CDM Workshop, Part 1

Common Data Model Working Group

2022-03-08

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

The Journey from Data to …

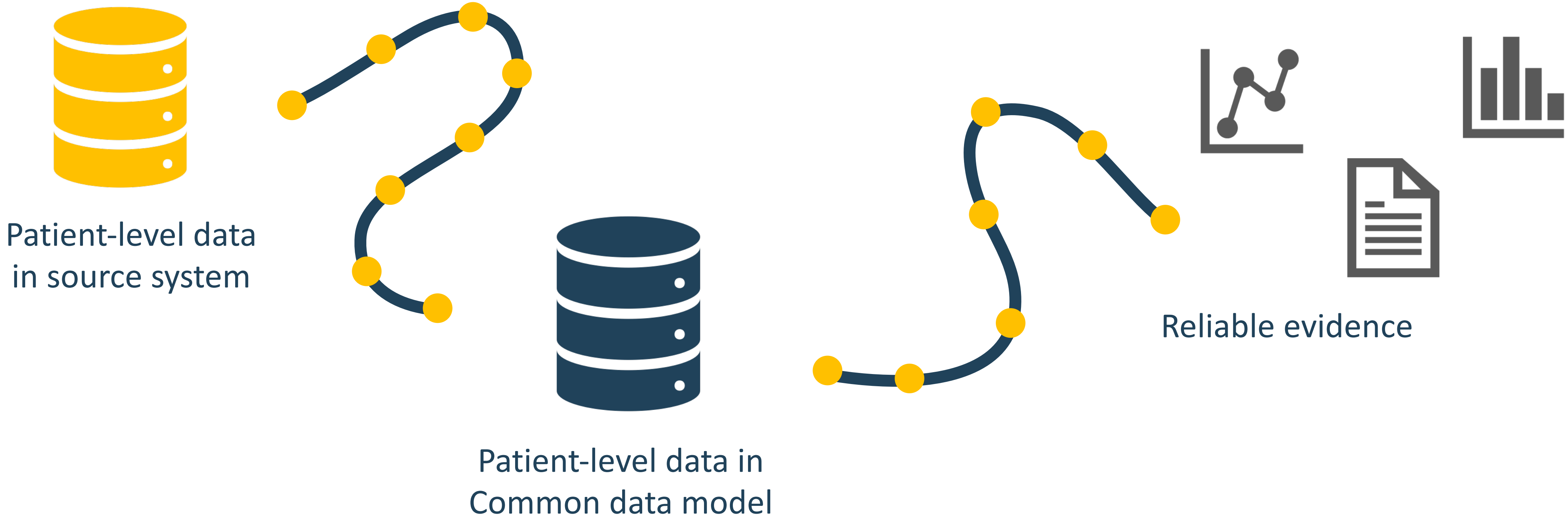
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

The Journey from Data to …

Data

The Journey from Data to Reliable Evidence

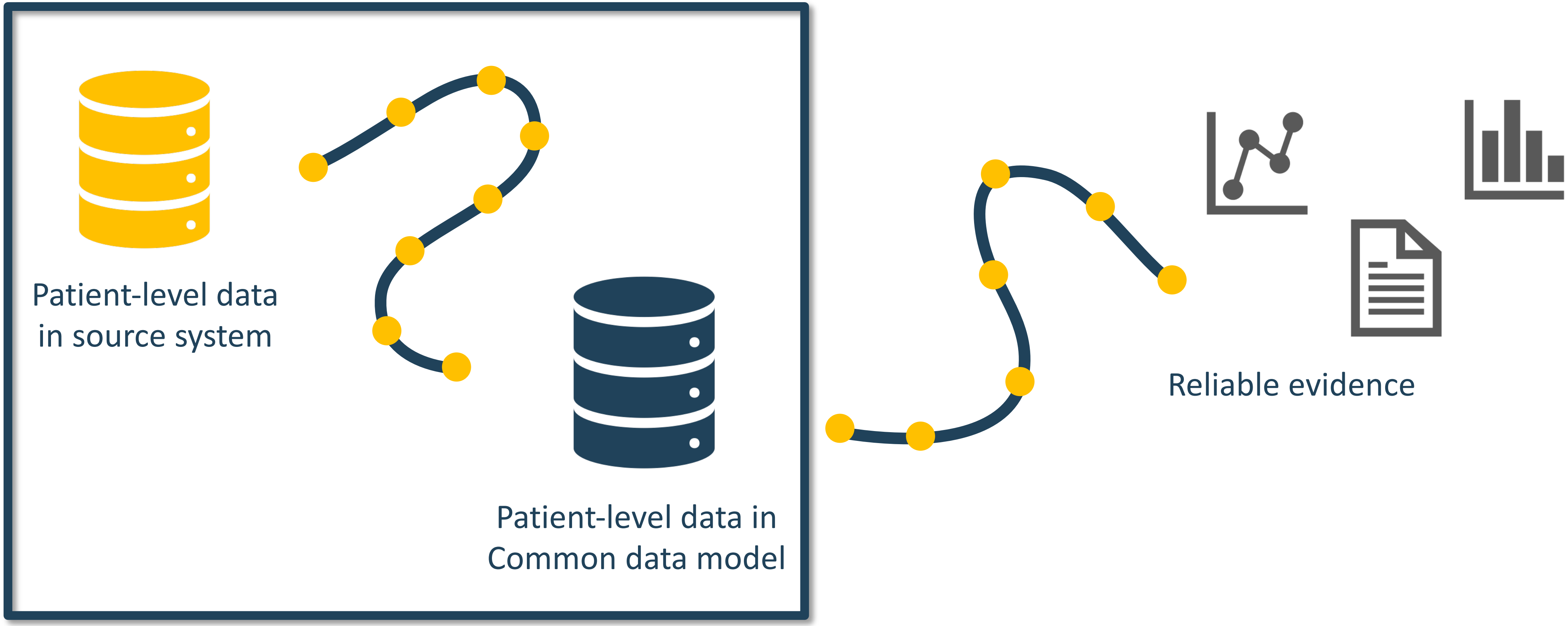


Patient-level data in source system

Patient-level data in Common data model

Reliable evidence

The Journey from Source to Standardized Data

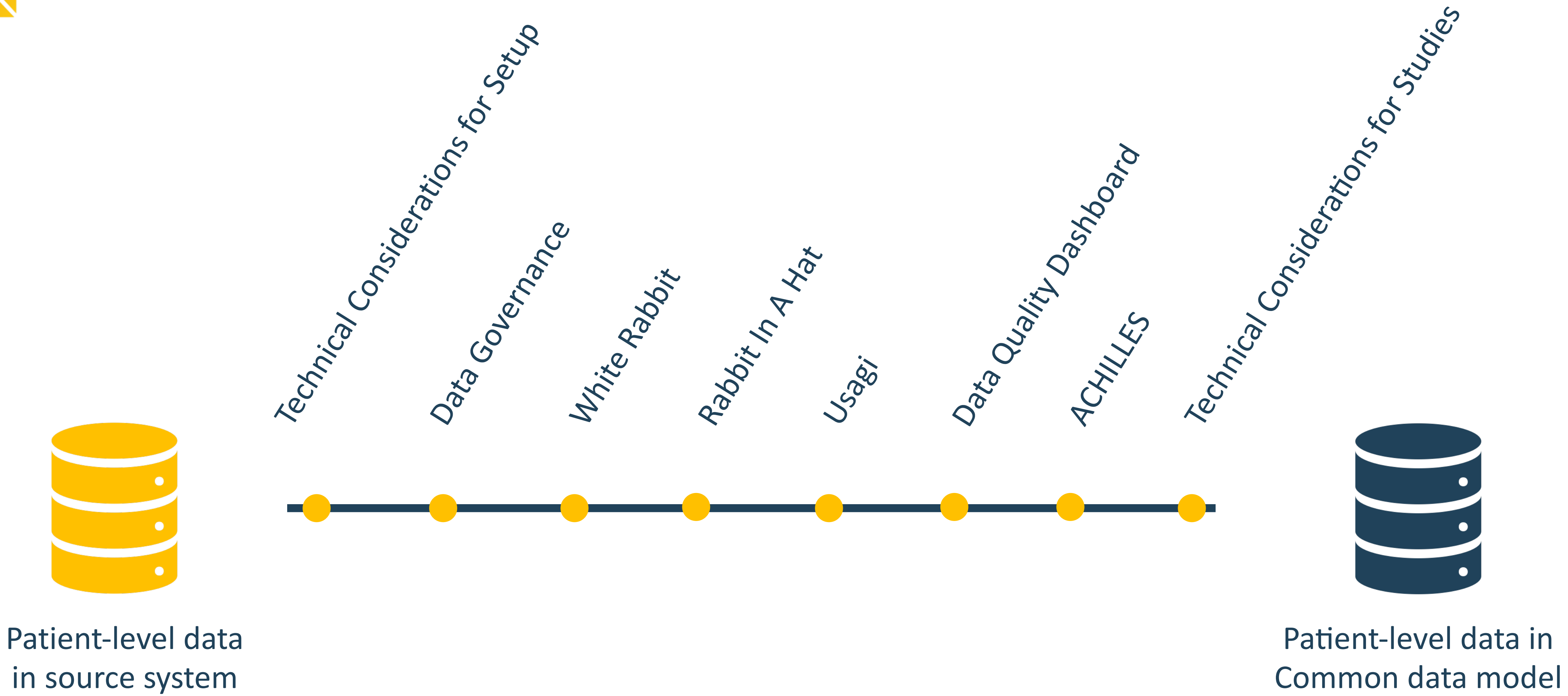


Patient-level data in source system

Patient-level data in Common data model

Reliable evidence

The Journey from Source to Standardized Data



Patient-level data in source system

Technical Considerations for Setup

Data Governance

White Rabbit

Rabbit In A Hat

Usagi

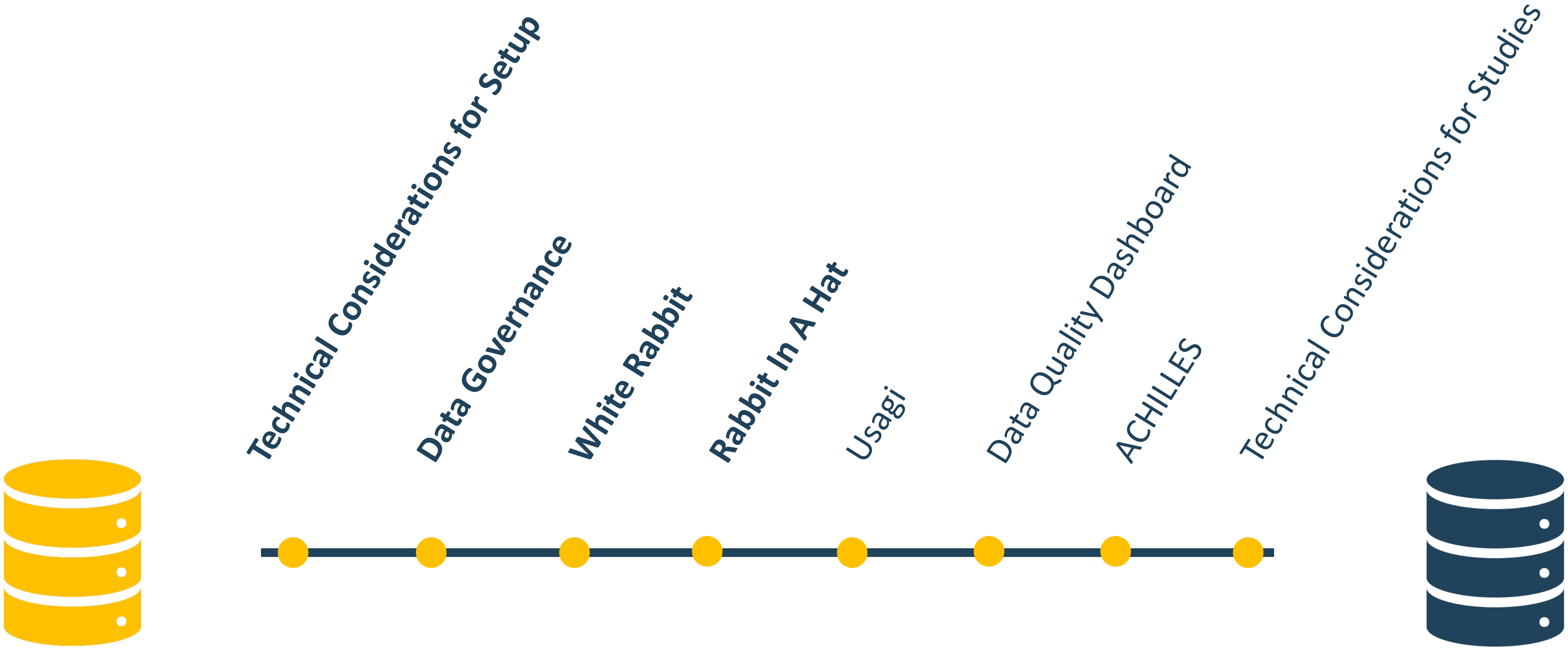
Data Quality Dashboard

ACHILLES

Technical Considerations for Studies

Patient-level data in Common data model

The Journey from Source to Standardized Data



Technical Considerations for Setup

Data Governance

White Rabbit

Rabbit In A Hat

Usagi

Data Quality Dashboard

ACHILLES

Technical Considerations for Studies

Patient-level data
in source system

Patient-level data in
Common data model

The Journey from Source to Standardized Data

Technical Considerations for Setup

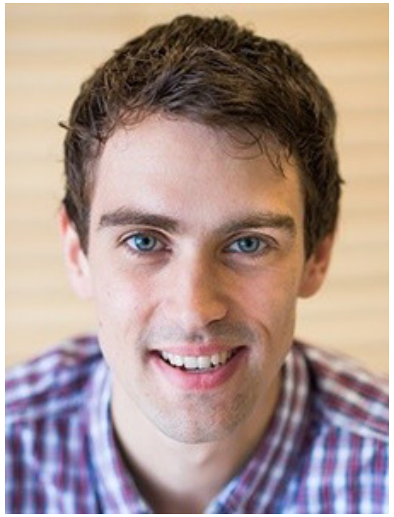
Frank DeFalco

Data Governance

Kristin Kostka

White Rabbit

Rabbit In A Hat



Maxim Moinat

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Technical Considerations for Setup

Frank DeFalco

CDM Workshop 2022

2022-03-08

Relational Database Requirements

- Start with a compatible relational database platform
  - Microsoft SQL Server
  - Oracle
  - PostgreSQL
  - Amazon RedShift
  - Impala
  - IBM Netezza
  - Google BigQuery
  - Microsoft PDW
  - Apache Spark
  - SQLite

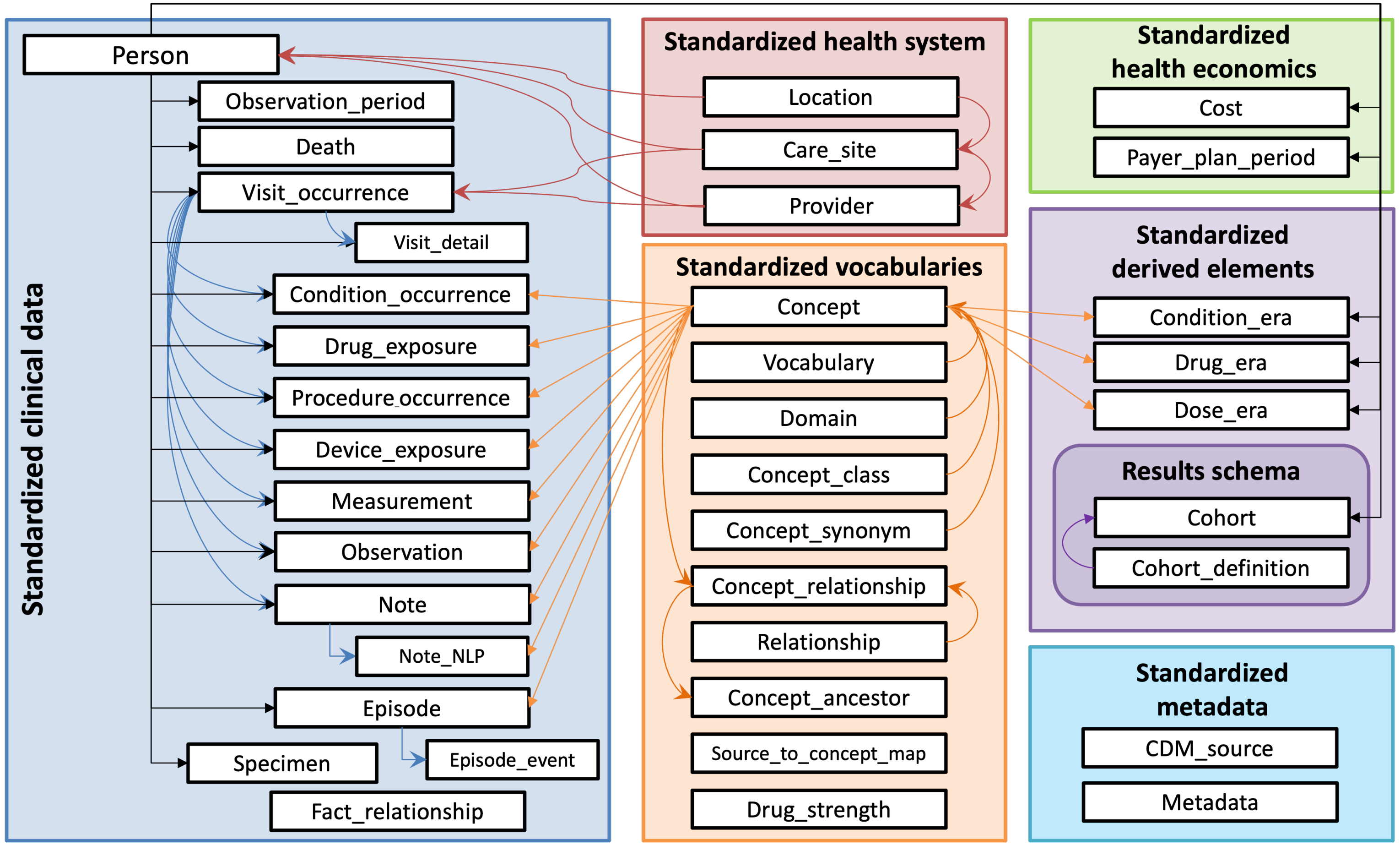
Setting up your CDM

- Choose a currently supported CDM Version (5.3, **5.4**)
  - CommonDataModel
    - R Package to generate a CDM on your database platform
    - https://github.com/OHDSI/CommonDataModel

```
CommonDataModel::executeDdl(
    connectionDetails = cd,
    cdmVersion = "5.4",
    cdmDatabaseSchema = "ohdsi_journey"
)
```

  - A vocabulary schema is required to exist and contain a vocabulary content distribution along with any instance of the CDM
  - Athena
    - https://athena.ohdsi.org/
    - Web site where you can download vocabulary content

CDM v5.4 Schema



Before you start loading your data, consider more than the technical...

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Governance Considerations with Your OMOP CDM

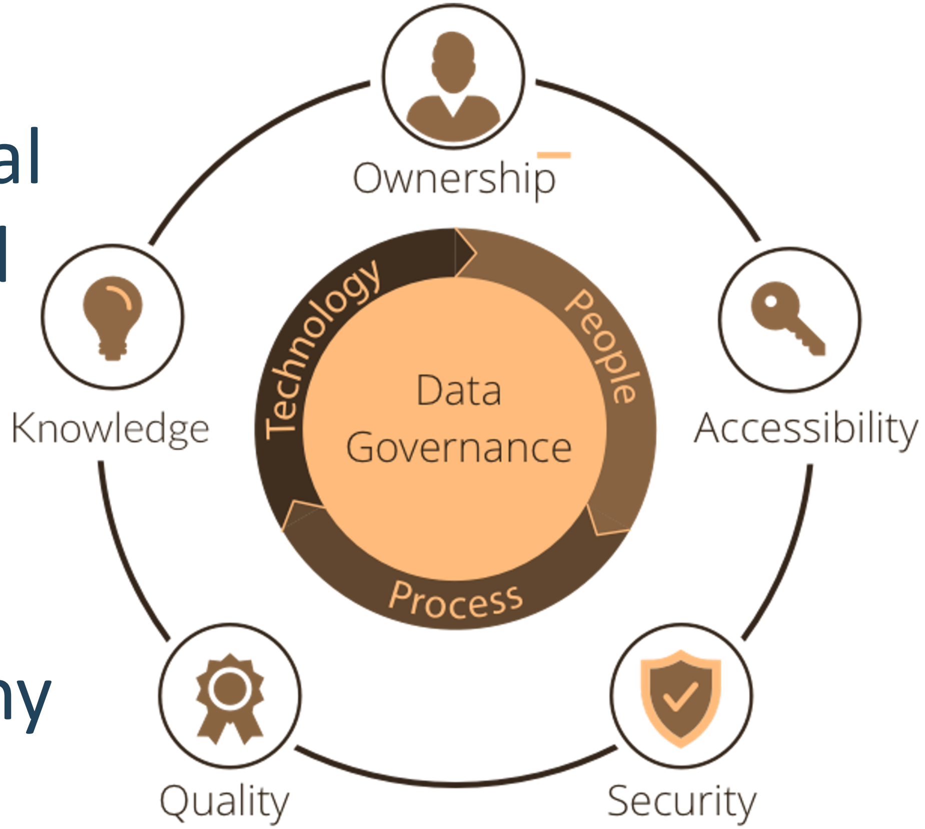
Kristin Kostka

CDM Workshop 2022

2022-03-08

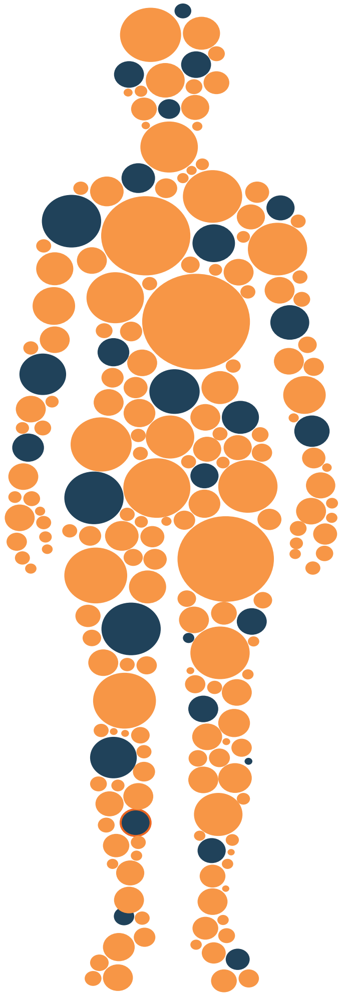
What is data governance?

- A process to formally outline how organizational data will be managed and controlled
- Ensures that data is consistent and trustworthy and is not misused



Why does it matter in the OMOP CDM?

- OMOP CDM is a person-centric model
- The model can retain attributes that may be considered personal identified information (PII) or protected health information (PHI)
- There are many ways a site may treat their OMOP CDM to uphold their governance protocols

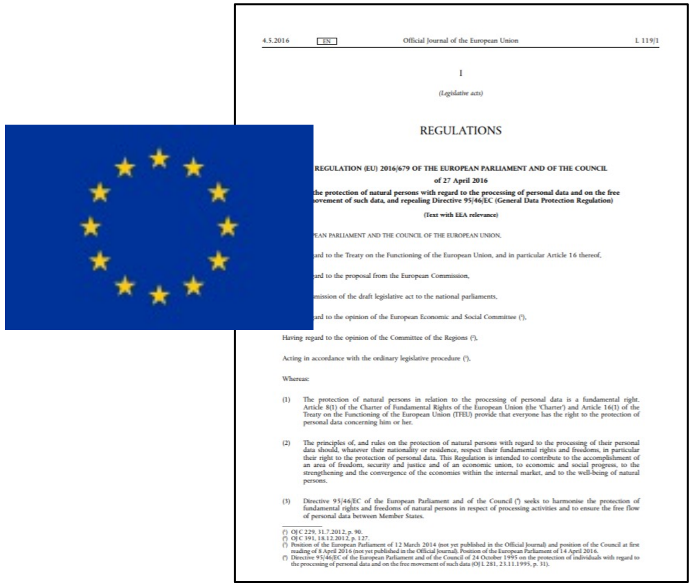


The OHDSI data holder’s responsibility

- In OHDSI research, it is the responsibility of each data holder to:
- **know,**
- **understand,**
- **and follow**
- local data governance processes related to use of your OMOP CDM instance.

Sites have different rules about patient data

- ***In Europe:***
- General Data Protection Regulation (GDPR) is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA)
  - *Implemented 25 May 2018*



4.5.2016 EN Official Journal of the European Union L 119/1

I

(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

he protection of natural persons with regard to the processing of personal data and on the free ovement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

EAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

ard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

ard to the proposal from the European Commission,

mission of the draft legislative act to the national parliaments,

ard to the opinion of the European Economic and Social Committee (1),

Having regard to the opinion of the Committee of the Regions (2),

Acting in accordance with the ordinary legislative procedure (3),

Whereas:

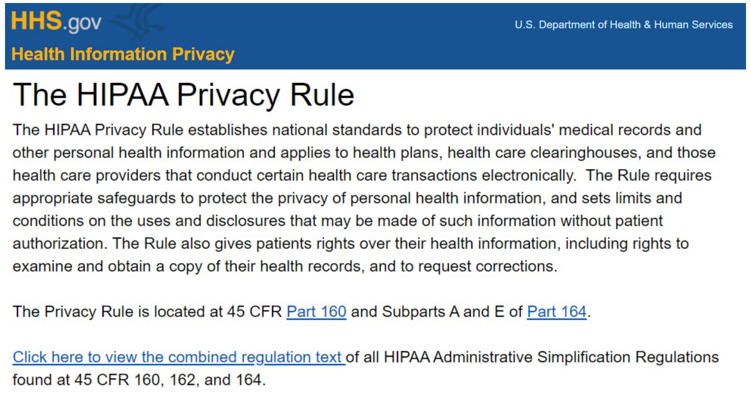
(1) The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

(2) The principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data. This Regulation is intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons.

(3) Directive 95/46/EC of the European Parliament and of the Council (4) seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data between Member States.

(1) OJ C 229, 31.7.2012, p. 90.
(2) OJ C 391, 18.12.2012, p. 127.
(3) Position of the European Parliament of 12 March 2014 (not yet published in the Official Journal) and position of the Council at first reading of 8 April 2016 (not yet published in the Official Journal). Position of the European Parliament of 14 April 2016.
(4) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ L 281, 23.11.1995, p. 31).

- ***In the US:***
- The Health Insurance Portability and Accountability Act (HIPAA) of 1996 establishes national standards to protect individuals' medical records and other personal health information



HHS.gov U.S. Department of Health & Human Services

Health Information Privacy

The HIPAA Privacy Rule

The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. The Rule requires appropriate safeguards to protect the privacy of personal health information, and sets limits and conditions on the uses and disclosures that may be made of such information without patient authorization. The Rule also gives patients rights over their health information, including rights to examine and obtain a copy of their health records, and to request corrections.

The Privacy Rule is located at 45 CFR Part 160 and Subparts A and E of Part 164.

Click here to view the combined regulation text of all HIPAA Administrative Simplification Regulations found at 45 CFR 160, 162, and 164.

Where do privacy concerns arise?

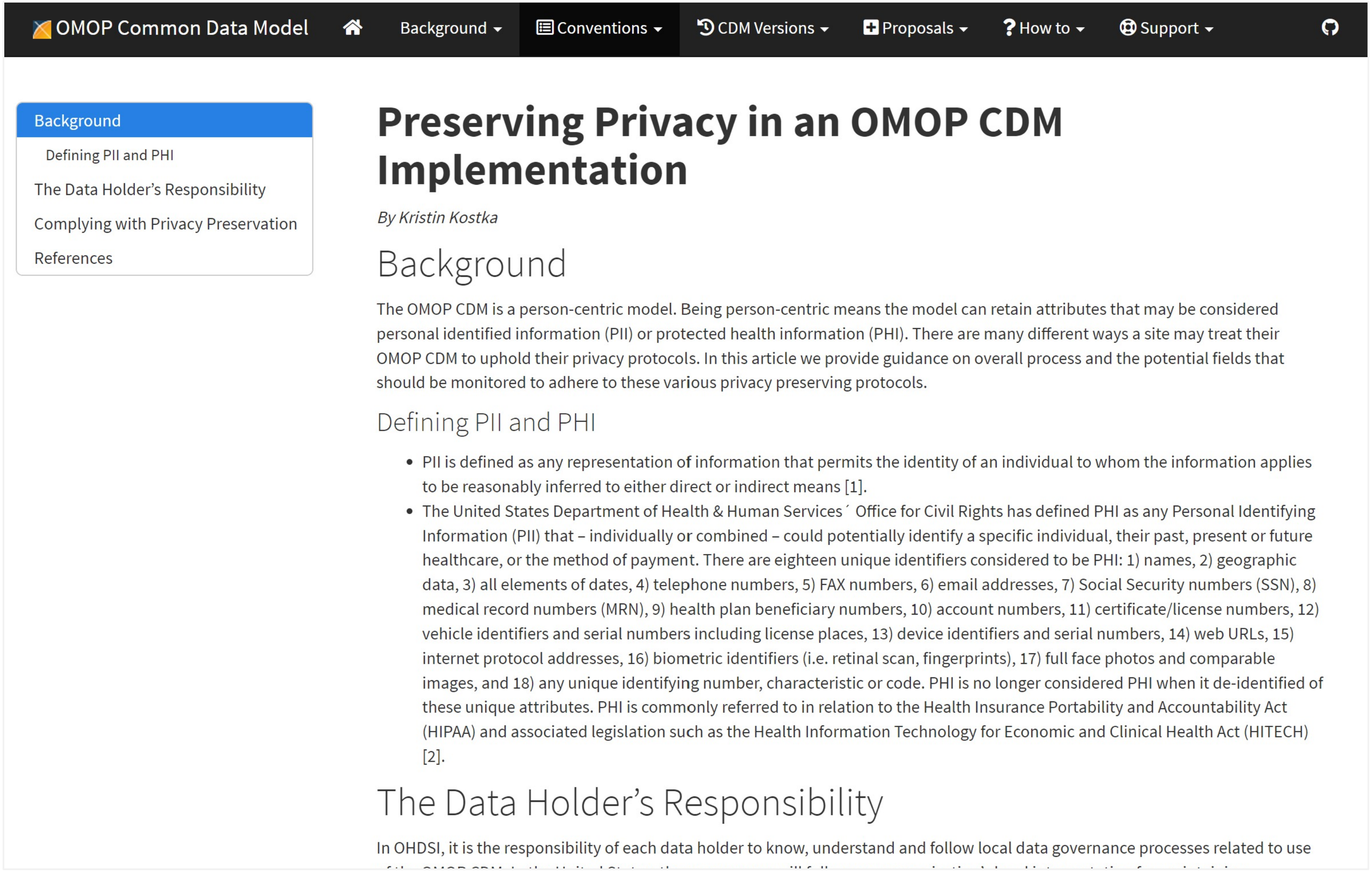
- Privacy issues usually happen around:
- Date of Birth
- Patient location
- Precise clinical event date
- Psychological or mentally related clinical conditions
- Death date
- Death cause (usually non-medical related death, murder, car accident etc.)

Where should I be checking my OMOP CDM?

- Date Fields Across Domains
- *_source_value
- Any string fields

- Extra sensitive tables: PERSON, LOCATION, PROVIDER, OBSERVATION, NOTE and NOTE_NLP

More detail on the OMOP CDM Wiki



Preserving Privacy in an OMOP CDM Implementation

By Kristin Kostka

Background

The OMOP CDM is a person-centric model. Being person-centric means the model can retain attributes that may be considered personal identified information (PII) or protected health information (PHI). There are many different ways a site may treat their OMOP CDM to uphold their privacy protocols. In this article we provide guidance on overall process and the potential fields that should be monitored to adhere to these various privacy preserving protocols.

Defining PII and PHI

- PII is defined as any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred to either direct or indirect means [1].
- The United States Department of Health & Human Services´ Office for Civil Rights has defined PHI as any Personal Identifying Information (PII) that – individually or combined – could potentially identify a specific individual, their past, present or future healthcare, or the method of payment. There are eighteen unique identifiers considered to be PHI: 1) names, 2) geographic data, 3) all elements of dates, 4) telephone numbers, 5) FAX numbers, 6) email addresses, 7) Social Security numbers (SSN), 8) medical record numbers (MRN), 9) health plan beneficiary numbers, 10) account numbers, 11) certificate/license numbers, 12) vehicle identifiers and serial numbers including license places, 13) device identifiers and serial numbers, 14) web URLs, 15) internet protocol addresses, 16) biometric identifiers (i.e. retinal scan, fingerprints), 17) full face photos and comparable images, and 18) any unique identifying number, characteristic or code. PHI is no longer considered PHI when it de-identified of these unique attributes. PHI is commonly referred to in relation to the Health Insurance Portability and Accountability Act (HIPAA) and associated legislation such as the Health Information Technology for Economic and Clinical Health Act (HITECH) [2].

The Data Holder’s Responsibility

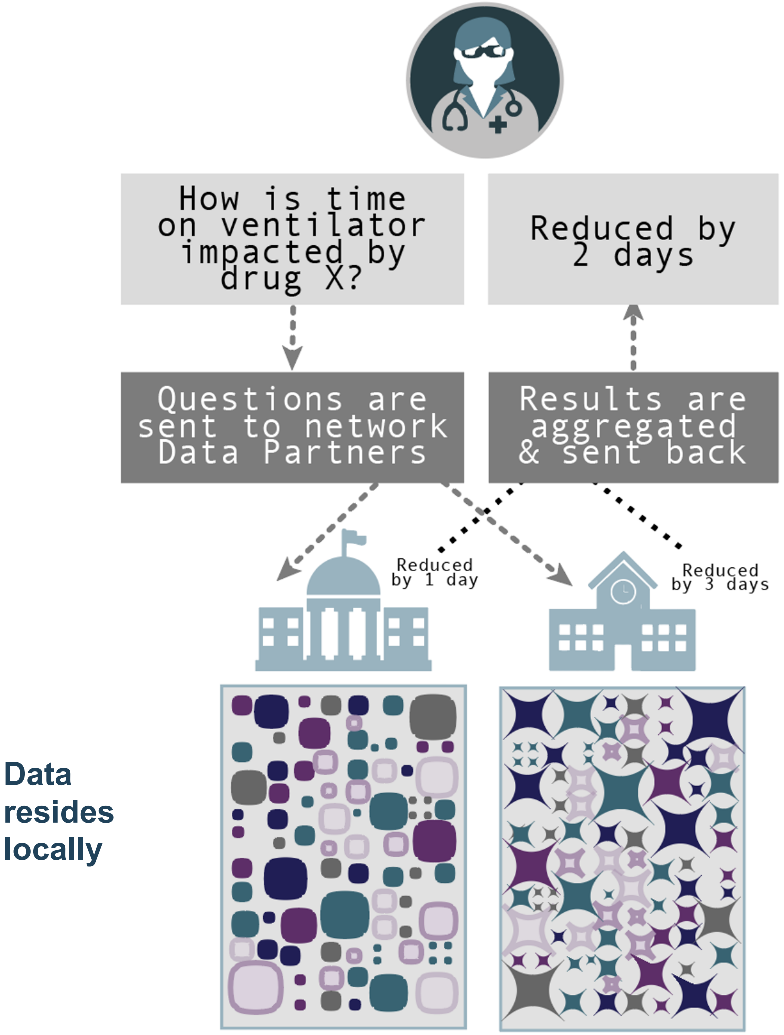
In OHDSI, it is the responsibility of each data holder to know, understand and follow local data governance processes related to use

https://ohdsi.github.io/CommonDataModel/cdmPrivacy.html

Alphabet soup: governance edition

- Institutional Review Board (IRB)
- Data use agreement (DUA)
- Data transfer agreements (DTA)
- Data use request (DUR)
- CITI training
- Data stewards
- Limited data set (LDS)
- Synthetic data set

Types of data querying



Federated querying

How is time on ventilator impacted by drug X?

Reduced by 2 days

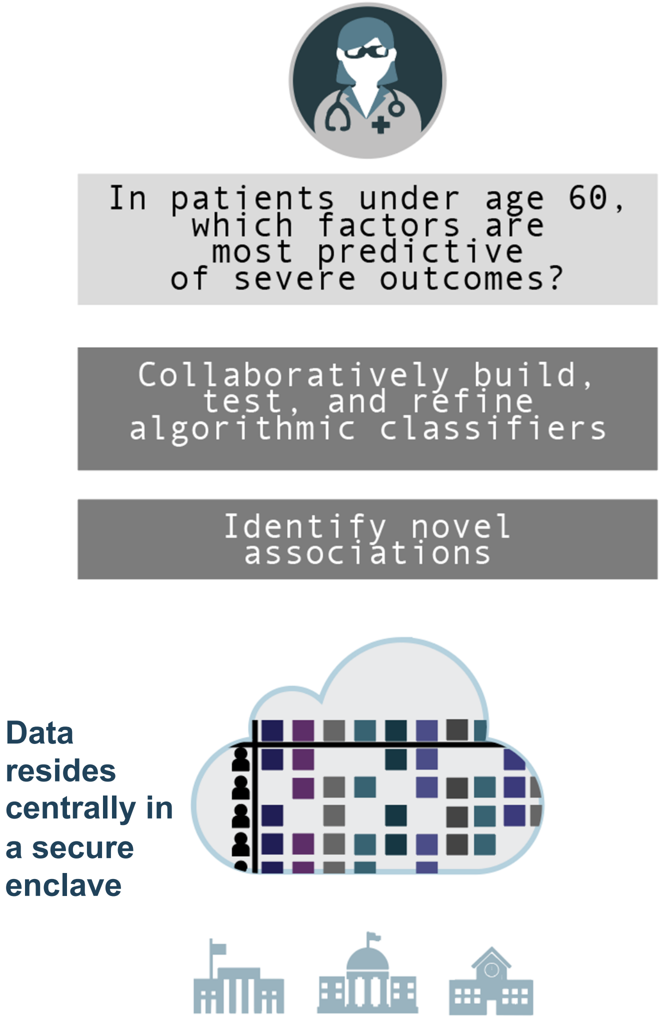
Questions are sent to network Data Partners

Results are aggregated & sent back

Reduced by 1 day

Reduced by 3 days

Data resides locally



Centralized analytics

In patients under age 60, which factors are most predictive of severe outcomes?

Collaboratively build, test, and refine algorithmic classifiers

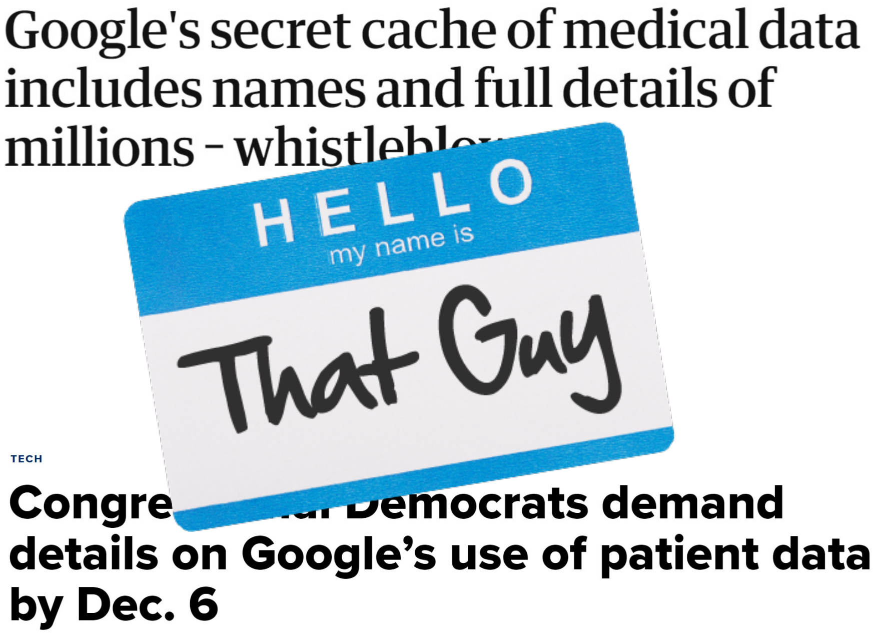
Identify novel associations

Data resides centrally in a secure enclave

Credit: N3C

Closing thoughts

- Governance is not one-size-fits-all
- Consult your local IRB and privacy office for guidance
- When in doubt, ask for help before you break any rules!



Google's secret cache of medical data includes names and full details of millions – whistlebl

TECH

Congre Democrats demand details on Google's use of patient data by Dec. 6

HELLO
my name is
That Guy

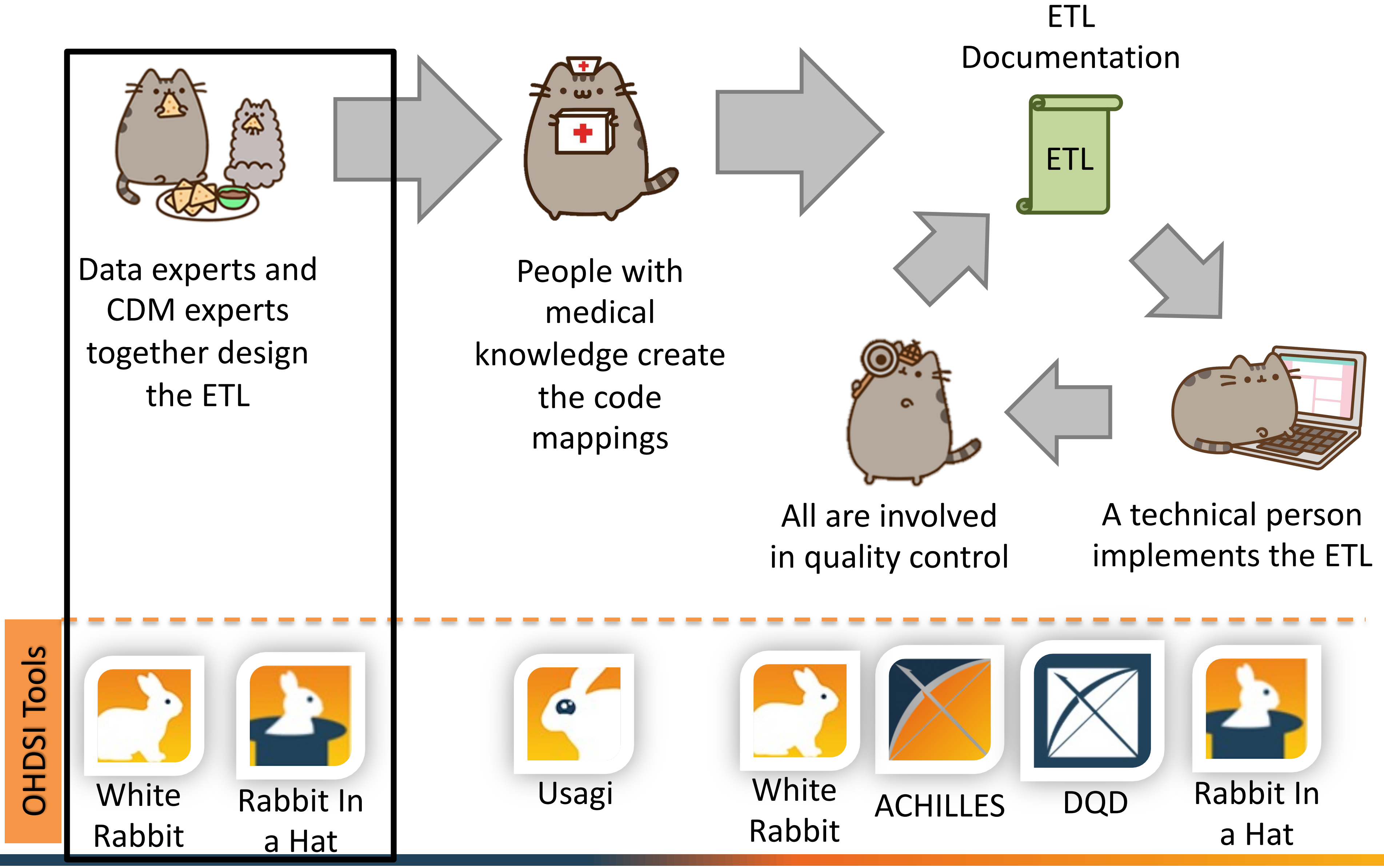
White Rabbit (WR)
Rabbit in a Hat (RiaH)

Maxim Moinat

CDM Workshop 2022

2022-03-08

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

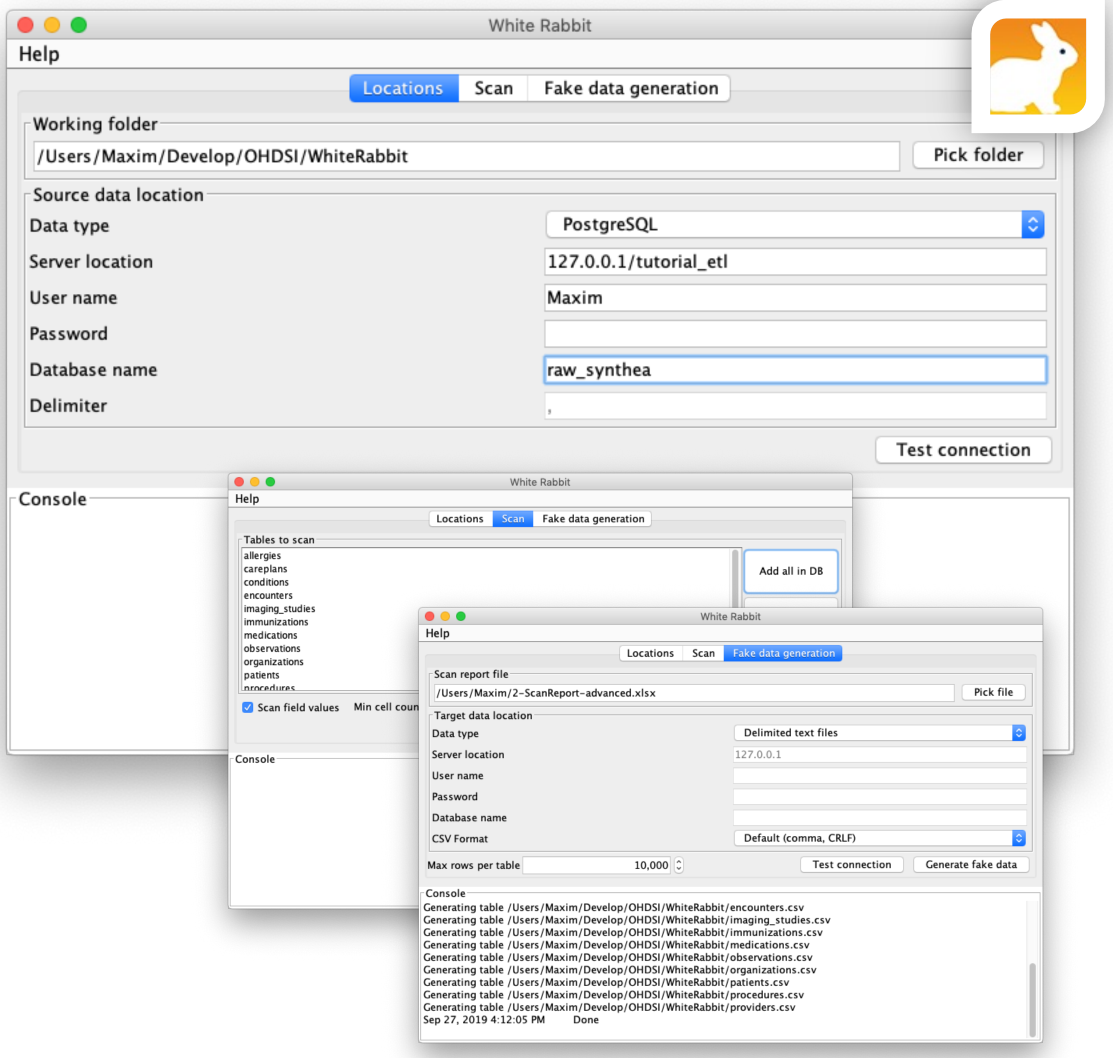
ACHILLES

DQD

Rabbit In a Hat

WR - Overview

- Create Scan Report
  - Overview of tables, fields and values
  - Record counts and statistics
- Support for:
  - Delimited text (csv)
  - Sas7bdat files
  - Eight RDBMS
- Fake data generator
  - Based on Scan Report
  - Random values per column



Working folder: /Users/Maxim/Develop/OHDSI/WhiteRabbit

Source data location

Data type: PostgreSQL
Server location: 127.0.0.1/tutorial_etl
User name: Maxim
Password:
Database name: raw_synthea
Delimiter: ,

Console

Tables to scan:
allergies
careplans
conditions
encounters
imaging_studies
immunizations
medications
observations
organizations
patients
procedures

☑ Scan field values Min cell coun

Console

Scan report file: /Users/Maxim/2-ScanReport-advanced.xlsx

Target data location

Data type: Delimited text files
Server location: 127.0.0.1
User name:
Password:
Database name:
CSV Format: Default (comma, CRLF)

Max rows per table: 10,000

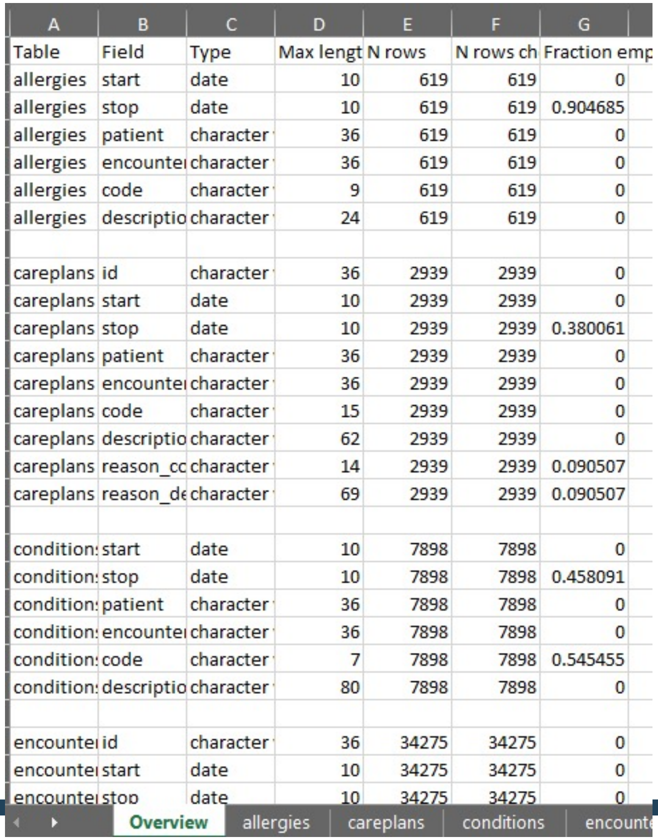
Console

```
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/encounters.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/imaging_studies.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/immunizations.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/medications.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/observations.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/organizations.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/patients.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/procedures.csv
Generating table /Users/Maxim/Develop/OHDSI/WhiteRabbit/providers.csv
Sep 27, 2019 4:12:05 PM        Done
```

Why create a scan report?

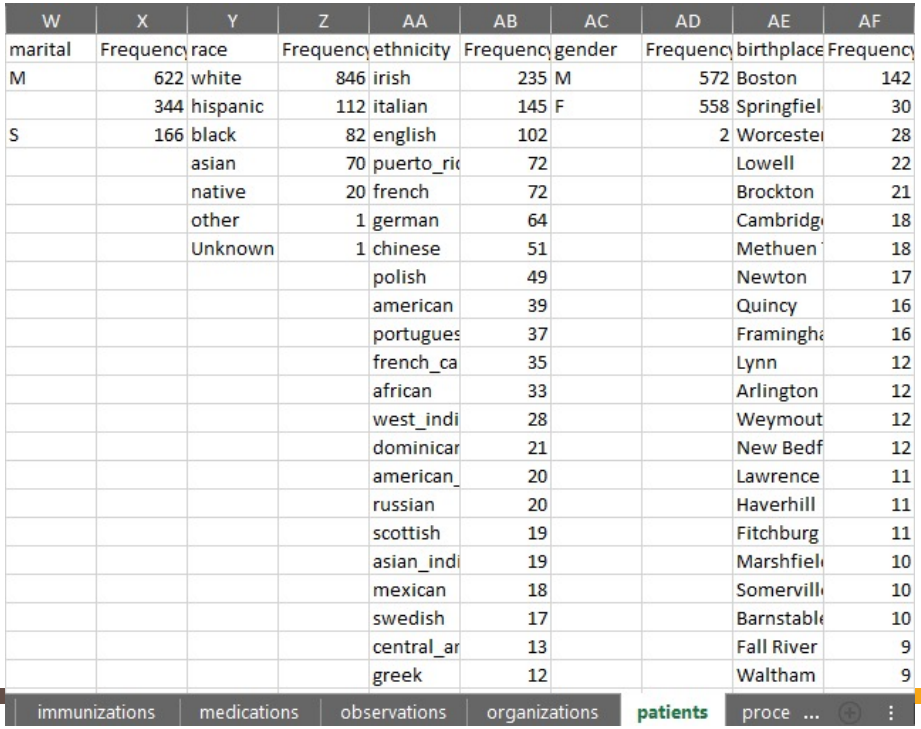
- Helps understanding the structure of your source data
- Input for Rabbit in a Hat

Table/Field Overview



A	B	C	D	E	F	G
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

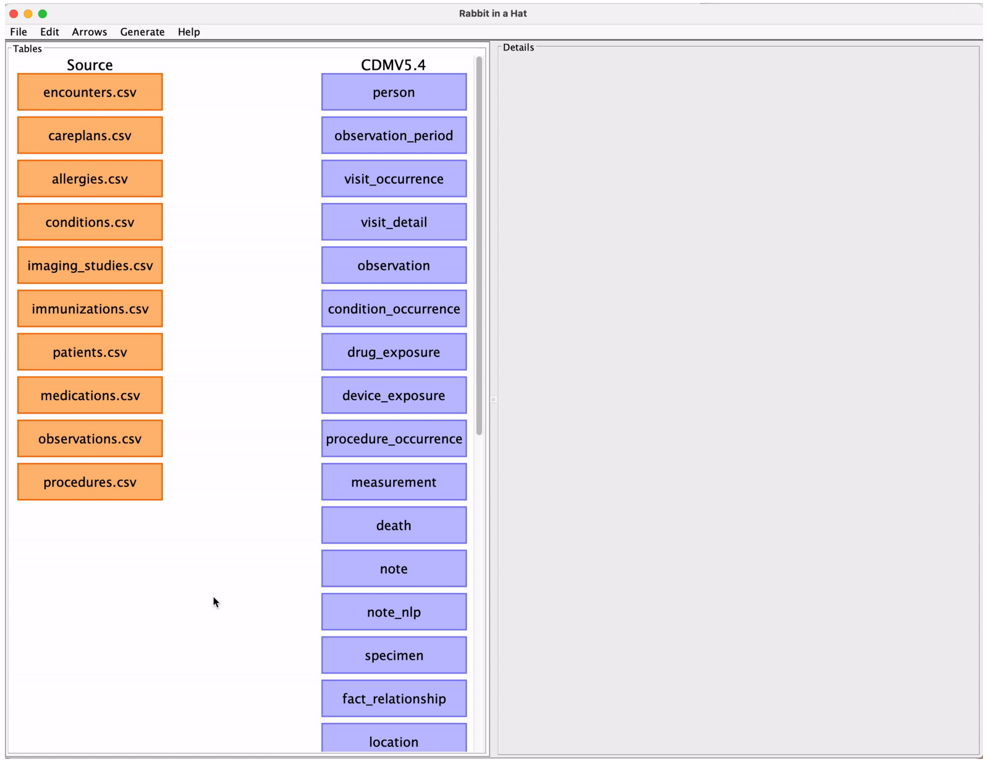
Value counts



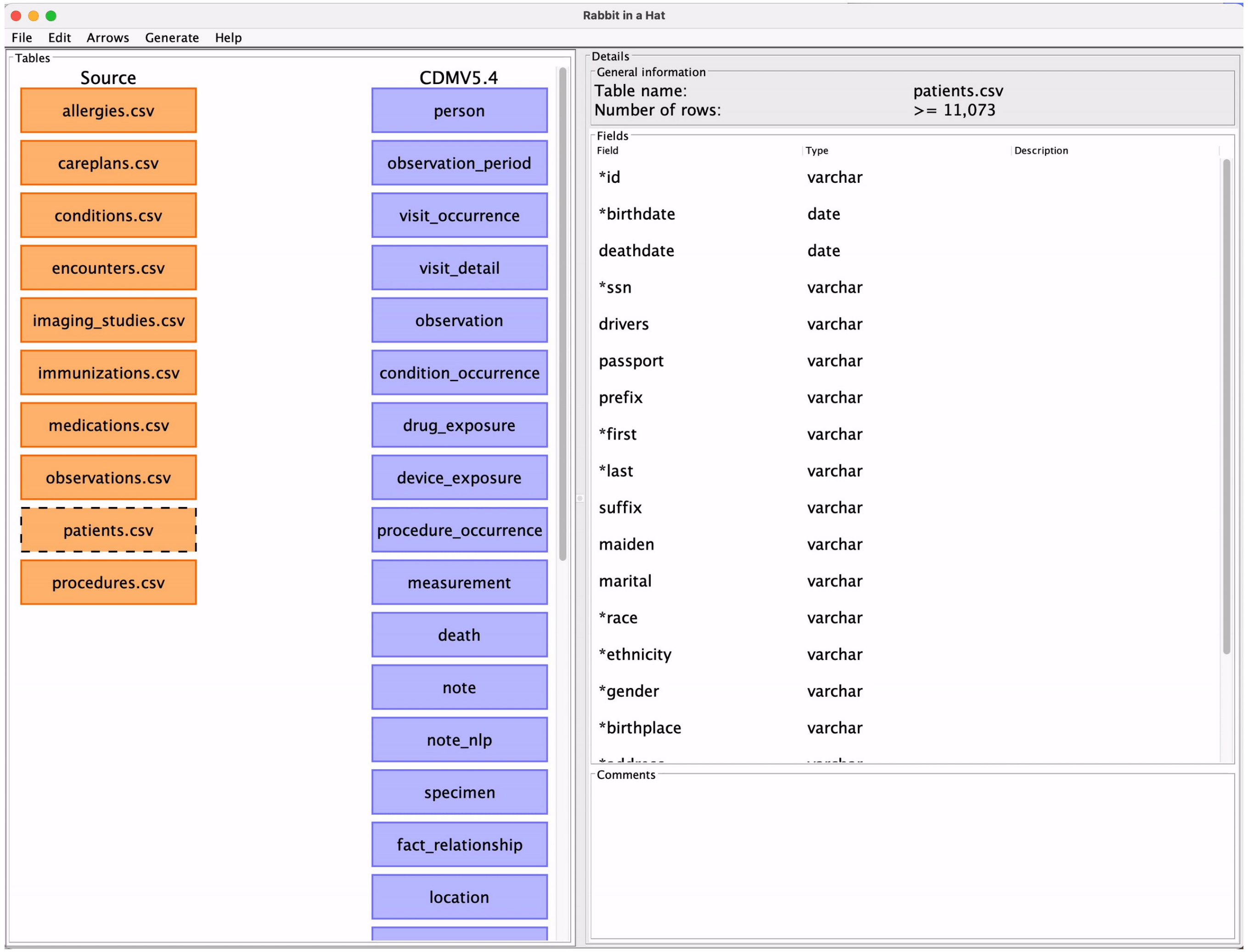
W	X	Y	Z	AA	AB	AC	AD	AE	AF
marital	Frequency	race	Frequency	ethnicity	Frequency	gender	Frequency	birthplace	Frequency
M	622	white	846	irish	235	M	572	Boston	142
	344	hispanic	112	italian	145	F	558	Springfiel	30
S	166	black	82	english	102		2	Worcester	28
		asian	70	puerto_ric	72			Lowell	22
		native	20	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
				dominicar	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			Barnstable	10
				central_ar	13			Fall River	9
				greek	12			Waltham	9

RiaH - Overview

- Facilitates mapping workshop
- Source and Target tables
  - Source tables read from WR Scan Report
  - Target tables from an OMOP CDM definition (or custom)
- Interactively document transformation rules by linking:
  - Table to Table
  - Field to Field



Rabbit in a Hat makes it easy to explore your source data structure!



General information

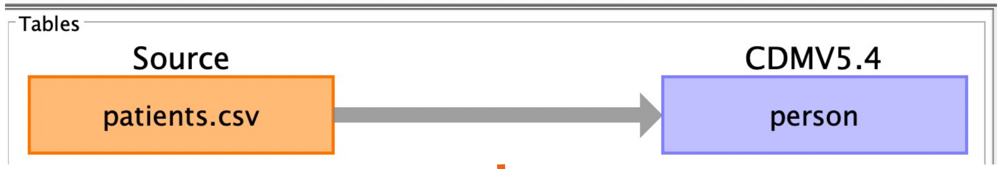
Table name:	patients.csv
Number of rows:	>= 11,073

Fields

Field	Type	Description
*id	varchar	
*birthdate	date	
deathdate	date	
*ssn	varchar	
drivers	varchar	
passport	varchar	
prefix	varchar	
*first	varchar	
*last	varchar	
suffix	varchar	
maiden	varchar	
marital	varchar	
*race	varchar	
*ethnicity	varchar	
*gender	varchar	
*birthplace	varchar	

Create the mapping specification

Table to Table



Tables

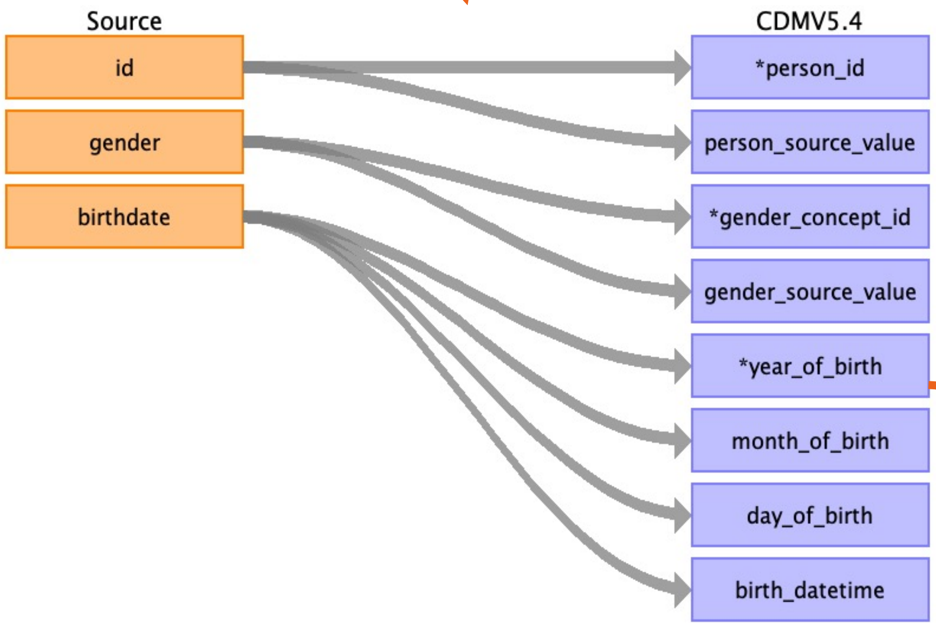
Source

patients.csv

CDMV5.4

person

Field to Field



Source

id

gender

birthdate

CDMV5.4

*person_id

person_source_value

*gender_concept_id

gender_source_value

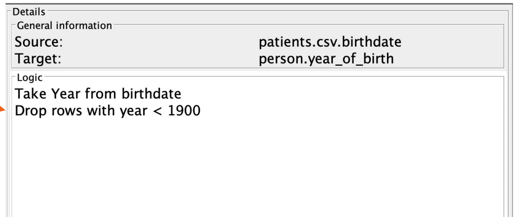
*year_of_birth

month_of_birth

day_of_birth

birth_datetime

Document mapping decisions and logic



Details

General information

Source: patients.csv.birthdate

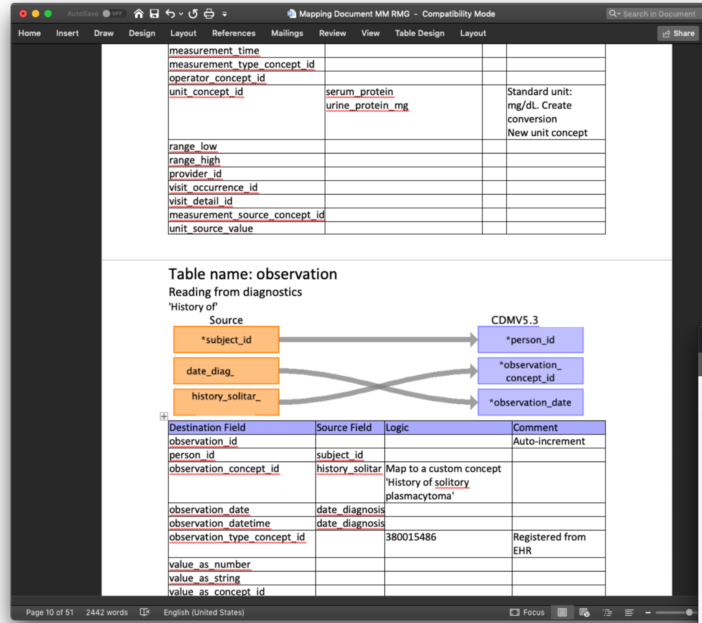
Target: person.year_of_birth

Logic

Take Year from birthdate
Drop rows with year < 1900

RiaH – Export mapping definition

Word document



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

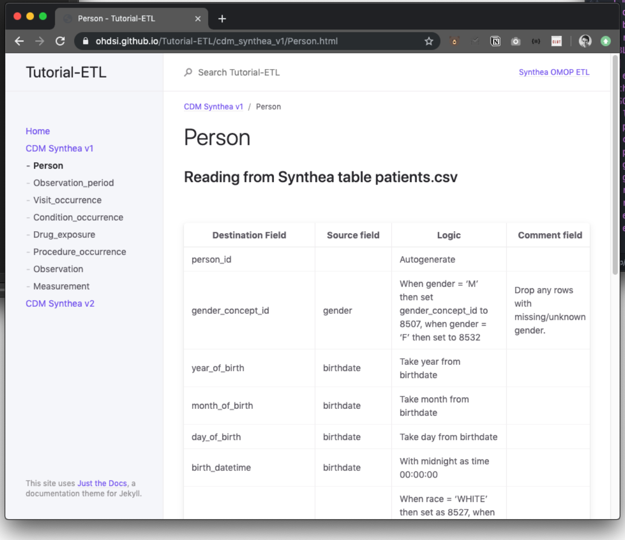
Table name: observation

Reading from diagnostics

'History of'

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

Html



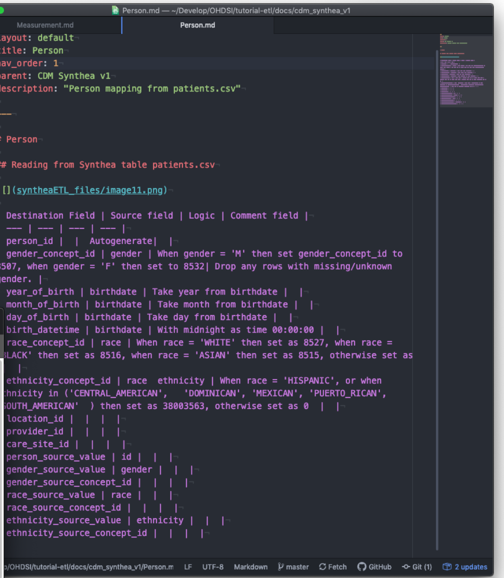
Person

Reading from Synthea table patients.csv

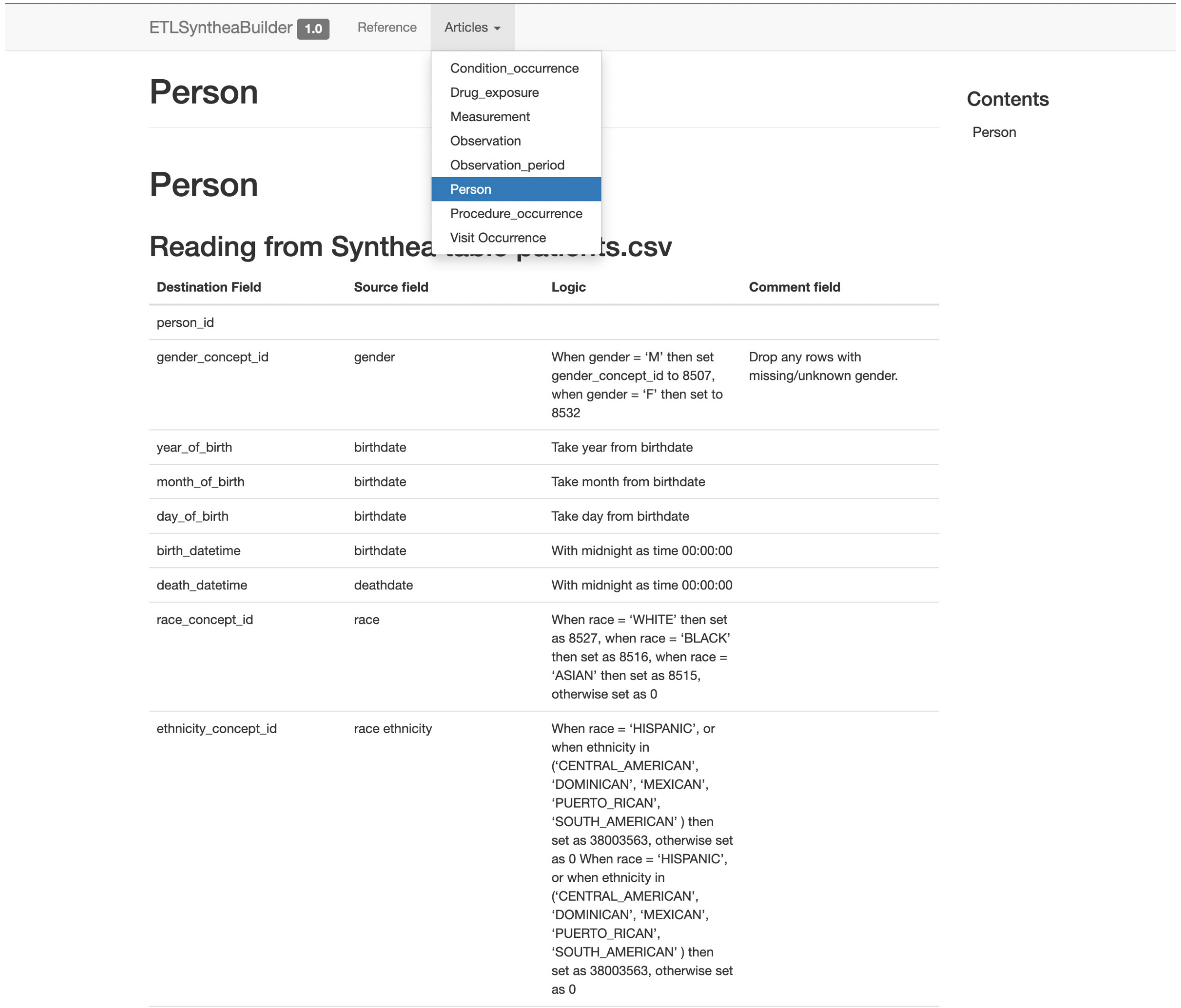
Destination Field	Source field	Logic	Comment field
person_id		Autogenerate	
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.
year_of_birth	birthdate	Take year from birthdate	
month_of_birth	birthdate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

This site uses Just the Docs, a documentation theme for Jekyll.

Markdown documents


Person.md — ~/Develop/OHDSI/tutorial-etl/docs/cdm_synthea_v1
Measurement.md
Person.md
/OHDSI/tutorial-etl/docs/cdm_synthea_v1/Person.m
LF
UTF-8
Markdown
master
Fetch
GitHub
Git (1)
2 updates

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

---

# Person

## Reading from Synthea table patients.csv

![](syntheaETL_files/image11.png)

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



Person

Person

Reading from Synthea table patients.csv

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.
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	
death_datetime	deathdate	With midnight as time 00:00:00	
race_concept_id	race	When race = 'WHITE' then set as 8527, when race = 'BLACK' then set as 8516, when race = 'ASIAN' then set as 8515, otherwise set as 0	
ethnicity_concept_id	race ethnicity	When race = 'HISPANIC', or when ethnicity in ('CENTRAL_AMERICAN', 'DOMINICAN', 'MEXICAN', 'PUERTO_RICAN', 'SOUTH_AMERICAN') then set as 38003563, otherwise set as 0 When race = 'HISPANIC', or when ethnicity in ('CENTRAL_AMERICAN', 'DOMINICAN', 'MEXICAN', 'PUERTO_RICAN', 'SOUTH_AMERICAN') then set as 38003563, otherwise set as 0	

Publish your ETL document!
Example: https://ohdsi.github.io/ETL-Synthea/articles/person.html

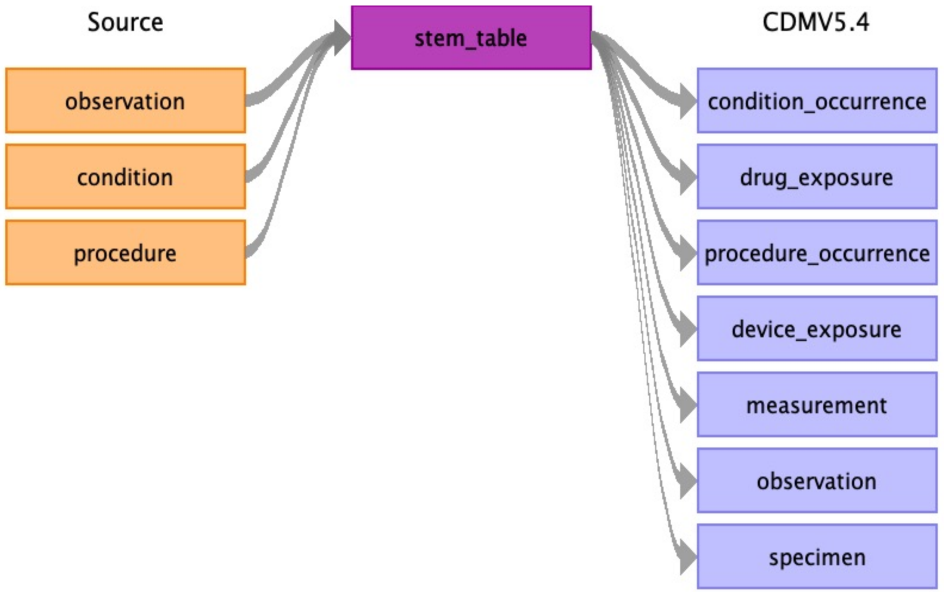
Vocabulary drives data movement: the RiaH Stem table

e.g. not all codes from condition table will be actual condition occurrences.

- SNOMED:161833006 “Abnormal weight gain ” is an Observation

The RiaH STEM Table is a domain-agnostic intermediate.

Relations to OMOP domains are provided.



Mapping Workshop Checklist

- ✓General database information
- ✓Technical specifications
- ✓Data Provenance
- ✓Use Cases
- ✓Goals of the transformation
- ✓Refreshes (data & ETL)
- ✓Data restrictions (ethical)
- ✓Contact for access (data manager)

💡 Get to know the source data!

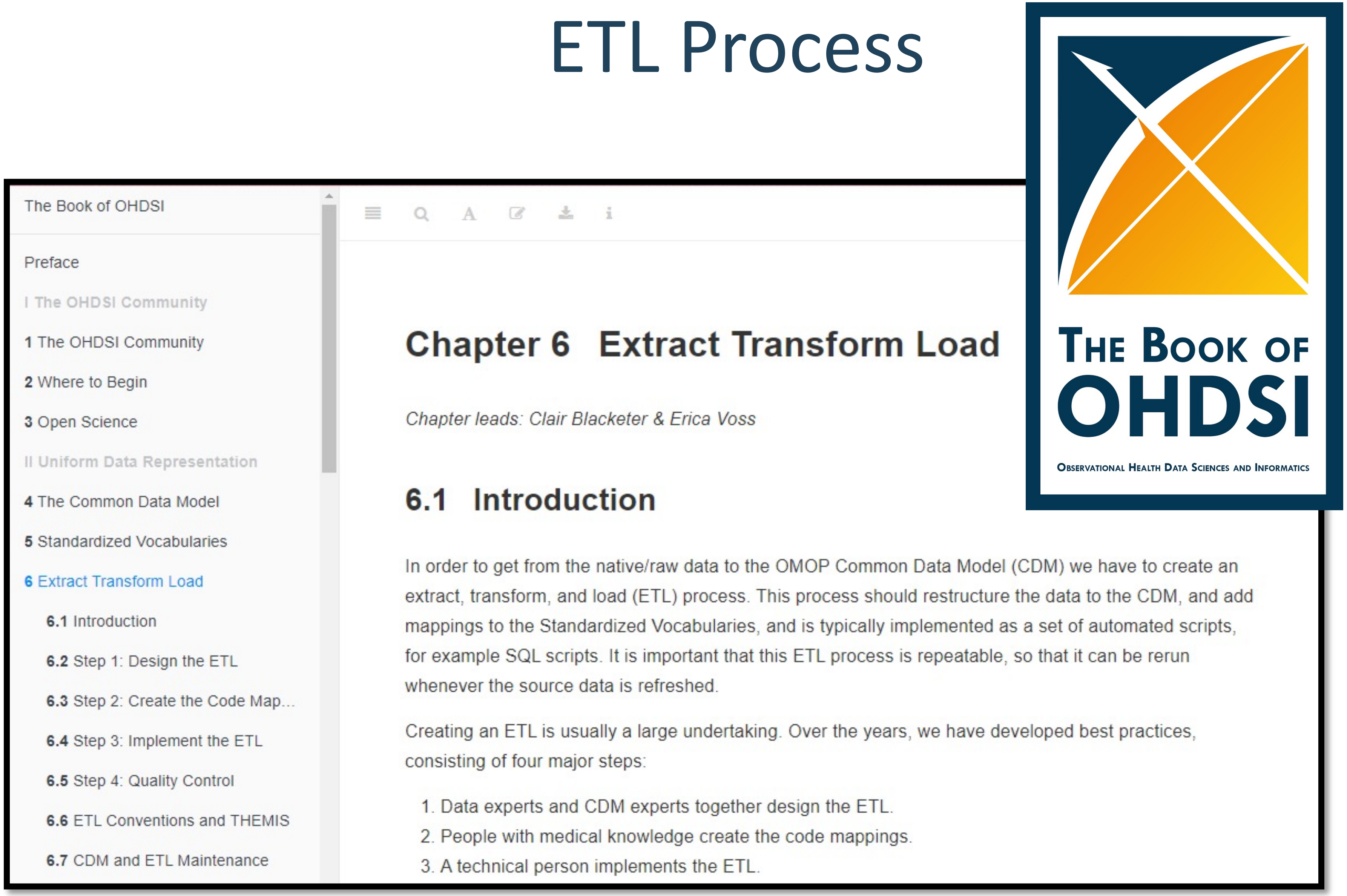
Final remark

Rabbit-in-a-Hat is only for creating the **structural mapping** specification and documenting mapping decisions.

"Which source fields are the input for which OMOP CDM fields."

⚠️ The mapping of values to standard concepts, semantic mapping, is a separate mapping process.
More on semantic mapping with Usagi next week!

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