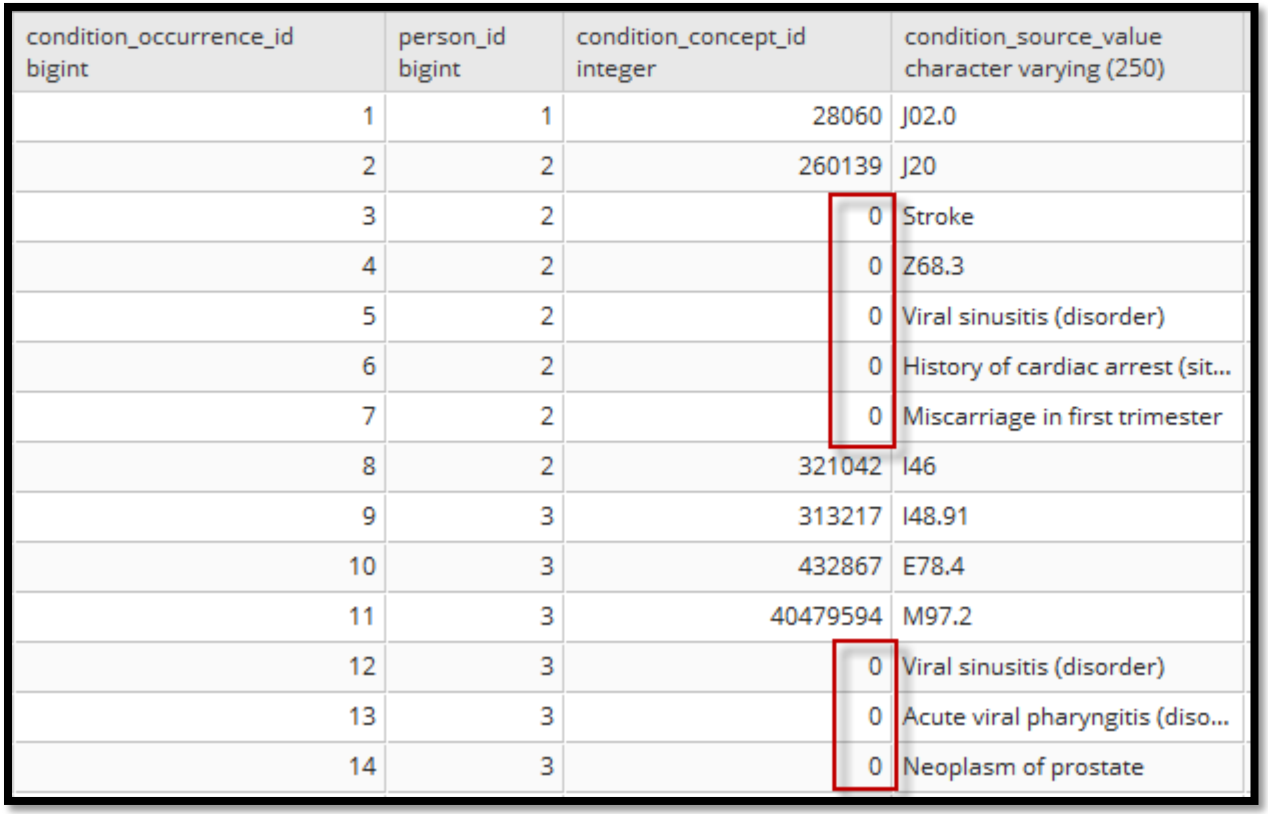
Results generated at 2019-09-10 01:19:09 in 4 mins

	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%).	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)

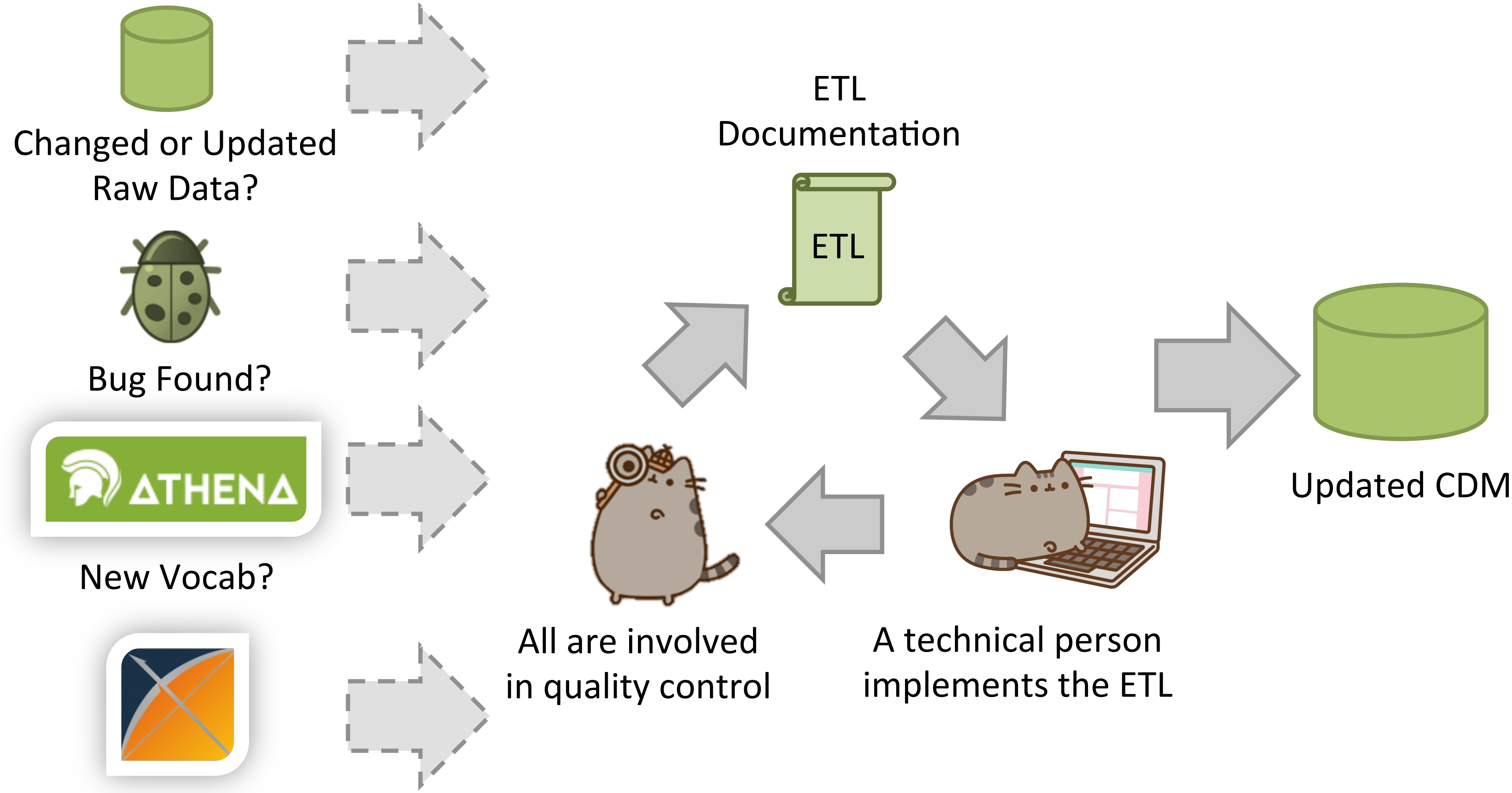
Maybe we have a bug?

- 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

ETL Maintenance



Changed or Updated Raw Data?

Bug Found?

New Vocab?

CDM Update?

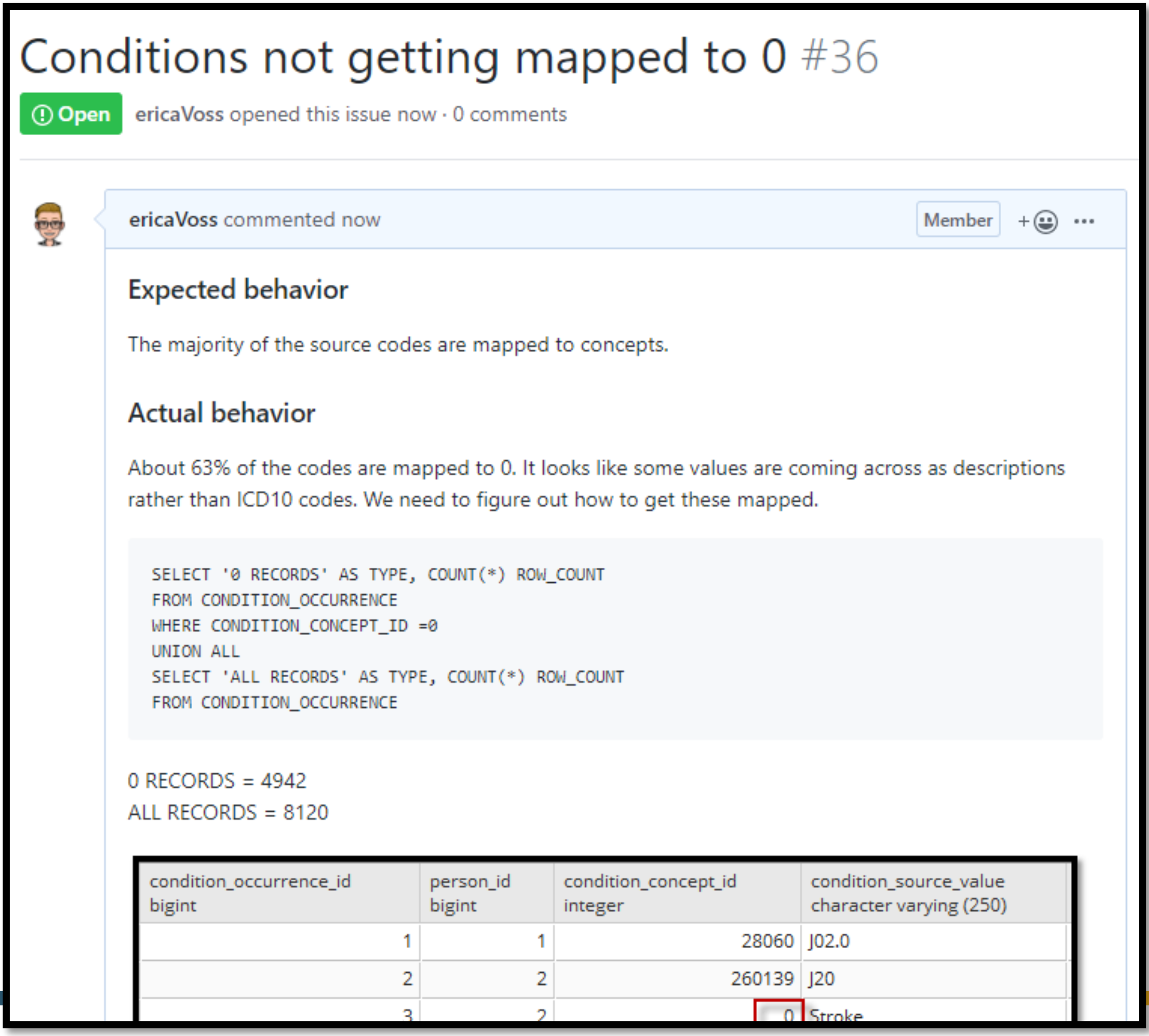
ETL Documentation

All are involved in quality control

A technical person implements the ETL

Updated CDM

Document the Bug



Conditions not getting mapped to 0 #36

ⓘ Open ericaVoss opened this issue now · 0 comments

ericaVoss commented now Member +☺ ···

Expected behavior

The majority of the source codes are mapped to concepts.

Actual behavior

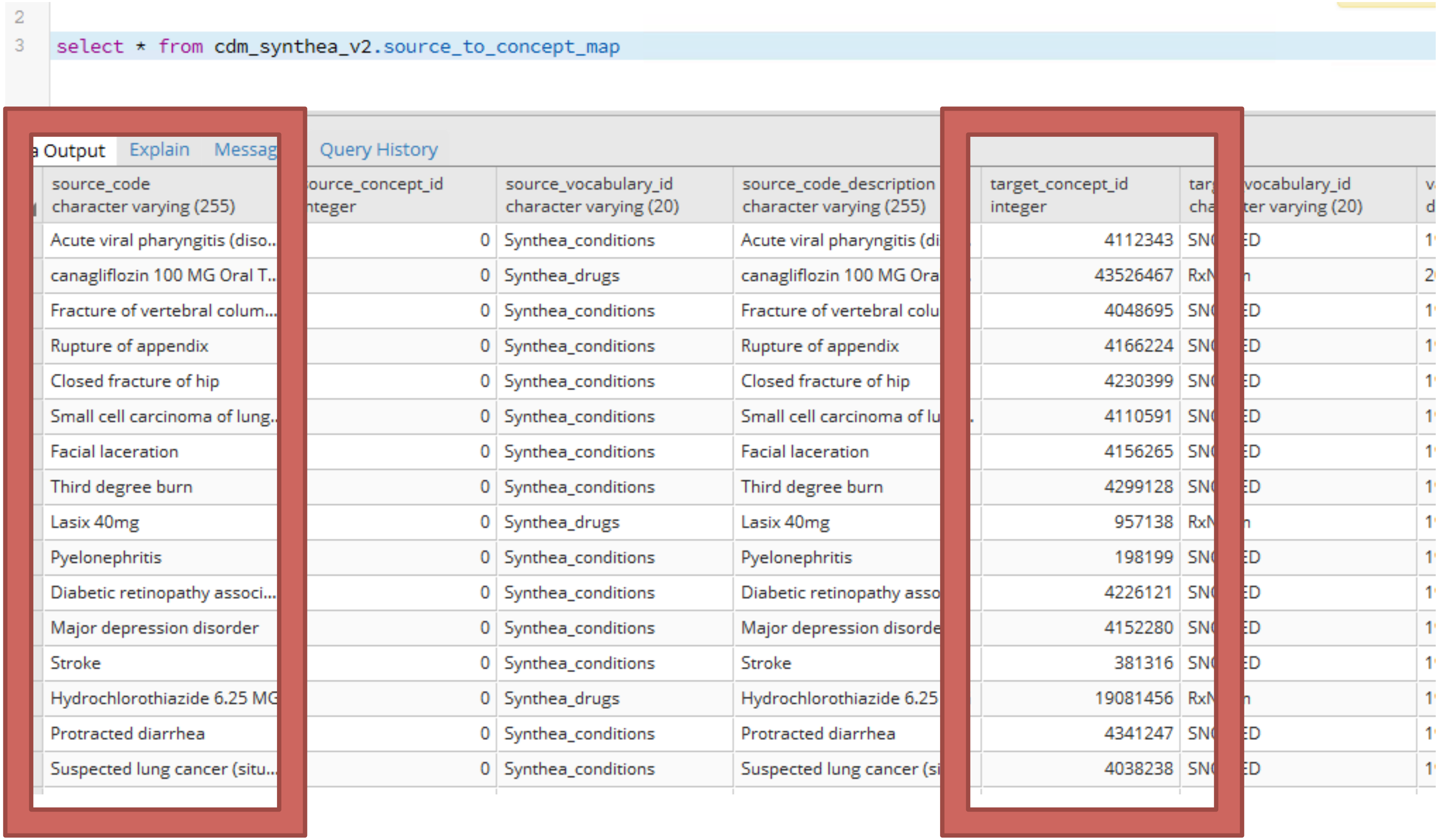
About 63% of the codes are mapped to 0. It looks like some values are coming across as descriptions rather than ICD10 codes. We need to figure out how to get these mapped.

```
SELECT '0 RECORDS' AS TYPE, COUNT(*) ROW_COUNT
FROM CONDITION_OCCURRENCE
WHERE CONDITION_CONCEPT_ID =0
UNION ALL
SELECT 'ALL RECORDS' AS TYPE, COUNT(*) ROW_COUNT
FROM CONDITION_OCCURRENCE
```

0 RECORDS = 4942
ALL RECORDS = 8120

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

Vocabulary to fix the problem



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

source_code character varying (255)	ource_concept_id nteger	source_vocabulary_id character varying (20)	source_code_description character varying (255)	target_concept_id integer	tar vocabulary_id cha ter varying (20)	v d
Acute viral pharyngitis (diso...	0	Synthea_conditions	Acute viral pharyngitis (di	4112343	SN ED	1
canagliflozin 100 MG Oral T...	0	Synthea_drugs	canagliflozin 100 MG Ora	43526467	RxN	2
Fracture of vertebral colum...	0	Synthea_conditions	Fracture of vertebral colu	4048695	SN ED	1
Rupture of appendix	0	Synthea_conditions	Rupture of appendix	4166224	SN ED	1
Closed fracture of hip	0	Synthea_conditions	Closed fracture of hip	4230399	SN ED	1
Small cell carcinoma of lung...	0	Synthea_conditions	Small cell carcinoma of lu	4110591	SN ED	1
Facial laceration	0	Synthea_conditions	Facial laceration	4156265	SN ED	1
Third degree burn	0	Synthea_conditions	Third degree burn	4299128	SN ED	1
Lasix 40mg	0	Synthea_drugs	Lasix 40mg	957138	RxN	1
Pyelonephritis	0	Synthea_conditions	Pyelonephritis	198199	SN ED	1
Diabetic retinopathy associ...	0	Synthea_conditions	Diabetic retinopathy asso	4226121	SN ED	1
Major depression disorder	0	Synthea_conditions	Major depression disorde	4152280	SN ED	1
Stroke	0	Synthea_conditions	Stroke	381316	SN ED	1
Hydrochlorothiazide 6.25 MG	0	Synthea_drugs	Hydrochlorothiazide 6.25	19081456	RxN	1
Protracted diarrhea	0	Synthea_conditions	Protracted diarrhea	4341247	SN ED	1
Suspected lung cancer (situ...	0	Synthea_conditions	Suspected lung cancer (si	4038238	SN ED	1

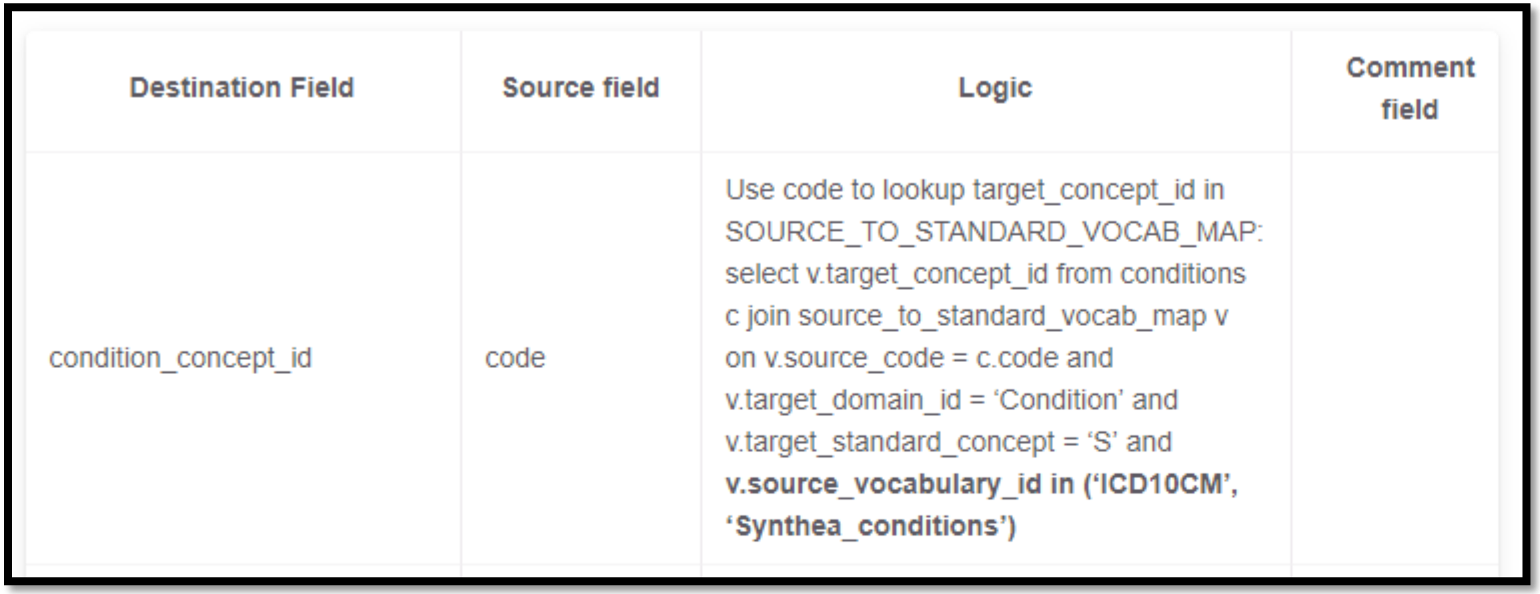
Vocabulary to fix the problem

```
WITH CTE_VOCAB_MAP AS (
      SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS SOURCE_CODE_DESCRIPTION,
c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,
c.VALID_START_DATE AS SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS
SOURCE_INVALID_REASON, c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS
TARGET_VOCABUALRY_ID,
c1.domain_id AS TARGET_DOMAIN_ID, c1.con                                                  D_REASON AS TARGET_INVALID_REASON,
c1.standard_concept AS TARGET_STANDARD_C
       FROM CONCEPT C
             JOIN CONCEPT_RELATIONSHIP C
                         ON C.CONCEPT_ID =
                         AND CR.invalid_r
                         AND cr.relations
              JOIN CONCEPT C1
                         ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID
                         AND C1.INVALID_REASON IS NULL
       UNION

SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id AS SOURCE_DOMAIN_ID,
c2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID,c1.VALID_START_DATE AS SOURCE_VALID_START_DATE,
c1.VALID_END_DATE AS SOURCE_VALID_END_DATE, stcm.INVALID_REASON AS SOURCE_INVALID_REASON,target_concept_id,
c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME, target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID,
c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c2.INVALID_REASON AS TARGET_INVALID_REASON,
c2.standard_concept AS TARGET_STANDARD_CONCEPT
       FROM source_to_concept_map stcm
              LEFT OUTER JOIN CONCEPT c1
                     ON c1.concept_id = stcm.source_concept_id
              LEFT OUTER JOIN CONCEPT c2
                     ON c2.CONCEPT_ID = stcm.target_concept_id
       WHERE stcm.INVALID_REASON IS NULL
)
SELECT TARGET_CONCEPT_ID, TARGET_CONCEPT_NAME, TARGET_DOMAIN_ID
FROM CTE_VOCAB_MAP
WHERE SOURCE_VOCABULARY_ID = ‘Synthea_conditions’
```

Look in the Source to Concept Map table for a map

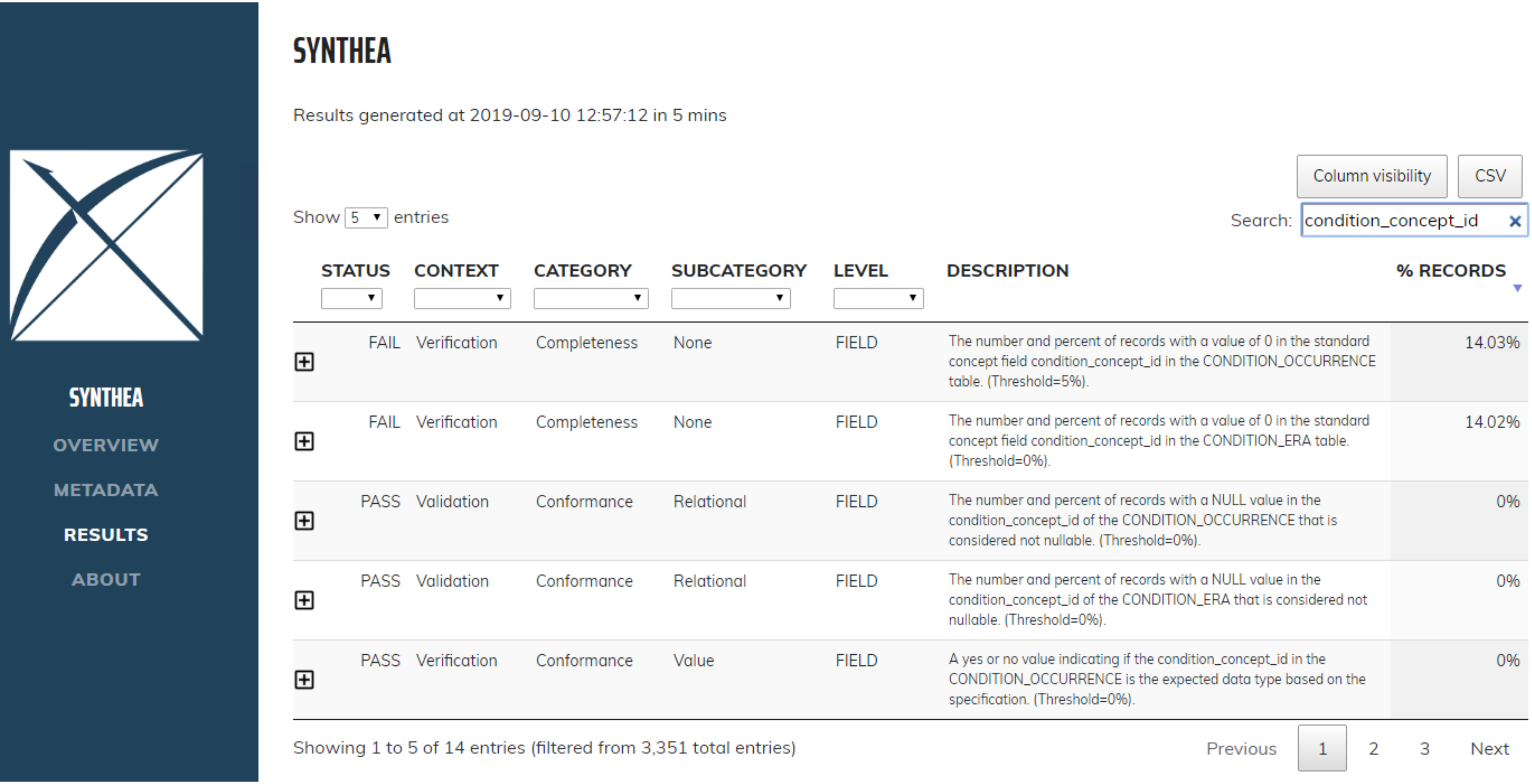
Update the ETL document

- https://ohdsi.github.io/Tutorial-ETL/docs/cdm_synthea_v2



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 (‘ICD10CM’, ‘Synthea_conditions’)**	

Re-run the DQD



SYNTHEA

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

Show 5 entries

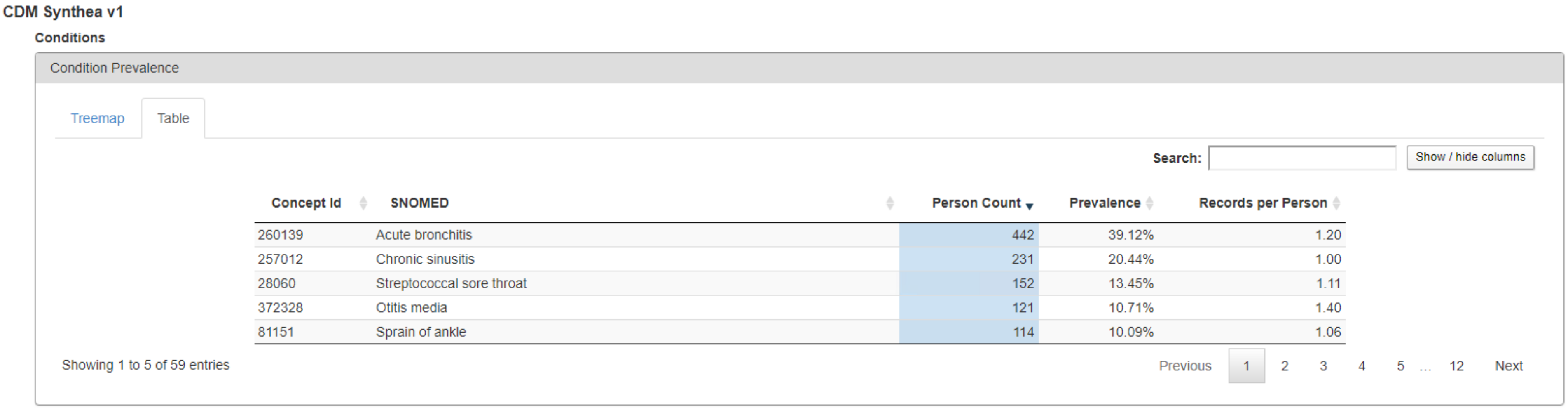
Search: condition_concept_id

	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)

Previous 1 2 3 Next

Re-run Achilles



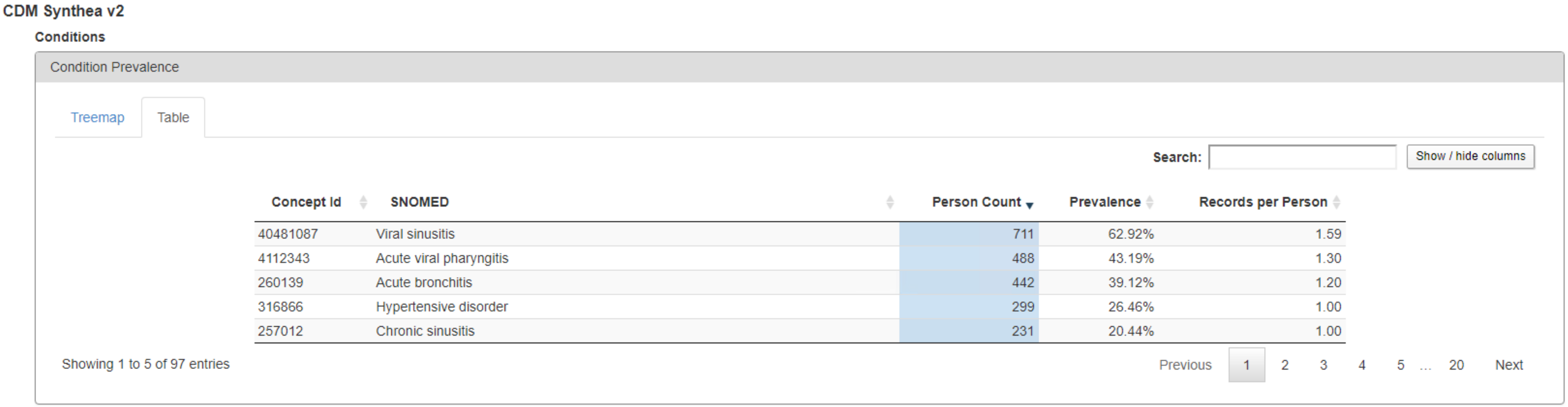
CDM Synthea v1

Conditions

Condition Prevalence

Concept Id	SNOMED	Person Count	Prevalence	Records per Person
260139	Acute bronchitis	442	39.12%	1.20
257012	Chronic sinusitis	231	20.44%	1.00
28060	Streptococcal sore throat	152	13.45%	1.11
372328	Otitis media	121	10.71%	1.40
81151	Sprain of ankle	114	10.09%	1.06

Showing 1 to 5 of 59 entries



CDM Synthea v2

Conditions

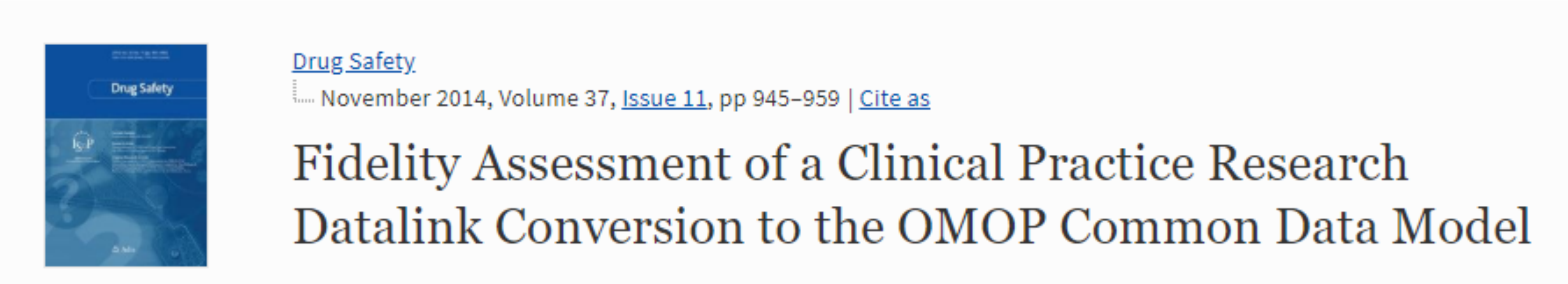
Condition Prevalence

Concept Id	SNOMED	Person Count	Prevalence	Records per Person
40481087	Viral sinusitis	711	62.92%	1.59
4112343	Acute viral pharyngitis	488	43.19%	1.30
260139	Acute bronchitis	442	39.12%	1.20
316866	Hypertensive disorder	299	26.46%	1.00
257012	Chronic sinusitis	231	20.44%	1.00

Showing 1 to 5 of 97 entries

Final Hard Lessons Learned

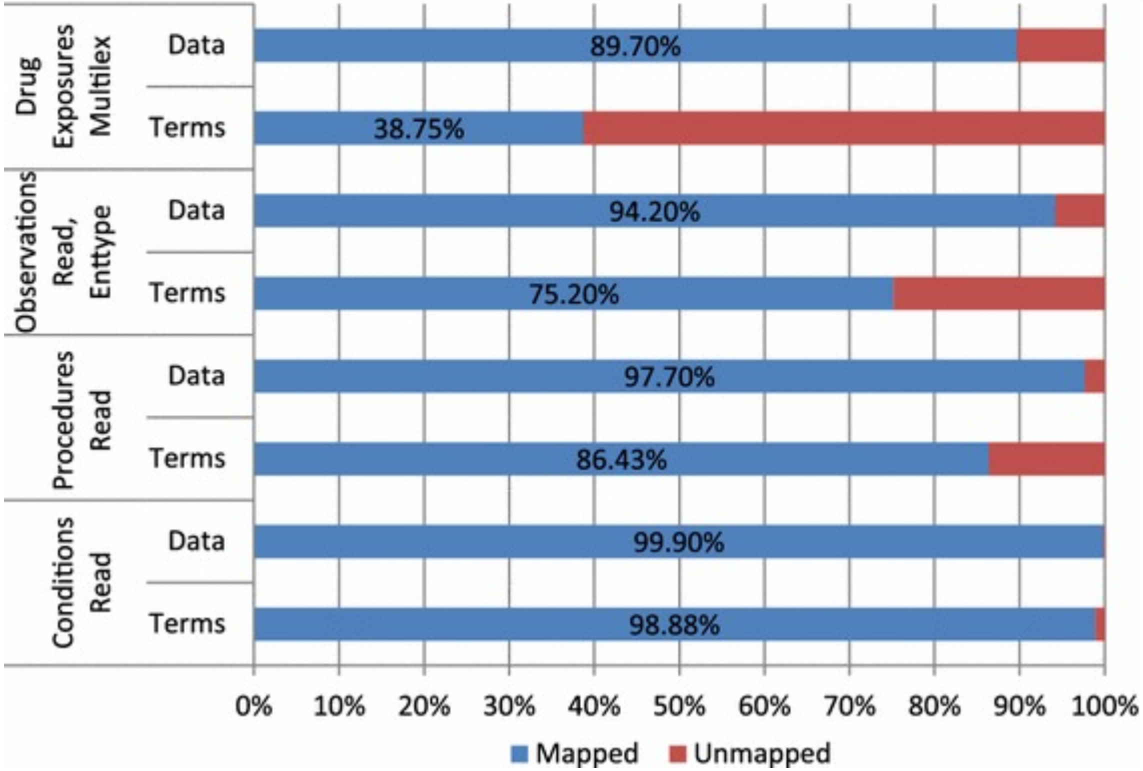
80/20 Rule



Drug Safety

November 2014, Volume 37, Issue 11, pp 945–959 | Cite as

Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model



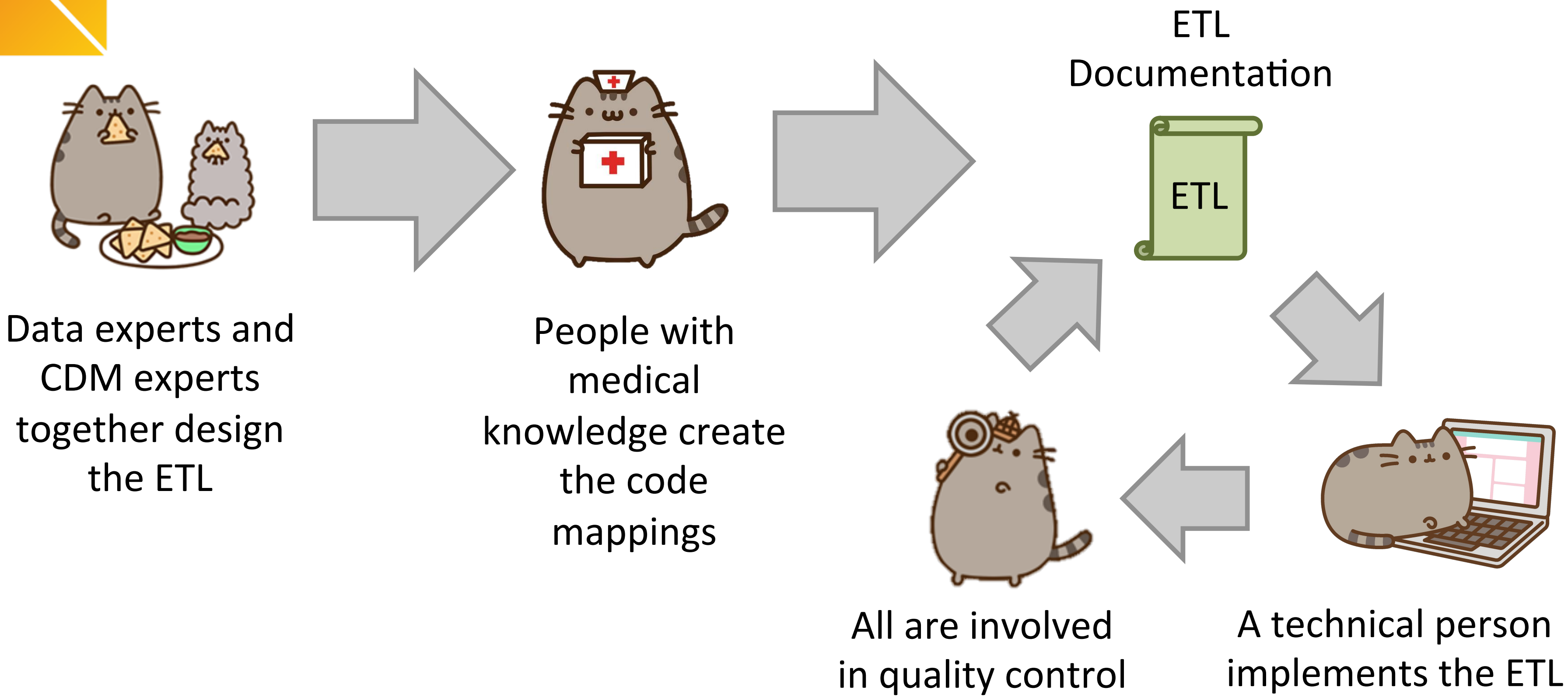
You don't need to map all terms to get good data coverage!

Comfort with Data Loss

- If there is data that is not of research quality or there are methods to adjust, use the ETL to standardize that

Example Patient Drop Counts from a CDM Build	
Reason to Drop Someone	**Person Count**
Unknown gender	23,592
Implausible year of birth - past	749
Implausible year of birth - post earliest observation period	3,836
Gender changes	2

ETL Process



Data experts and CDM experts together design the ETL

People with medical knowledge create the code mappings

ETL Documentation

ETL

A technical person implements the ETL

All are involved in quality control

OHDSI Tools

White Rabbit

Rabbit In a Hat

Usagi

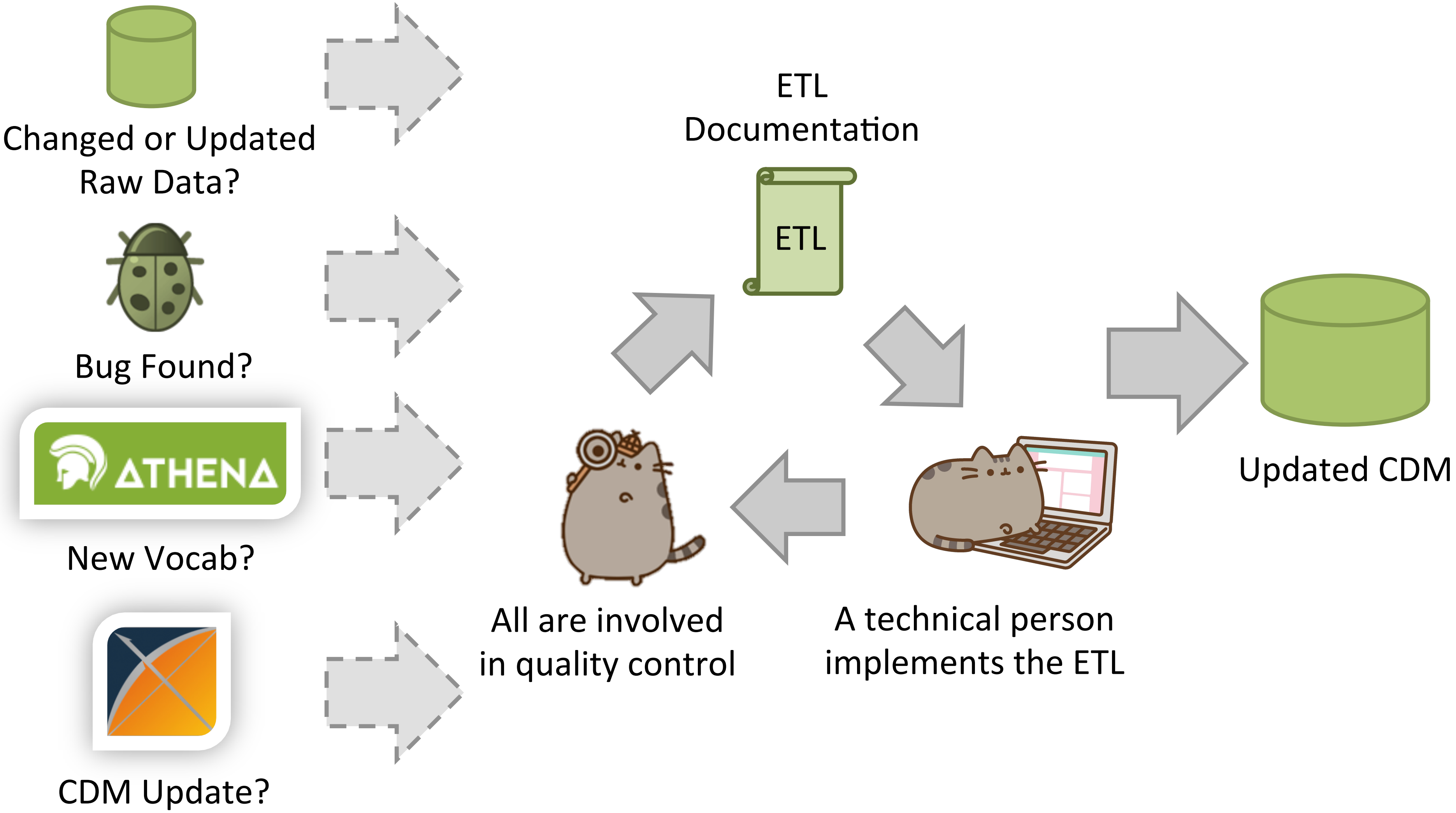
White Rabbit

ACHILLES

DQD

Rabbit In a Hat

ETL Maintenance



Changed or Updated Raw Data?

Bug Found?

New Vocab?

CDM Update?

ETL Documentation

All are involved in quality control

A technical person implements the ETL

Updated CDM

Thank you!

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

This tutorial would not have been possible without the contribution of many collaborators in the OHDSI Community

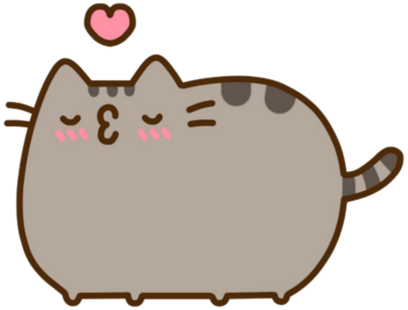
aws

We like to thank Amazon Web Services for their valuable technical support and resources

Acknowledgements

Anthony Molinaro who wrote the Synthea CDM Builder

James Wiggins who helps us prepare an AWS instance for use today



Pusheen the Cat

http://pusheen.com/

SEPTEMBER 16TH, 2019

2019 OHDSI SYMPOSIUM

TUTORIALS:
SEPTEMBER
15TH & 17TH

Bethesda North Marriott
& Conference Center
5701 Marinelli Rd
North Bethesda, MD 20852

WWW.OHDSI.ORG/EVENTS/2019-OHDSI-SYMPOSIUM