OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

ETL Training

Agenda

Aug 12 (Korea Time)	Contents	Speakers
9:00 – 9:30 AM	Introduction to ETL / Agile Methodology	Mui Van Zandt
9:30 – 11:30 AM	Source Data Analysis (Lecture, Exercise, Review)	Mui Van Zandt
11:30 – 12:30 PM	Break	
12:30 – 14:30 PM	Vocabulary Mapping (Lecture, Exercise, Review)	Prof. Seng Chan You
14:30 – 14:45 PM	Break	
14:45 – 16:45 PM	ETL Specification Writing (Lecture, Exercise, Review)	Jing Li

Speakers

Seng Chan You (Chan), MD, PhD

Translational Research Assistant Professor

Department of Preventive Medicine, Yonsei University, College of Medicine

Selva Muthu Kumaran Sathappan

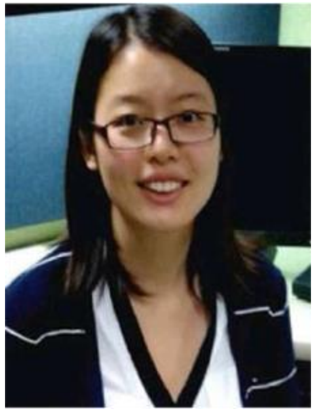
Data Analyst

Saw Swee Hock School of Public Health, National University of Singapore

Mui Van Zandt

Senior Director

OMOP Data Networks, IQVIA



Jing Li

Senior Data Scientist

OMOP Studies, IQVIA

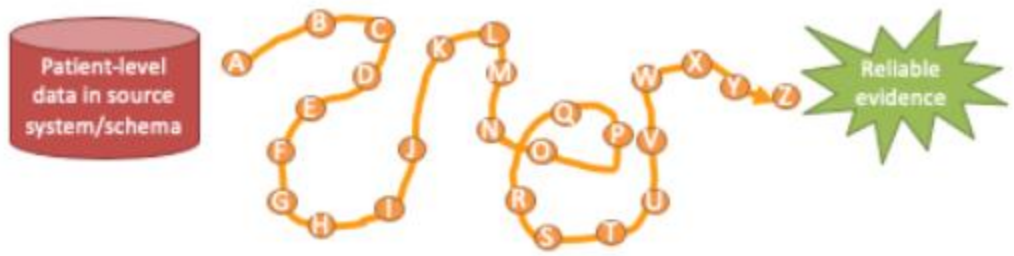
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Introduction to ETL

ETL

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

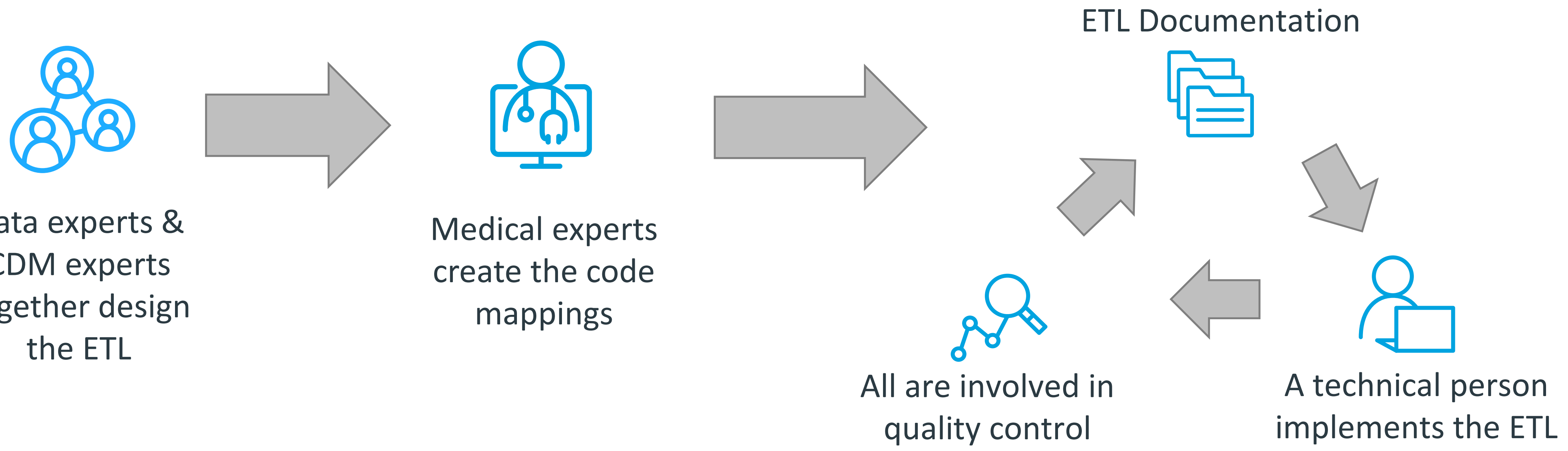


- Goal in ETLing is to standardize the format and terminology
- This tutorial

  - 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 & CDM experts together design the ETL

Medical experts create the code mappings

ETL Documentation

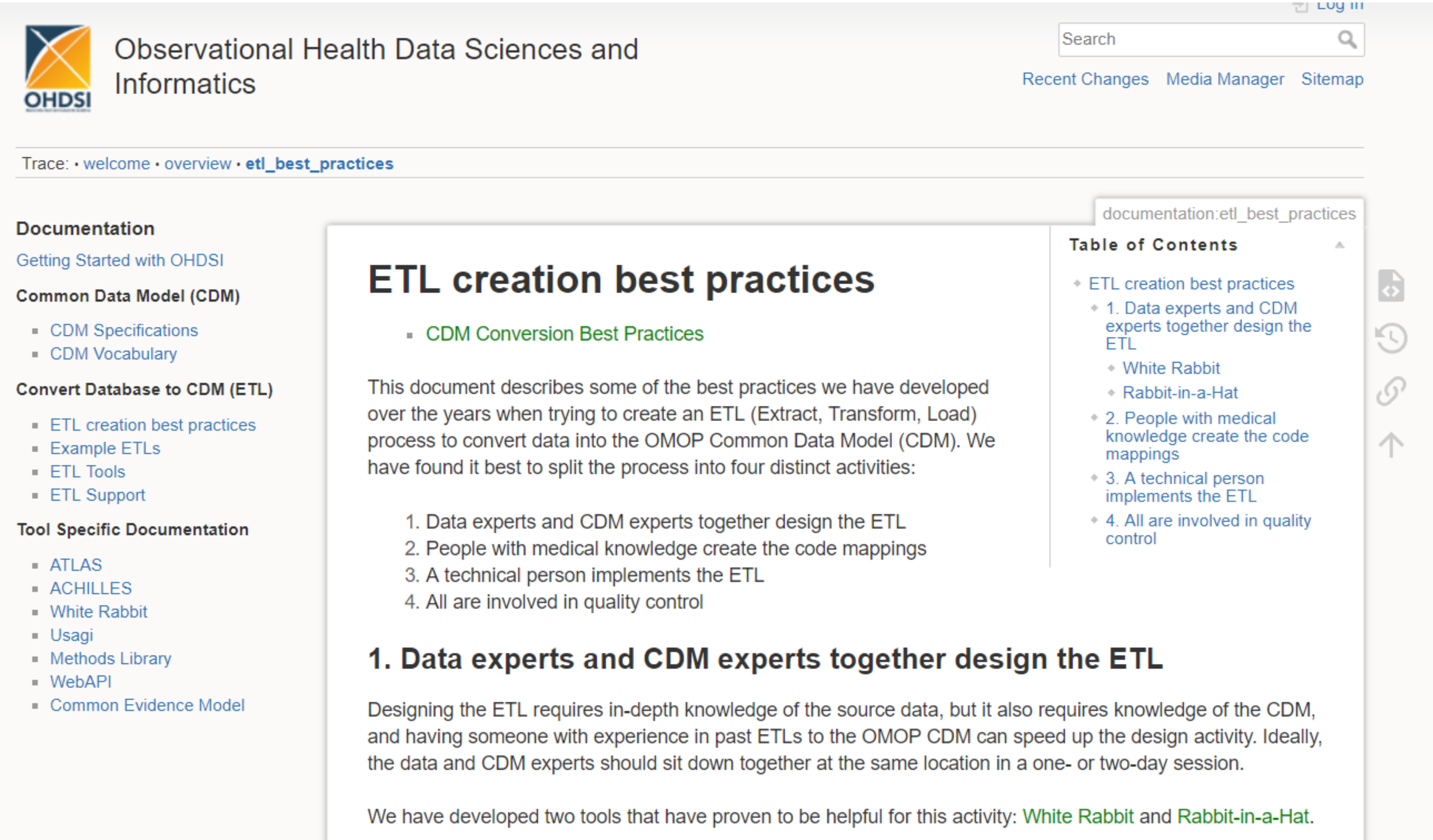
A technical person implements the ETL

All are involved in quality control

Tools

Analysis			Quality Control			Development	
White Rabbit	Rabbit In a Hat	Usagi	Internal Quality Checks	Achilles	Data Quality Dashboard	Jenkins	Code Repository

ETL Process



Observational Health Data Sciences and Informatics

Trace: • welcome • overview • **etl_best_practices**

Documentation

Getting Started with OHDSI

Common Data Model (CDM)

- CDM Specifications
- CDM Vocabulary

Convert Database to CDM (ETL)

- ETL creation best practices
- Example ETLs
- ETL Tools
- ETL Support

Tool Specific Documentation

- ATLAS
- ACHILLES
- White Rabbit
- Usagi
- Methods Library
- WebAPI
- Common Evidence Model

documentation:etl_best_practices

Table of Contents

- ETL creation best practices
  - 1. Data experts and CDM experts together design the ETL
    - White Rabbit
    - Rabbit-in-a-Hat
  - 2. People with medical knowledge create the code mappings
  - 3. A technical person implements the ETL
  - 4. All are involved in quality control

ETL creation best practices

- CDM Conversion Best Practices

This document describes some of the best practices we have developed over the years when trying to create an ETL (Extract, Transform, Load) process to convert data into the OMOP Common Data Model (CDM). We have found it best to split the process into four distinct activities:

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
4. All are involved in quality control

1. Data experts and CDM experts together design the ETL

Designing the ETL requires in-depth knowledge of the source data, but it also requires knowledge of the CDM, and having someone with experience in past ETLs to the OMOP CDM can speed up the design activity. Ideally, the data and CDM experts should sit down together at the same location in a one- or two-day session.

We have developed two tools that have proven to be helpful for this activity: White Rabbit and Rabbit-in-a-Hat.

http://www.ohdsi.org/web/wiki/doku.php?id=documentation:etl_best_practices

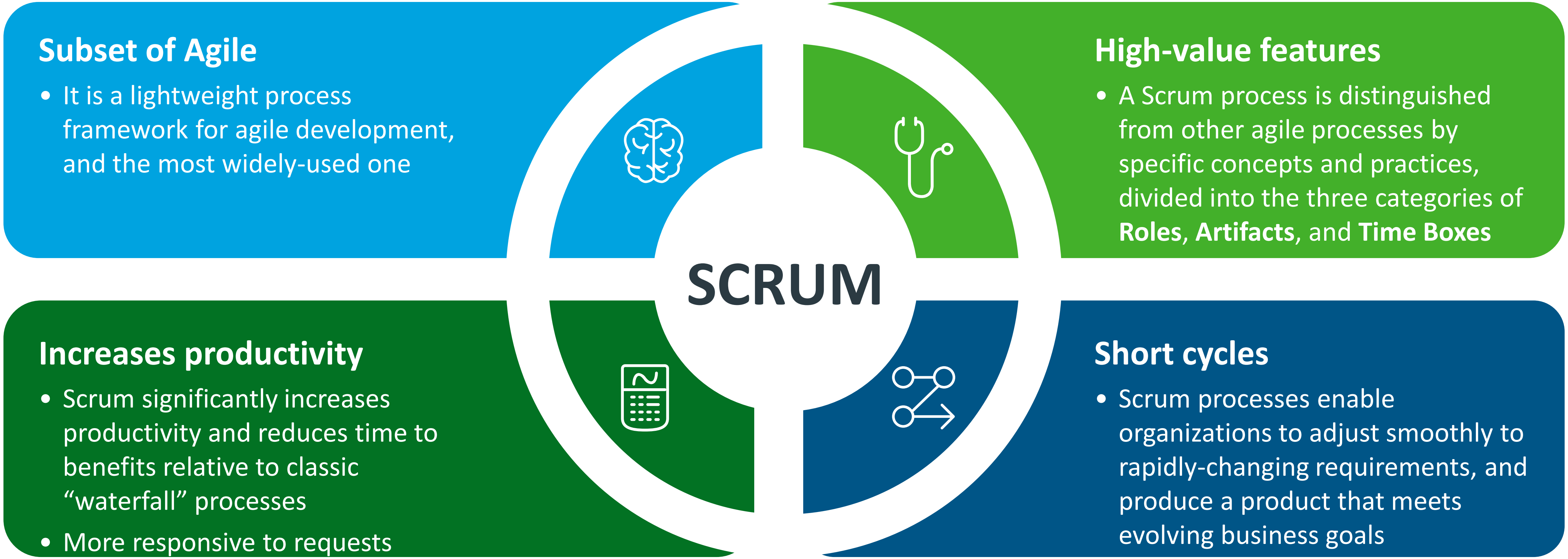
Agile Methodology

What is Agile Scrum

1. Software development methodology
2. Iterative approach
3. Evolves through collaboration
4. Self organizing cross functional team

https://www.cprime.com/resources/what-is-agile-what-is-scrum/

Benefits of Agile Scrum



SCRUM

Subset of Agile

- It is a lightweight process framework for agile development, and the most widely-used one

High-value features

- A Scrum process is distinguished from other agile processes by specific concepts and practices, divided into the three categories of **Roles**, **Artifacts**, and **Time Boxes**

Increases productivity

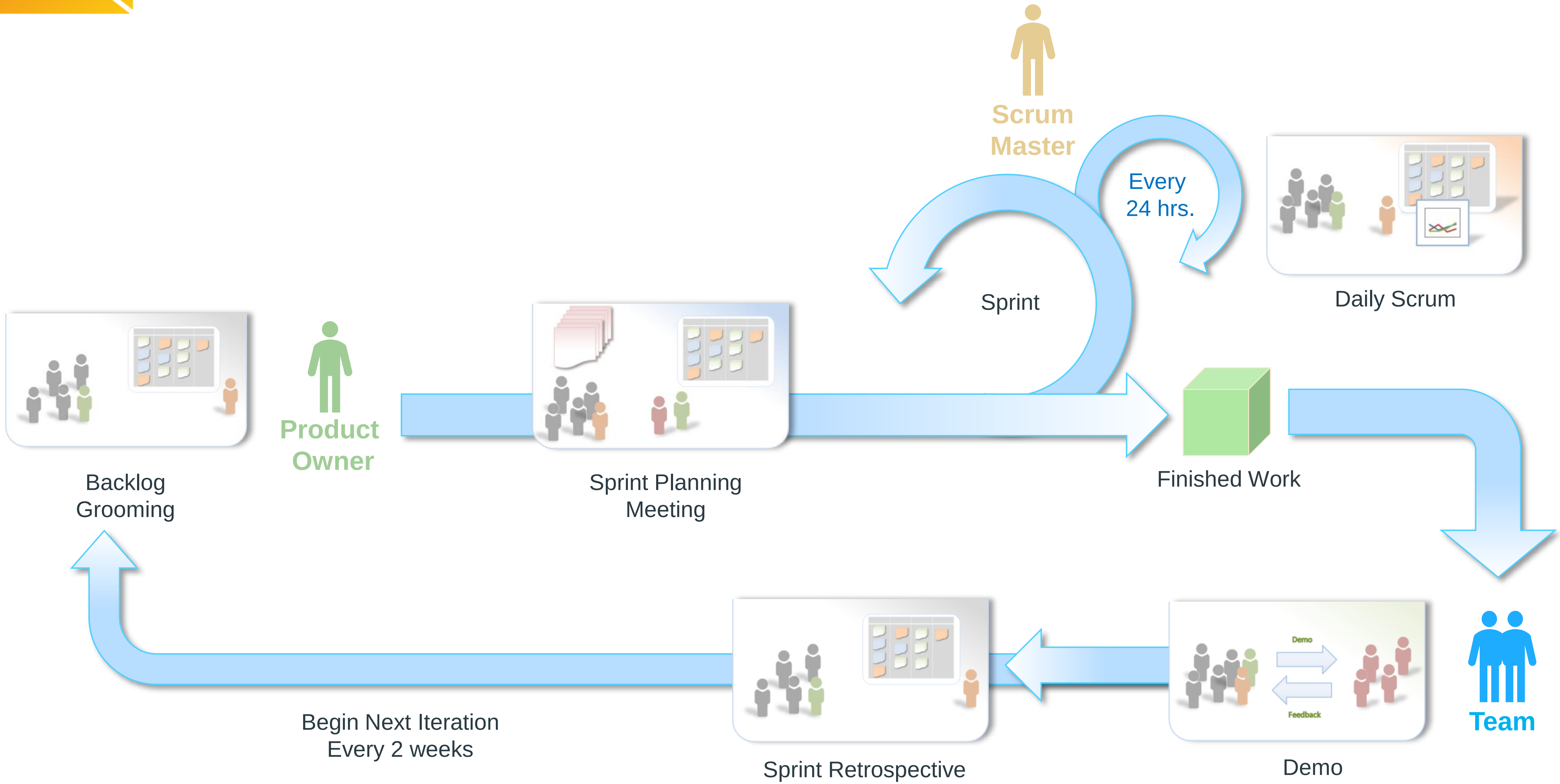
- Scrum significantly increases productivity and reduces time to benefits relative to classic “waterfall” processes
- More responsive to requests

Short cycles

- Scrum processes enable organizations to adjust smoothly to rapidly-changing requirements, and produce a product that meets evolving business goals

https://www.cprime.com/resources/what-is-agile-what-is-scrum/

Agile Scrum framework



Roles in Agile Scrum

Product Owner

- Leads product definition
- Create, maintain, prioritize Product Backlog
- Communicates status and updates to clients/other stakeholders
- Prioritized defect

Scrum Master

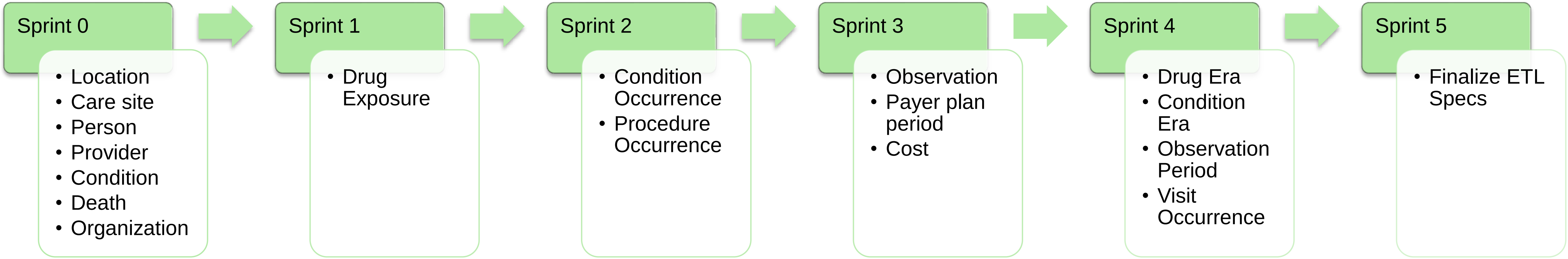
- Responsible for overall status of Sprint
- Help identify and remove impediments
- Blocks “noise” from team
- Ensures retrospective recommendations are executed
- Facilitate all ceremonies

Scrum Team

- “The Do-ers” consisting of 5 people, plus or minus 2
- Co-located - Cross-Functional - Dedicated
- Self-organizing / self-managing, without externally assigned roles
- Communicates commitments with the Product Owner, one Sprint at a time

Typical OMOP Conversion Process

Analysis – Creation of ETL Specs/Stories



Sprint 0
- Location
- Care site
- Person
- Provider
- Condition
- Death
- Organization

Sprint 1
- Drug Exposure

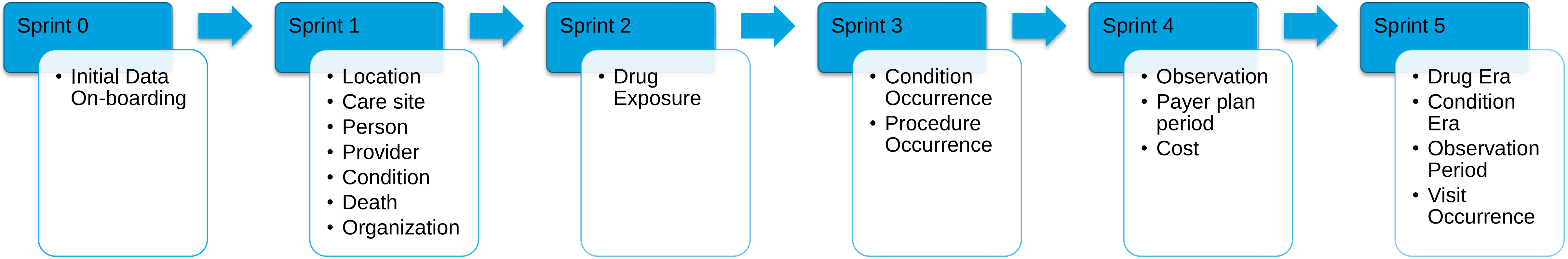
Sprint 2
- Condition Occurrence
- Procedure Occurrence

Sprint 3
- Observation
- Payer plan period
- Cost

Sprint 4
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence

Sprint 5
- Finalize ETL Specs

Development – Implementation/Validation of ETL Specs



Sprint 0
- Initial Data On-boarding

Sprint 1
- Location
- Care site
- Person
- Provider
- Condition
- Death
- Organization

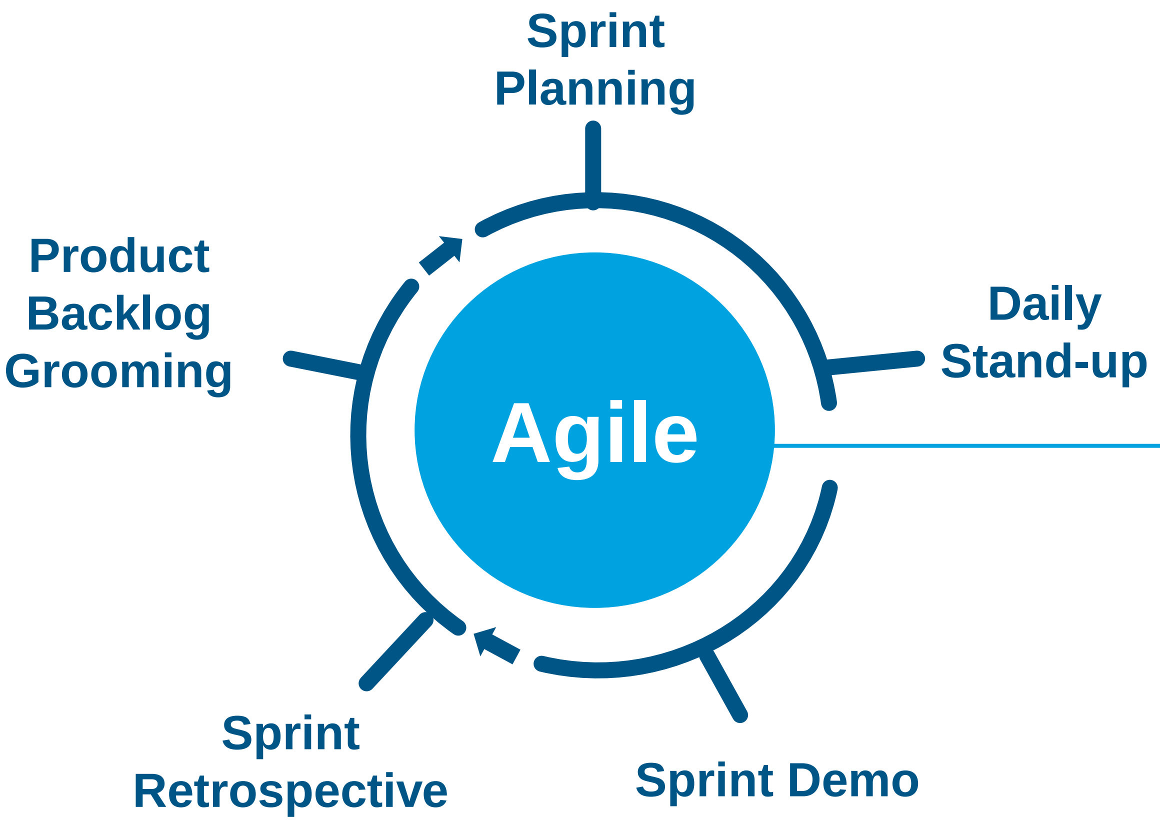
Sprint 2
- Drug Exposure

Sprint 3
- Condition Occurrence
- Procedure Occurrence

Sprint 4
- Observation
- Payer plan period
- Cost

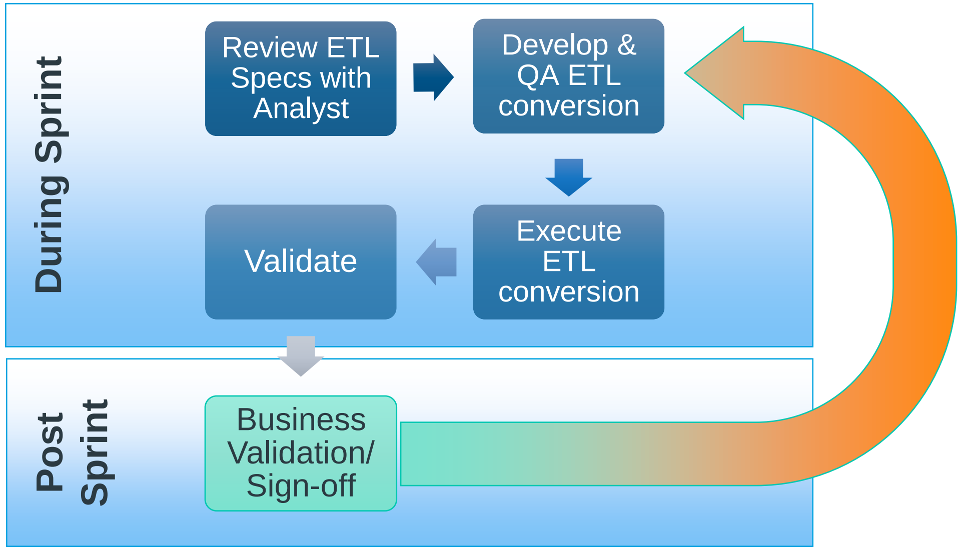
Sprint 5
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence

OMOP Agile Conversion Process



What is Agile?

- Project management & software development
- 2 week sprints
- Promotes continuous adaptation


During Sprint
Review ETL Specs with Analyst
Develop & QA ETL conversion
Execute ETL conversion
Validate
Post Sprint
Business Validation/ Sign-off

Cultural and behavioural changes

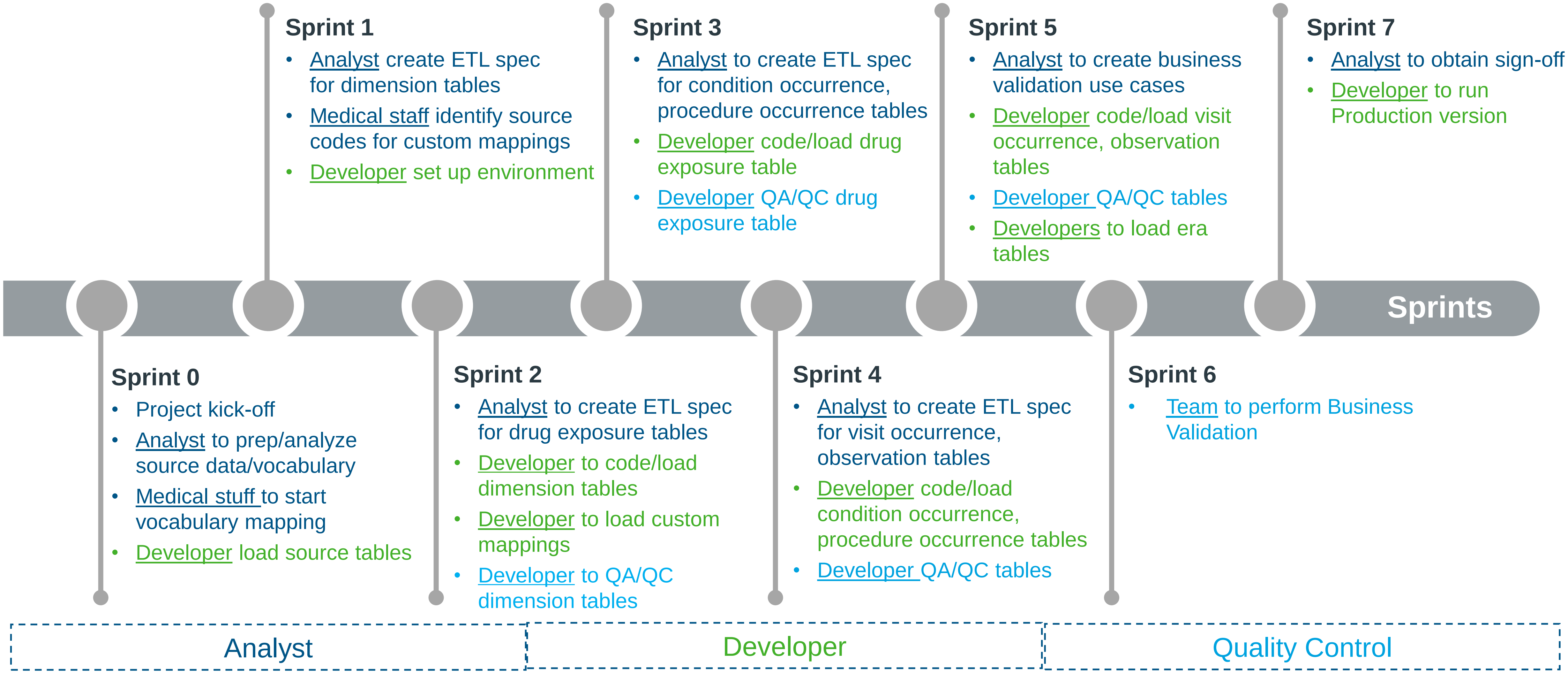
Waterfall

- ✗ Formal Milestone
- ✗ One or two big bang deployments
- ✗ Team spans location and time zones
- ✗ Decision by committee
- ✗ Controlled project management
- ✗ Make a plan and follow it
- ✗ Change requests process management system
- ✗ Not cross functional

Agile

- ✓ Sprint releases
- ✓ Small & frequent MVP deployments
- ✓ Predominately co-located teams
- ✓ Team are empowered to make decisions
- ✓ Scope changes made iteratively
- ✓ Plan continuously and iteratively
- ✓ Adapting change based on need and understanding
- ✓ Cross functional teams

Conversion timeline in sprint – Example



Sprint 0
- Project kick-off
- Analyst to prep/analyze source data/vocabulary
- Medical stuff to start vocabulary mapping
- Developer load source tables

Sprint 1
- Analyst create ETL spec for dimension tables
- Medical staff identify source codes for custom mappings
- Developer set up environment

Sprint 2
- Analyst to create ETL spec for drug exposure tables
- Developer to code/load dimension tables
- Developer to load custom mappings
- Developer to QA/QC dimension tables

Sprint 3
- Analyst to create ETL spec for condition occurrence, procedure occurrence tables
- Developer code/load drug exposure table
- Developer QA/QC drug exposure table

Sprint 4
- Analyst to create ETL spec for visit occurrence, observation tables
- Developer code/load condition occurrence, procedure occurrence tables
- Developer QA/QC tables

Sprint 5
- Analyst to create business validation use cases
- Developer code/load visit occurrence, observation tables
- Developer QA/QC tables
- Developers to load era tables

Sprint 6
- Team to perform Business Validation

Sprint 7
- Analyst to obtain sign-off
- Developer to run Production version

Analyst

Developer

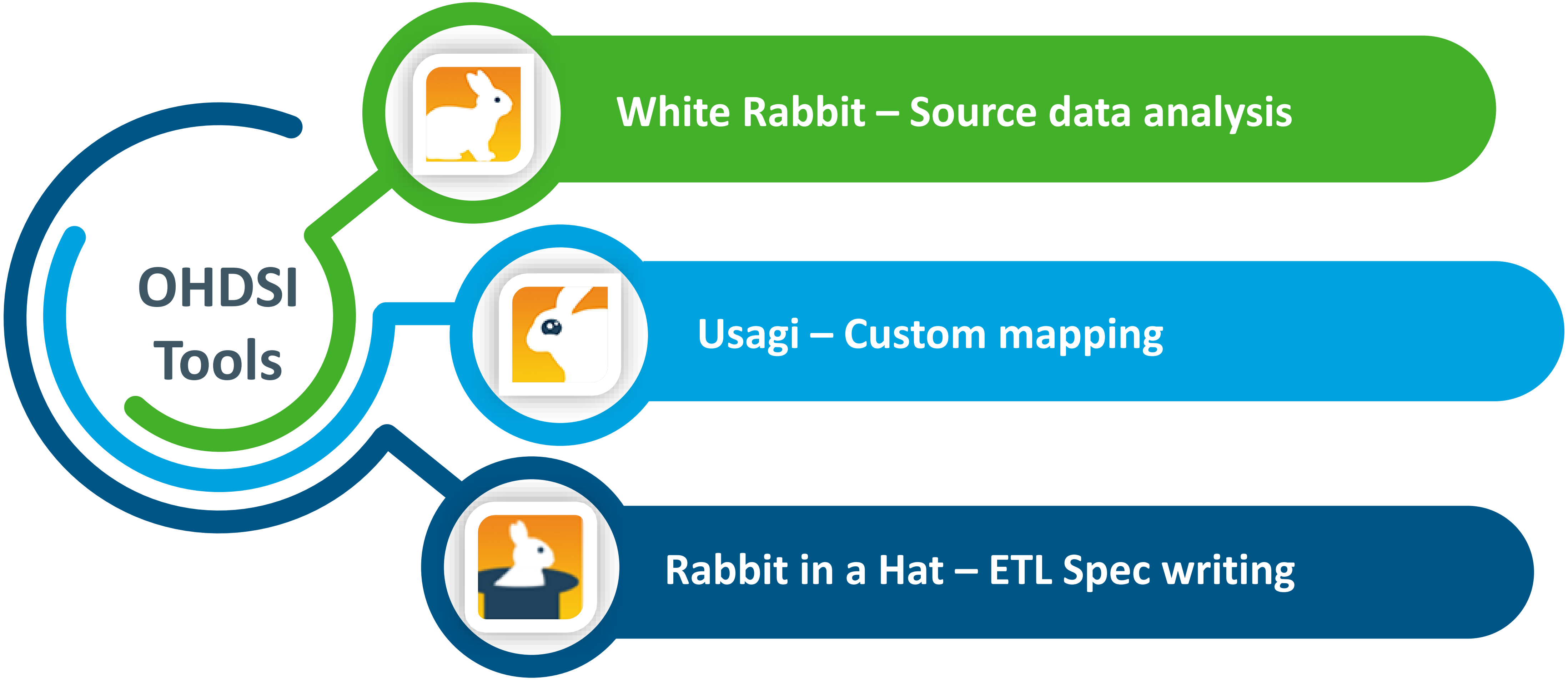
Quality Control

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Source Data Analysis

OHDSI Tools for Analysis



OHDSI Tools

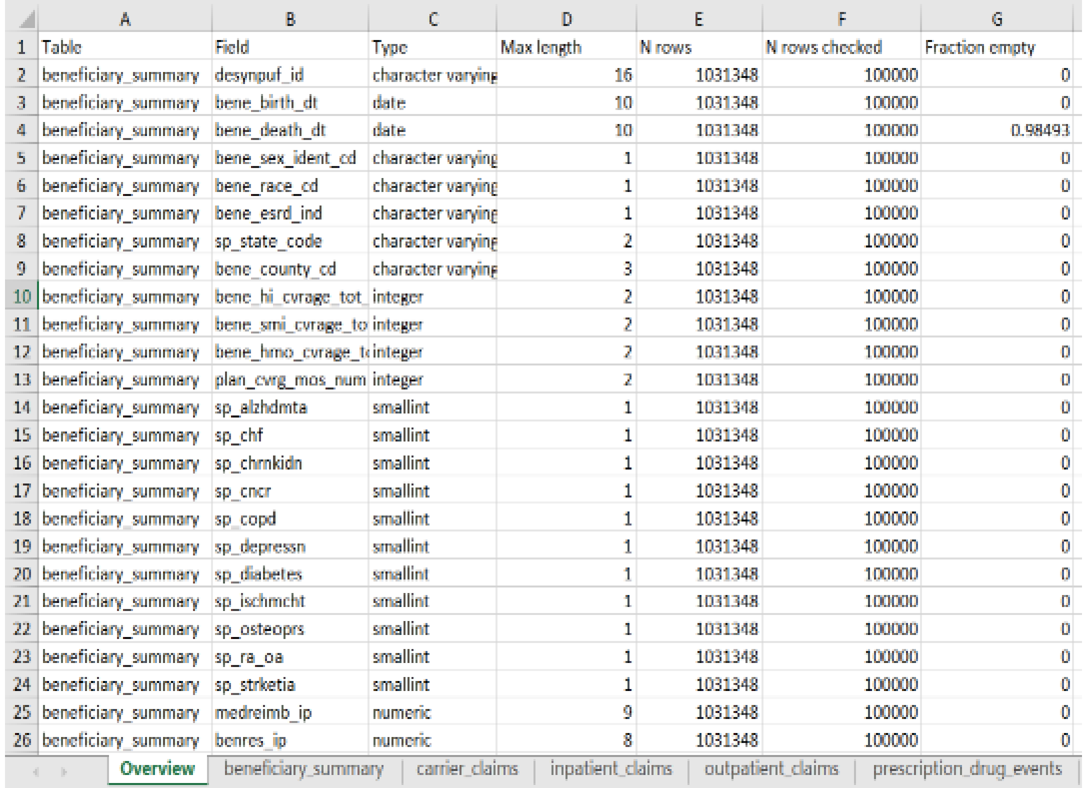
White Rabbit – Source data analysis

Usagi – Custom mapping

Rabbit in a Hat – ETL Spec writing

Source data analysis

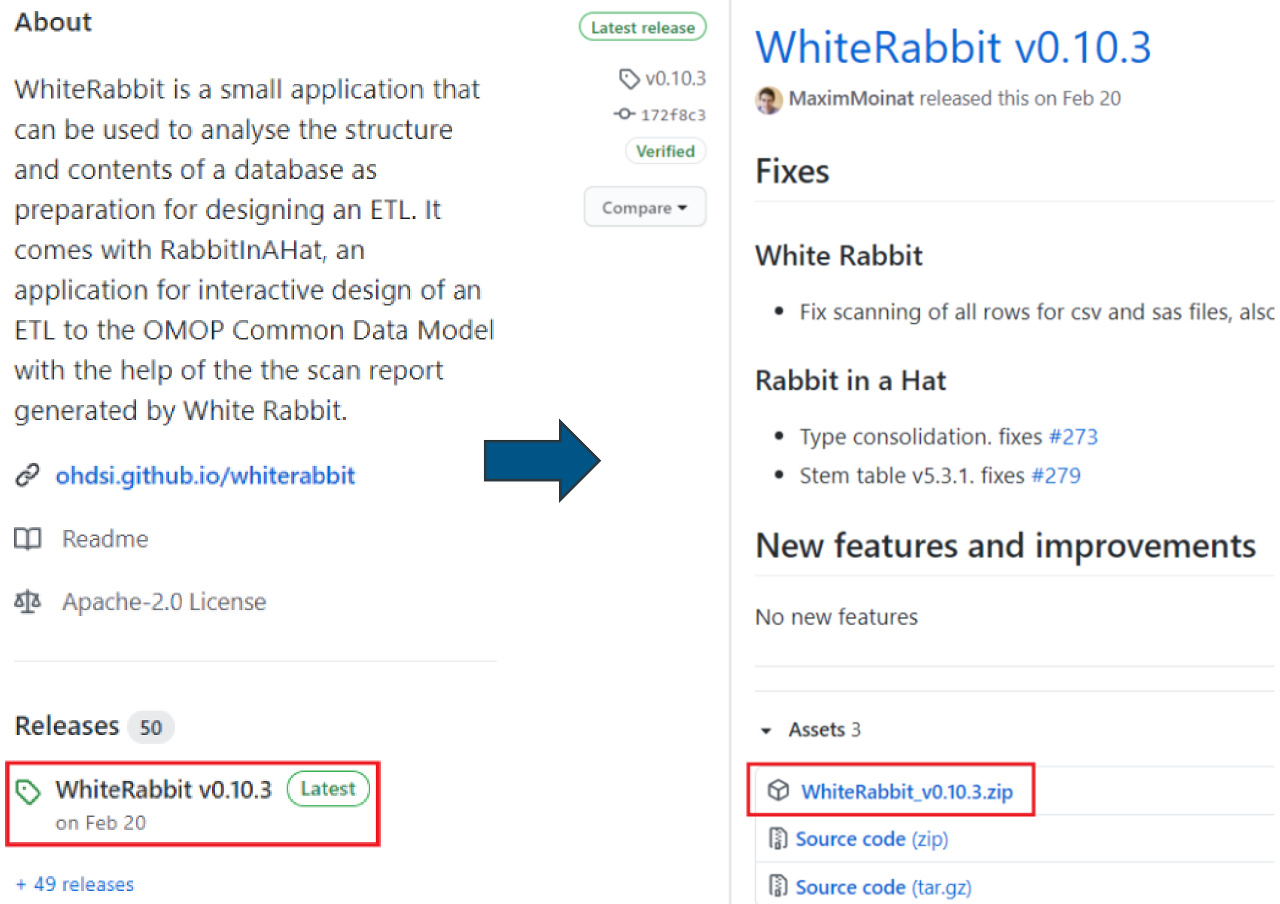
- Used to analyze the structure and content of source data
- Assists with data types, values, frequency, anomalies
- Creates scan report of tables, columns, values
- Starts/continues investigation of source data with data owner
- Used in preparation for creating ETL specification



	A	B	C	D	E	F	G
1	Table	Field	Type	Max length	N rows	N rows checked	Fraction empty
2	beneficiary_summary	desynpuf_id	character varying	16	1031348	100000	0
3	beneficiary_summary	bene_birth_dt	date	10	1031348	100000	0
4	beneficiary_summary	bene_death_dt	date	10	1031348	100000	0.98493
5	beneficiary_summary	bene_sex_ident_cd	character varying	1	1031348	100000	0
6	beneficiary_summary	bene_race_cd	character varying	1	1031348	100000	0
7	beneficiary_summary	bene_esrd_ind	character varying	1	1031348	100000	0
8	beneficiary_summary	sp_state_code	character varying	2	1031348	100000	0
9	beneficiary_summary	bene_county_cd	character varying	3	1031348	100000	0
10	beneficiary_summary	bene_hi_cvrage_tot_	integer	2	1031348	100000	0
11	beneficiary_summary	bene_smi_cvrage_to	integer	2	1031348	100000	0
12	beneficiary_summary	bene_hmo_cvrage_t	integer	2	1031348	100000	0
13	beneficiary_summary	plan_cvrg_mos_num	integer	2	1031348	100000	0
14	beneficiary_summary	sp_alzhdmta	smallint	1	1031348	100000	0
15	beneficiary_summary	sp_chf	smallint	1	1031348	100000	0
16	beneficiary_summary	sp_chrnkidn	smallint	1	1031348	100000	0
17	beneficiary_summary	sp_cncr	smallint	1	1031348	100000	0
18	beneficiary_summary	sp_copd	smallint	1	1031348	100000	0
19	beneficiary_summary	sp_depressn	smallint	1	1031348	100000	0
20	beneficiary_summary	sp_diabetes	smallint	1	1031348	100000	0
21	beneficiary_summary	sp_ischmcht	smallint	1	1031348	100000	0
22	beneficiary_summary	sp_osteoprs	smallint	1	1031348	100000	0
23	beneficiary_summary	sp_ra_oa	smallint	1	1031348	100000	0
24	beneficiary_summary	sp_strketia	smallint	1	1031348	100000	0
25	beneficiary_summary	medreimb_ip	numeric	9	1031348	100000	0
26	beneficiary_summary	benres_ip	numeric	8	1031348	100000	0

Getting White Rabbit

1. White Rabbit Download
   https://github.com/OHDSI/WhiteRabbit
2. Find the “Latest Release” and download the WhiteRabbit zip file
3. Unzip the download
4. Double-click on *bin/whiteRabbit.bat* on Windows to start White Rabbit



About

WhiteRabbit is a small application that can be used to analyse the structure and contents of a database as preparation for designing an ETL. It comes with RabbitInAHat, an application for interactive design of an ETL to the OMOP Common Data Model with the help of the scan report generated by White Rabbit.

ohdsi.github.io/whiterabbit

Readme

Apache-2.0 License

Releases 50

WhiteRabbit v0.10.3 Latest
on Feb 20

+ 49 releases

WhiteRabbit v0.10.3

MaximMoinat released this on Feb 20

Fixes

White Rabbit

- Fix scanning of all rows for csv and sas files, alsc

Rabbit in a Hat

- Type consolidation. fixes #273
- Stem table v5.3.1. fixes #279

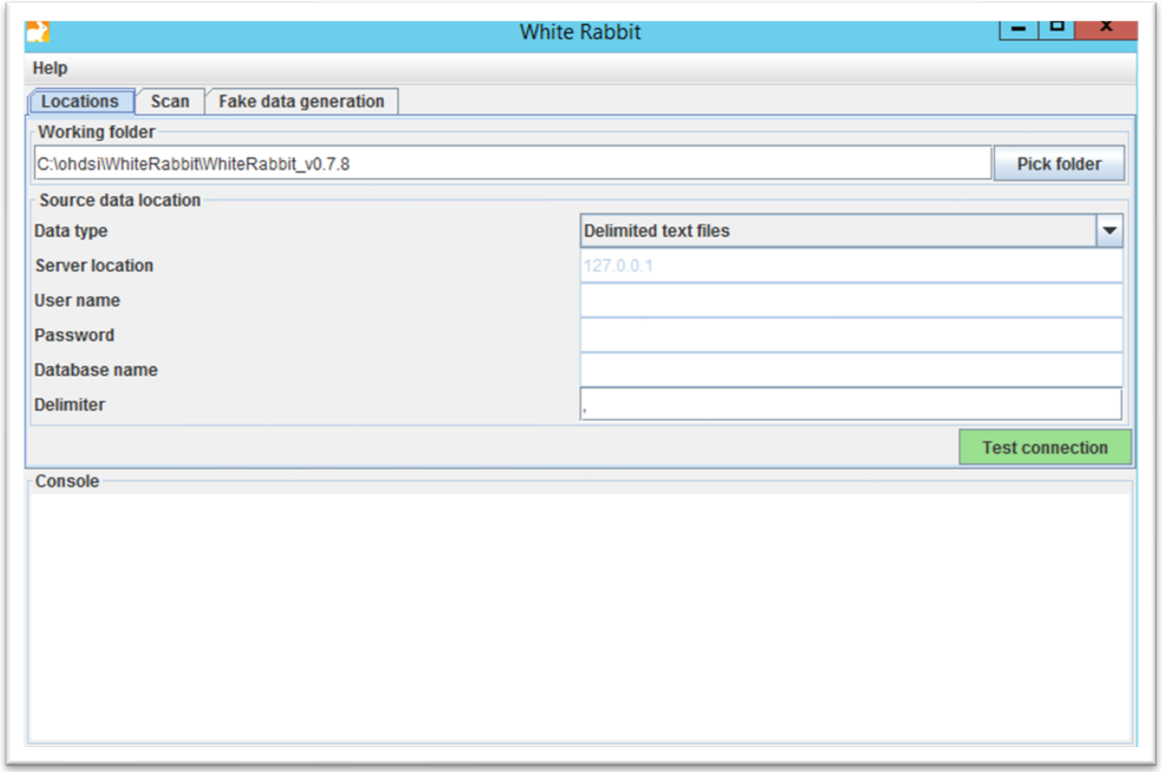
New features and improvements

No new features

Assets 3

- WhiteRabbit_v0.10.3.zip
- Source code (zip)
- Source code (tar.gz)

White Rabbit – Location and Scan



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

Data type: Delimited text files

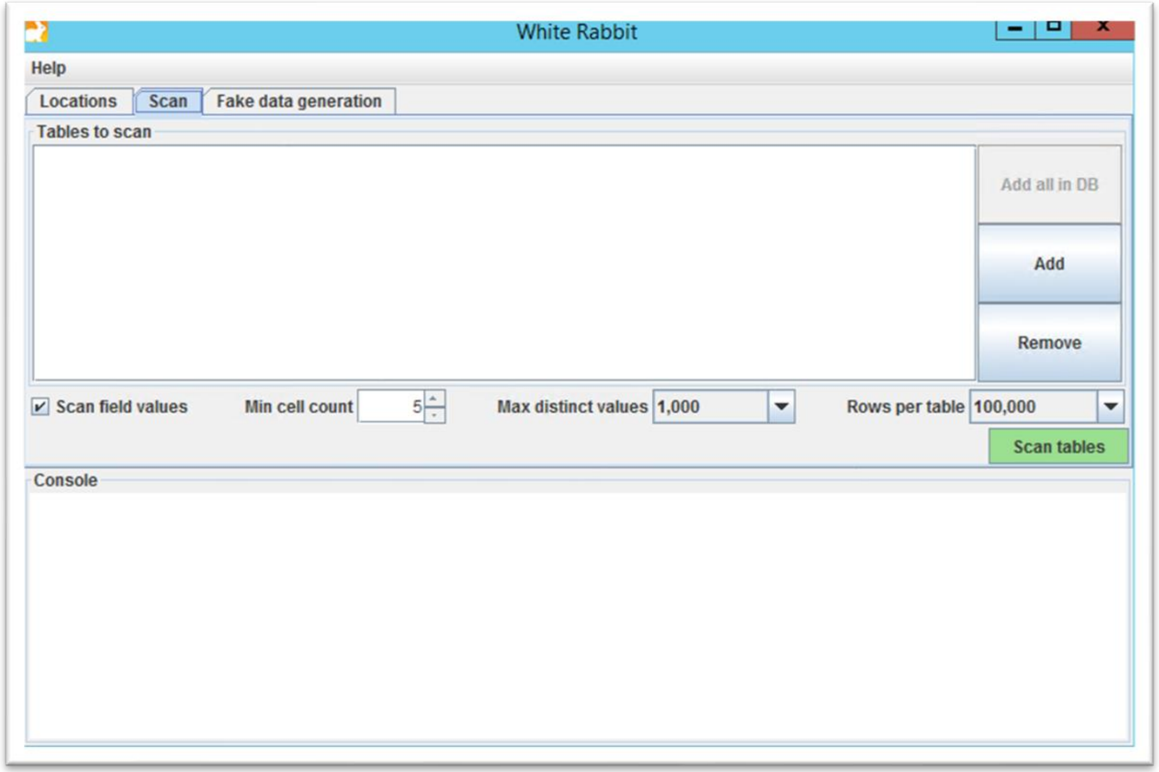
Server location: 127.0.0.1

User name:

Password:

Database name:

Delimiter: ,



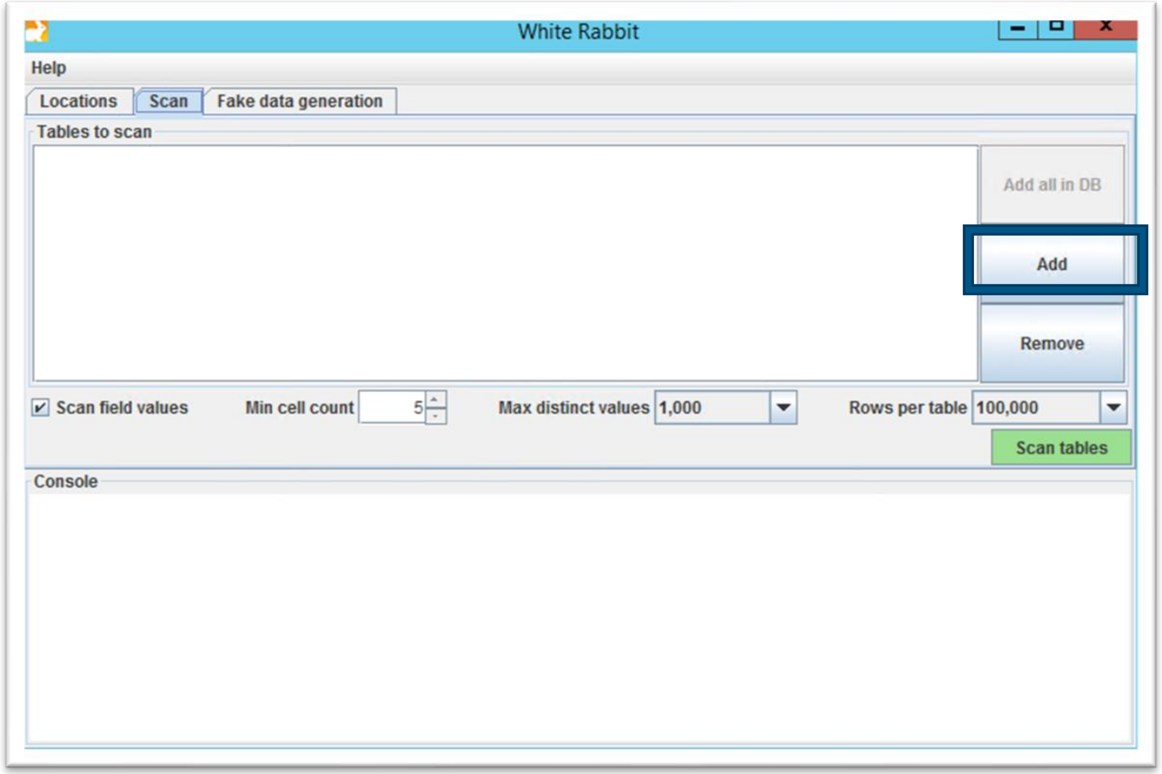
☑ Scan field values

Min cell count: 5

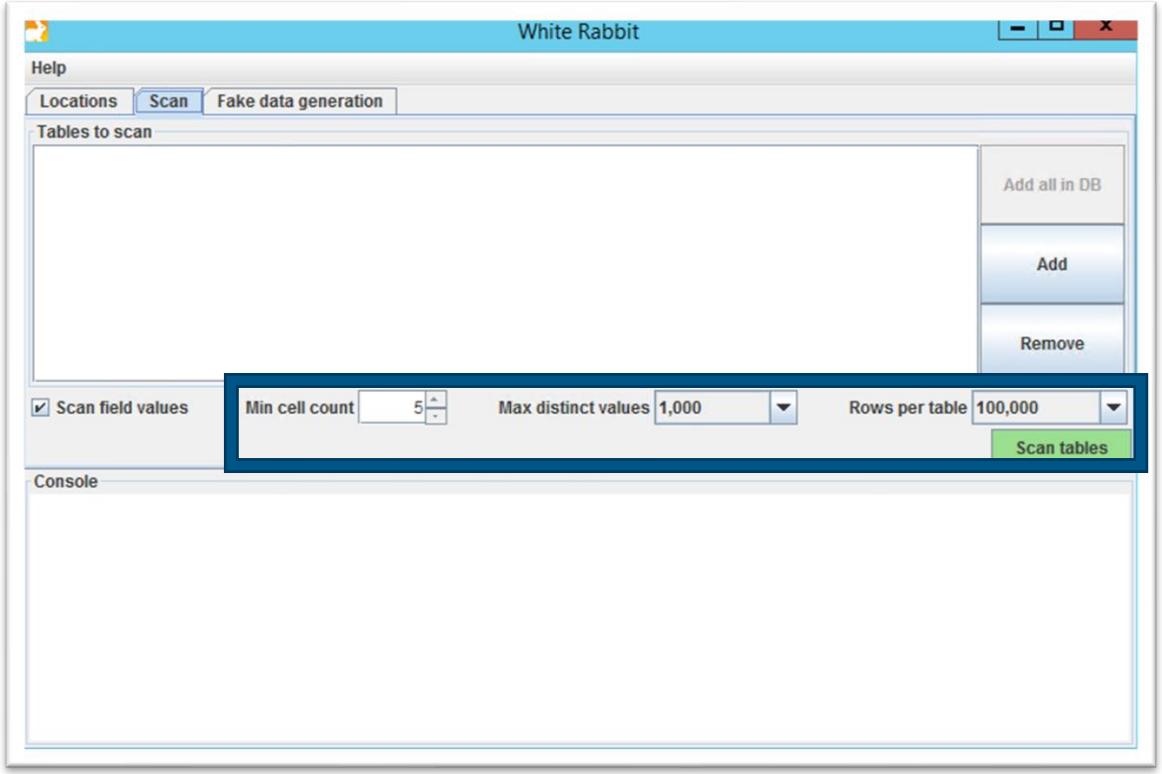
Max distinct values: 1,000

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



☑ Scan field values
Min cell count 5
Max distinct values 1,000
Rows per table 100,000

Reading the Scan

Overview Tab

Provides the definition of each table analyzed, there will only be one tab of this type

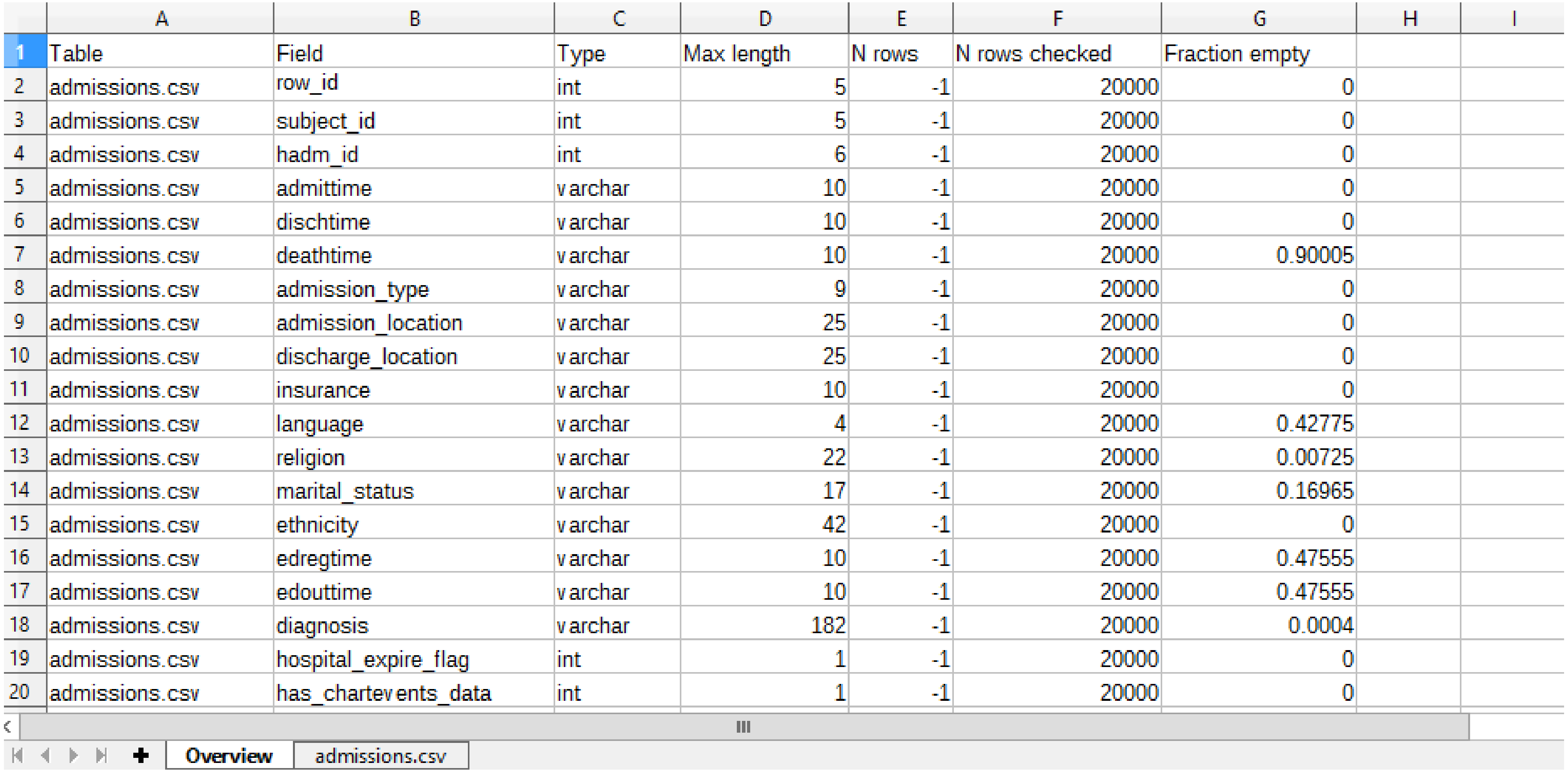
Series of tabs in an XLSX file

Table Tabs

A summary column for each field, there will be as many tabs as tables selected to analyze

Overview Tab

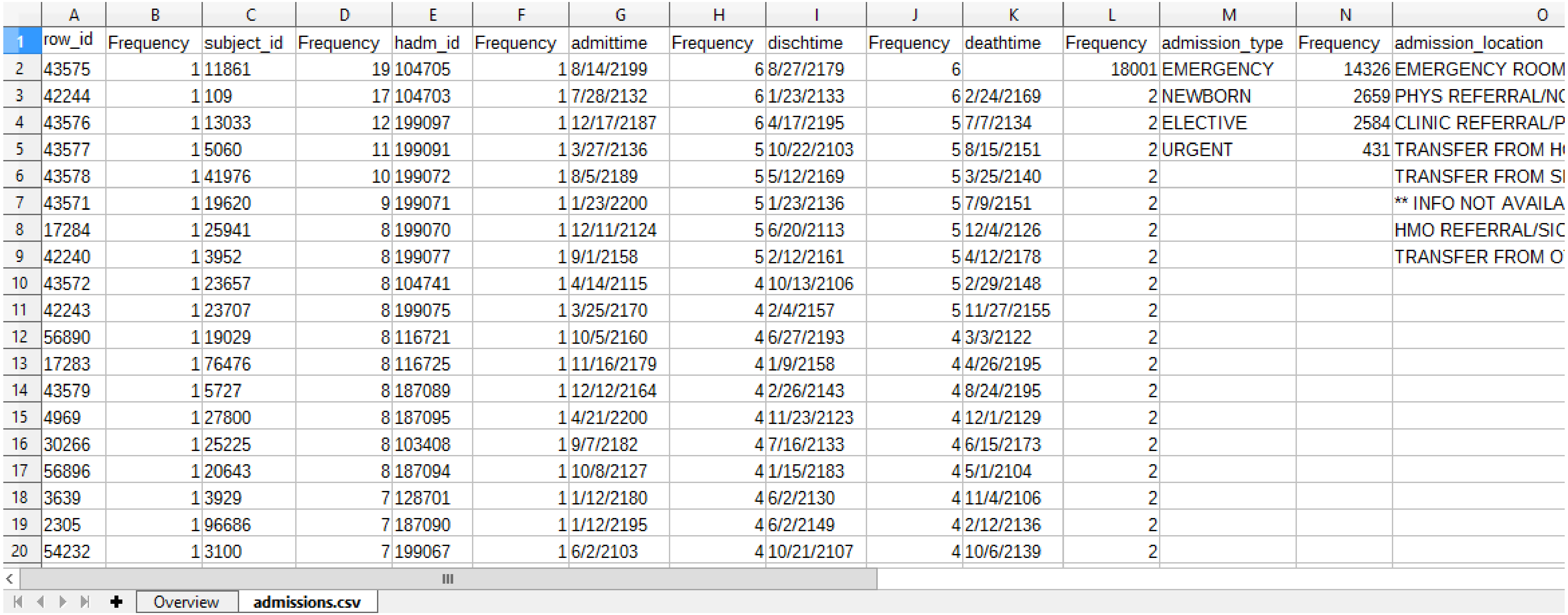
- Defines the tables you scanned



	A	B	C	D	E	F	G	H	I
1	Table	Field	Type	Max length	N rows	N rows checked	Fraction empty		
2	admissions.csv	row_id	int	5	-1	20000	0		
3	admissions.csv	subject_id	int	5	-1	20000	0		
4	admissions.csv	hadm_id	int	6	-1	20000	0		
5	admissions.csv	admittime	varchar	10	-1	20000	0		
6	admissions.csv	dischtime	varchar	10	-1	20000	0		
7	admissions.csv	deathtime	varchar	10	-1	20000	0.90005		
8	admissions.csv	admission_type	varchar	9	-1	20000	0		
9	admissions.csv	admission_location	varchar	25	-1	20000	0		
10	admissions.csv	discharge_location	varchar	25	-1	20000	0		
11	admissions.csv	insurance	varchar	10	-1	20000	0		
12	admissions.csv	language	varchar	4	-1	20000	0.42775		
13	admissions.csv	religion	varchar	22	-1	20000	0.00725		
14	admissions.csv	marital_status	varchar	17	-1	20000	0.16965		
15	admissions.csv	ethnicity	varchar	42	-1	20000	0		
16	admissions.csv	edregtime	varchar	10	-1	20000	0.47555		
17	admissions.csv	edouttime	varchar	10	-1	20000	0.47555		
18	admissions.csv	diagnosis	varchar	182	-1	20000	0.0004		
19	admissions.csv	hospital_expire_flag	int	1	-1	20000	0		
20	admissions.csv	has_chartevents_data	int	1	-1	20000	0		

Table Tabs

- A summary column for each field, there will be as many tabs as tables selected to analyze

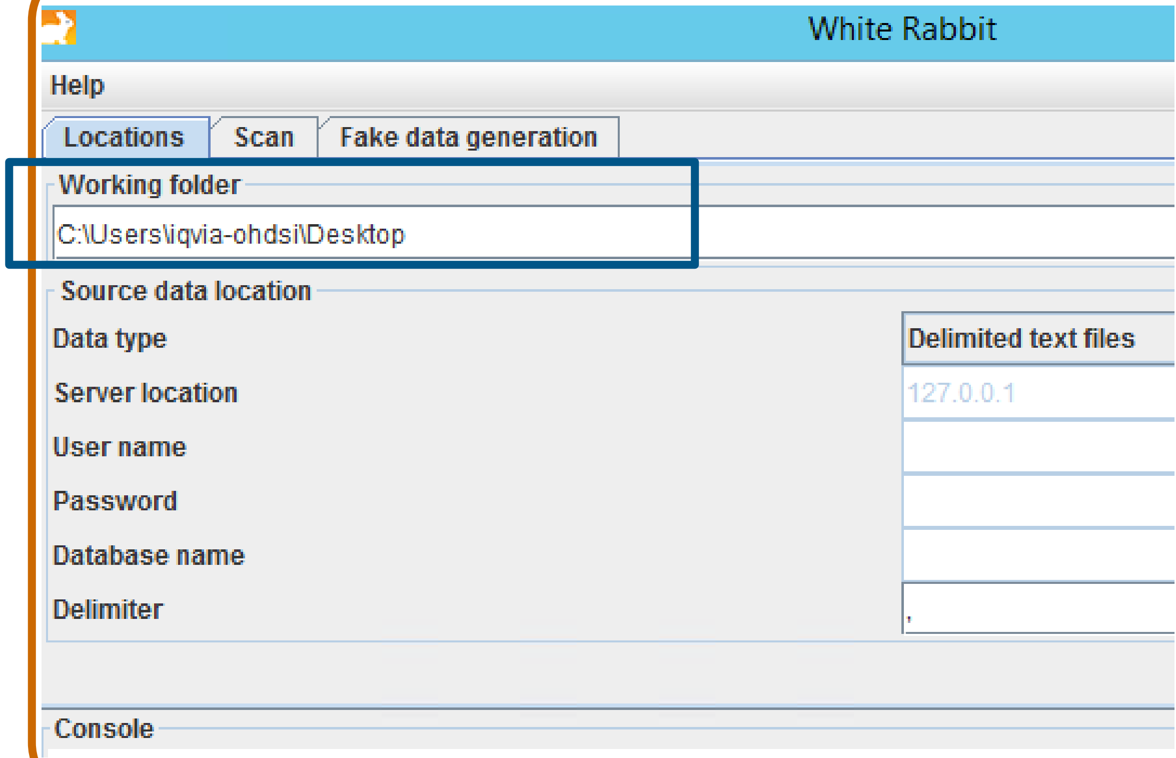


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	row_id	Frequency	subject_id	Frequency	hadm_id	Frequency	admittime	Frequency	dischtime	Frequency	deathtime	Frequency	admission_type	Frequency	admission_location
2	43575	1	11861	19	104705	1	8/14/2199	6	8/27/2179	6		18001	EMERGENCY	14326	EMERGENCY ROOM
3	42244	1	109	17	104703	1	7/28/2132	6	1/23/2133	6	2/24/2169	2	NEWBORN	2659	PHYS REFERRAL/NC
4	43576	1	13033	12	199097	1	12/17/2187	6	4/17/2195	5	7/7/2134	2	ELECTIVE	2584	CLINIC REFERRAL/P
5	43577	1	5060	11	199091	1	3/27/2136	5	10/22/2103	5	8/15/2151	2	URGENT	431	TRANSFER FROM H
6	43578	1	41976	10	199072	1	8/5/2189	5	5/12/2169	5	3/25/2140	2			TRANSFER FROM SI
7	43571	1	19620	9	199071	1	1/23/2200	5	1/23/2136	5	7/9/2151	2			** INFO NOT AVAILA
8	17284	1	25941	8	199070	1	12/11/2124	5	6/20/2113	5	12/4/2126	2			HMO REFERRAL/SIC
9	42240	1	3952	8	199077	1	9/1/2158	5	2/12/2161	5	4/12/2178	2			TRANSFER FROM O
10	43572	1	23657	8	104741	1	4/14/2115	4	10/13/2106	5	2/29/2148	2			
11	42243	1	23707	8	199075	1	3/25/2170	4	2/4/2157	5	11/27/2155	2			
12	56890	1	19029	8	116721	1	10/5/2160	4	6/27/2193	4	3/3/2122	2			
13	17283	1	76476	8	116725	1	11/16/2179	4	1/9/2158	4	4/26/2195	2			
14	43579	1	5727	8	187089	1	12/12/2164	4	2/26/2143	4	8/24/2195	2			
15	4969	1	27800	8	187095	1	4/21/2200	4	11/23/2123	4	12/1/2129	2			
16	30266	1	25225	8	103408	1	9/7/2182	4	7/16/2133	4	6/15/2173	2			
17	56896	1	20643	8	187094	1	10/8/2127	4	1/15/2183	4	5/1/2104	2			
18	3639	1	3929	7	128701	1	1/12/2180	4	6/2/2130	4	11/4/2106	2			
19	2305	1	96686	7	187090	1	1/12/2195	4	6/2/2149	4	2/12/2136	2			
20	54232	1	3100	7	199067	1	6/2/2103	4	10/21/2107	4	10/6/2139	2			

Source Data Analysis – Exercise

Exercise – Scan Mimic data

- Click on WhiteRabbit shortcut
- Select Working folder to save ScanReport
- Go to the “Scan” tab
- Press “Add” button to choose Mimic tables,
  set “Min cell count” to 0,
  set “Max distinct values” to 100,000,
  set “Rows per table” to 100,000,
  last press “Scan tabs” button



Working folder

C:\Users\iqvia-ohdsi\Desktop

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

User name

Password

Database name

Delimiter: ,

Exercise – Using White Rabbit to Scan Mimic Data

Background

Using **White Rabbit** to scan Mimic Data and answer the following questions.

Exercises

- How many patients are there in Patients table?
- How many patients do not have date of death (dod) information?
- What is the most common condition (code) among patients?
- How many admission types are there in Admission table? What are they?
- How many patients have no insurance, just “Self Pay”?
- What is the most common drug (drug) patients use?

Exercise Answers

Background

Using **White Rabbit** to scan Mimic Data and answer the following questions.

Exercises

- How many patients are there in Patients table?
  91
- How many patients do not have date of death (dod) information?
  65
- What is the most common condition (code) among patients?
  4019
- How many admission types are there in Admission table? What are they?
  4; EMERGENCY, NEWBORN, ELECTIVE, URGENT
- How many patients have no insurance, just “Self Pay”?
  192
- What is the most common drug (drug) patients use?
  D5W

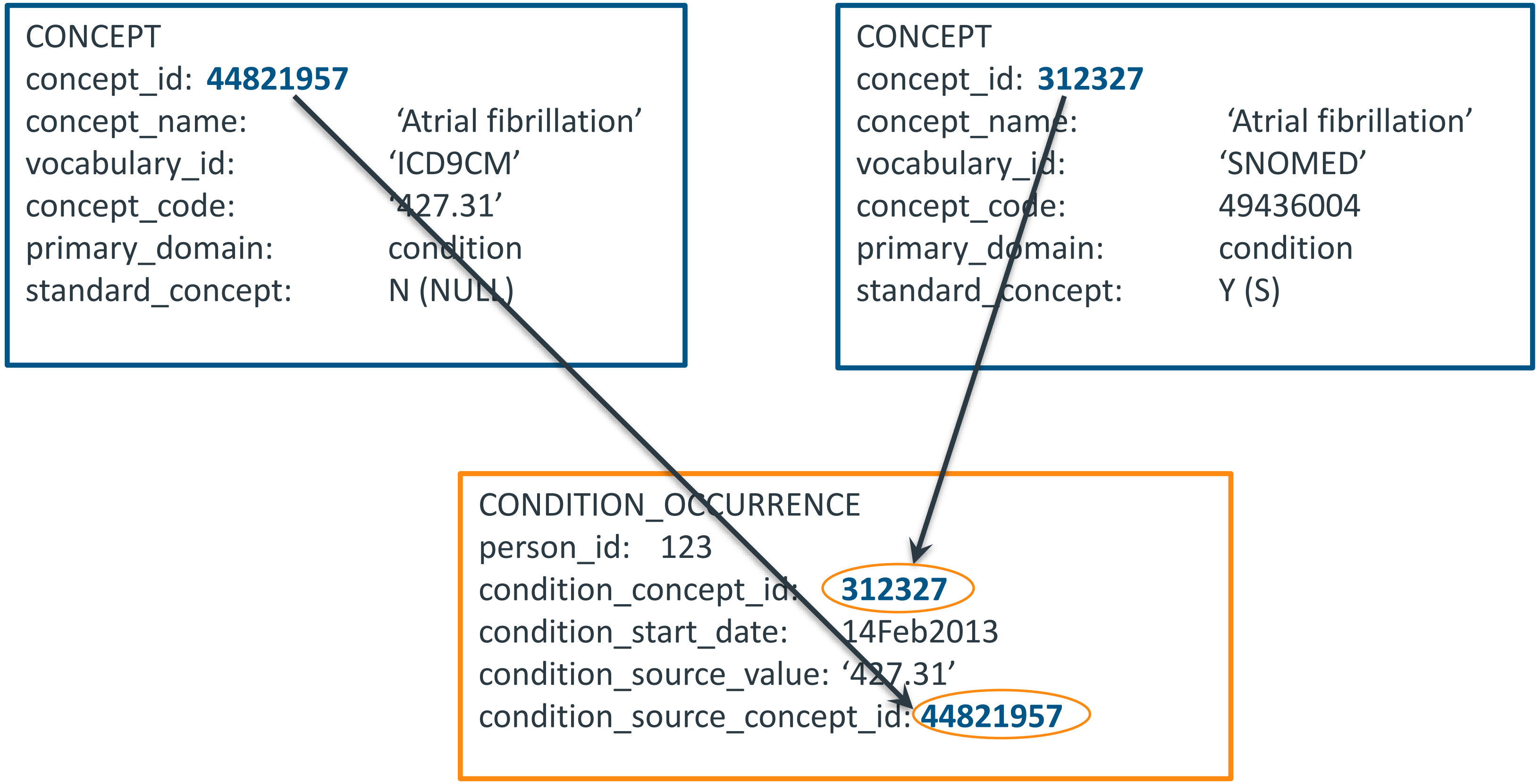
OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Break – 1 hour

Vocabulary Mapping

Integration of CDM and Vocabulary



CONCEPT
concept_id: **44821957**
concept_name: ‘Atrial fibrillation’
vocabulary_id: ‘ICD9CM’
concept_code: ‘427.31’
primary_domain: condition
standard_concept: N (NULL)

CONCEPT
concept_id: **312327**
concept_name: ‘Atrial fibrillation’
vocabulary_id: ‘SNOMED’
concept_code: 49436004
primary_domain: condition
standard_concept: Y (S)

CONDITION_OCCURRENCE
person_id: 123
condition_concept_id: **312327**
condition_start_date: 14Feb2013
condition_source_value: ‘427.31’
condition_source_concept_id: **44821957**

Mapping to Standard Concept #1

Step 1. Lookup the Source Concept

```
SELECT * FROM concept WHERE concept_code = ‘427.31 ';
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code		**427.31**

Step 2. Translate to Standard

```
SELECT * FROM concept_relationship WHERE concept_id_1 = 44821957 AND relationship_id = 'Maps to';
```

CONCEPT _ID_1	CONCEPT _ID_2	RELATIONSHIP _ID	VALID_START _DATE	VALID_END _DATE	INVALID _REASON
44821957	**313217**	Maps to	1970-01-01	2099-12-31	

Step 3. Check out the standard Concept

```
SELECT * FROM concept WHERE concept_id = 313217;
```

Determines place in CDM

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	**S**	49436004

Mapping to Standard Concept #2

Step 1. Lookup the Source Concept

SELECT * **FROM** concept **WHERE** concept_code = '67544050474';

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
45867731	clopidogrel 75 MG Oral Tablet [Plavix]	Drug	NDC	11-digit NDC		**67544050474**

Step 2. Translate to Standard

SELECT * **FROM** concept_relationship **WHERE** concept_id_1 = 45867731 **AND** relationship_id = **'Maps to';**

CONCEPT _ID_1	CONCEPT _ID_2	RELATIONSHIP _ID	VALID_START _DATE	VALID_END _DATE	INVALID _REASON
45867731	**1322185**	Maps to	2015-01-29	2099-12-31	

Step 3. Check out the standard Concept

SELECT * **FROM** concept **WHERE** concept_id = 1322185;

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
1322185	clopidogrel 75 MG Oral Tablet [Plavix]	Drug	RxNorm	Branded Drug	**S**	213169

Exercise – Write SQL Query to Find Standard Concept

Write the SQL query to find the standard concept for this source code: R26.2

Hint:

- This is an ICD10 code
- It belongs to Condition domain
- Use Concept table to find source_concept_id
- Use Concept_relationship table and ‘Maps to’ relationship_id to find standard concept_id

Answer to the Exercise

Step 1. Lookup the Source Concept

```
SELECT * FROM concept WHERE concept_code = ‘R26.2 ';
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
45602016	Difficulty in walking, not elsewhere classified	Condition	ICD10	ICD10 code		**R26.2**

Step 2. Translate to Standard

```
SELECT * FROM concept_relationship WHERE concept_id_1 = 45602016 AND relationship_id = 'Maps to';
```

CONCEPT _ID_1	CONCEPT _ID_2	RELATIONSHIP _ID	VALID_START _DATE	VALID_END _DATE	INVALID _REASON
45602016	**36714126**	Maps to	2018-11-28	2099-12-31	

Step 3. Check out the standard Concept

```
SELECT * FROM concept WHERE concept_id = 36714126;
```

CONCEPT _ID	CONCEPT_ NAME	DOMAIN _ID	VOCABULARY _ID	CONCEPT_ CLASS_ID	STANDARD_ CONCEPT	CONCEPT_ CODE
36714126	Difficulty walking	Condition	SNOMED	Clinical Finding	**S**	719232003

One source field can go to multiple CDM domains

This is an example showing source Diagnosis table (diagnosis_code) can be mapped to different domains

diagnosis_code (ICD10)	diagnosis_description
I48.2	Chronic atrial fibrillation
Z31.5	Genetic counseling
Z82.3	Family history of stroke
R71	Abnormality of red blood cells

concept_id (standard)	concept_name (standard)	domain_id
4141360	Chronic atrial fibrillation	**Condition**
4196362	Genetic counseling	**Procedure**
4169009	Family history of stroke	**Observation**
4098353	Red blood cell test	**Measurement**

Exercise – Find out Domains for Following Codes

Find out the destination table (domain) for following diagnosis data:

diagnosis_code (ICD10)	diagnosis_description
R10.0	Acute abdomen
Z01.1	Examination of ears and hearing
Z85.6	Personal history of leukaemia
R77.0	Abnormality of albumin

Answer to Exercise

diagnosis_code (ICD10)	diagnosis_description
R10.0	Acute abdomen
Z01.1	Examination of ears and hearing
Z85.6	Personal history of leukaemia
R77.0	Abnormality of albumin

concept_id (standard)	concept_name (standard)	domain_id
4241033	Acute abdomen	**Condition**
4134565	Hearing examination	**Procedure**
4058706	History of leukemia	**Observation**
4097664	Albumin measurement	**Measurement**

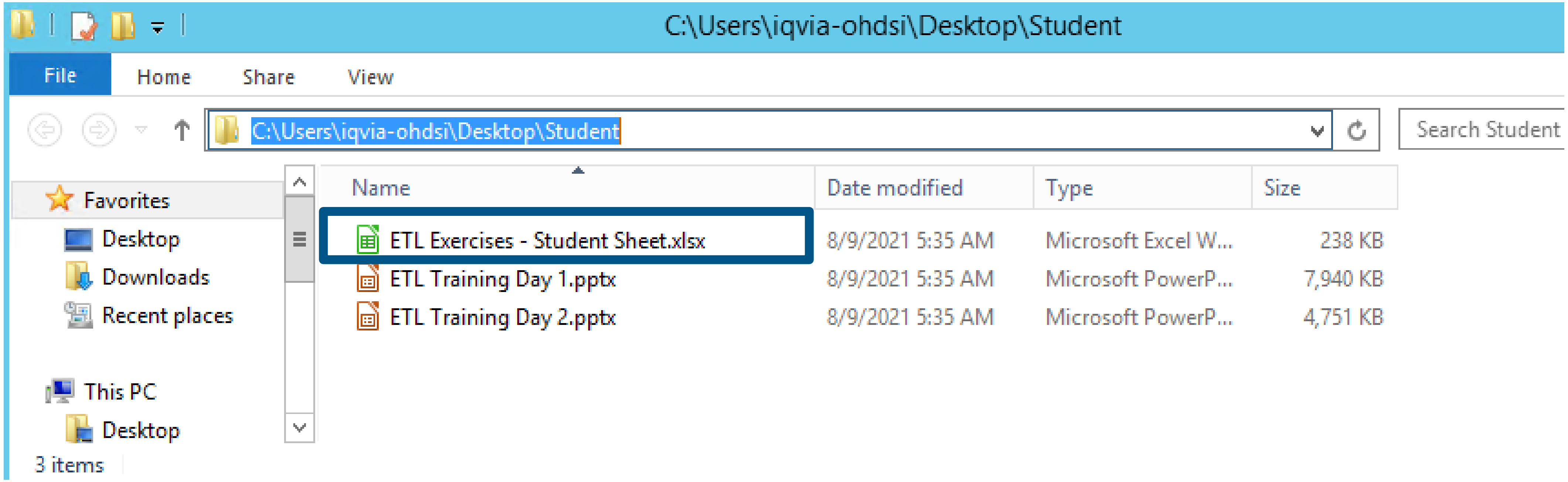
Vocabulary Mapping – Exercise

Vocabulary Mapping Exercise

- On the Box, go to folder ‘C:\Users\iqvia-ohdsi\Desktop\Student’

- Open file ‘ETL Exercises - Student Sheet’

- Do exercise in **Day1 Vocabulary Mapping** tab



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

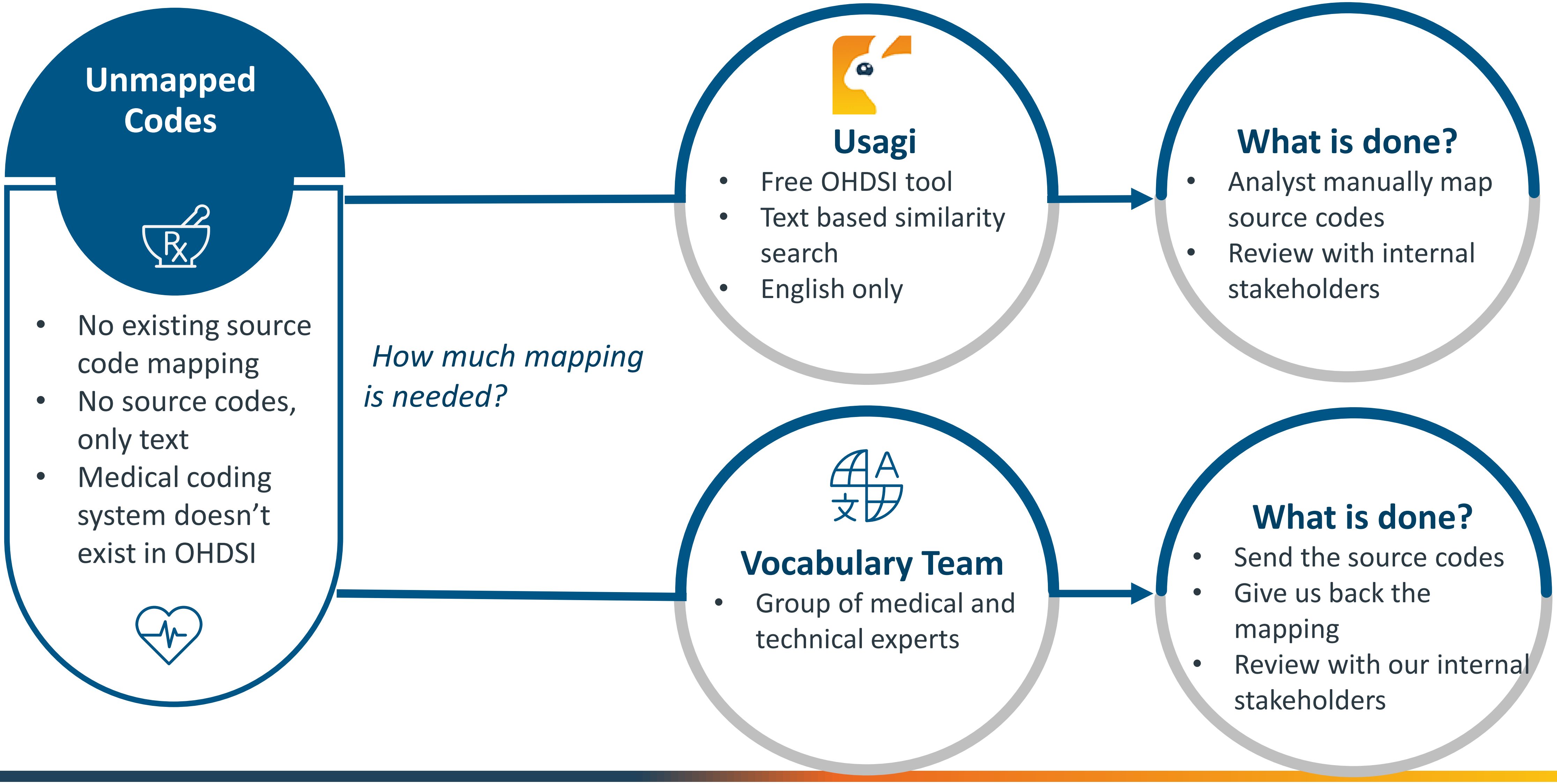
C:\Users\iqvia-ohdsi\Desktop\Student

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

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Custom Mapping of Unmapped Codes Using Usagi

Custom source code mapping



Unmapped Codes

- No existing source code mapping
- No source codes, only text
- Medical coding system doesn't exist in OHDSI

How much mapping is needed?

Usagi

- Free OHDSI tool
- Text based similarity search
- English only

What is done?

- Analyst manually map source codes
- Review with internal stakeholders

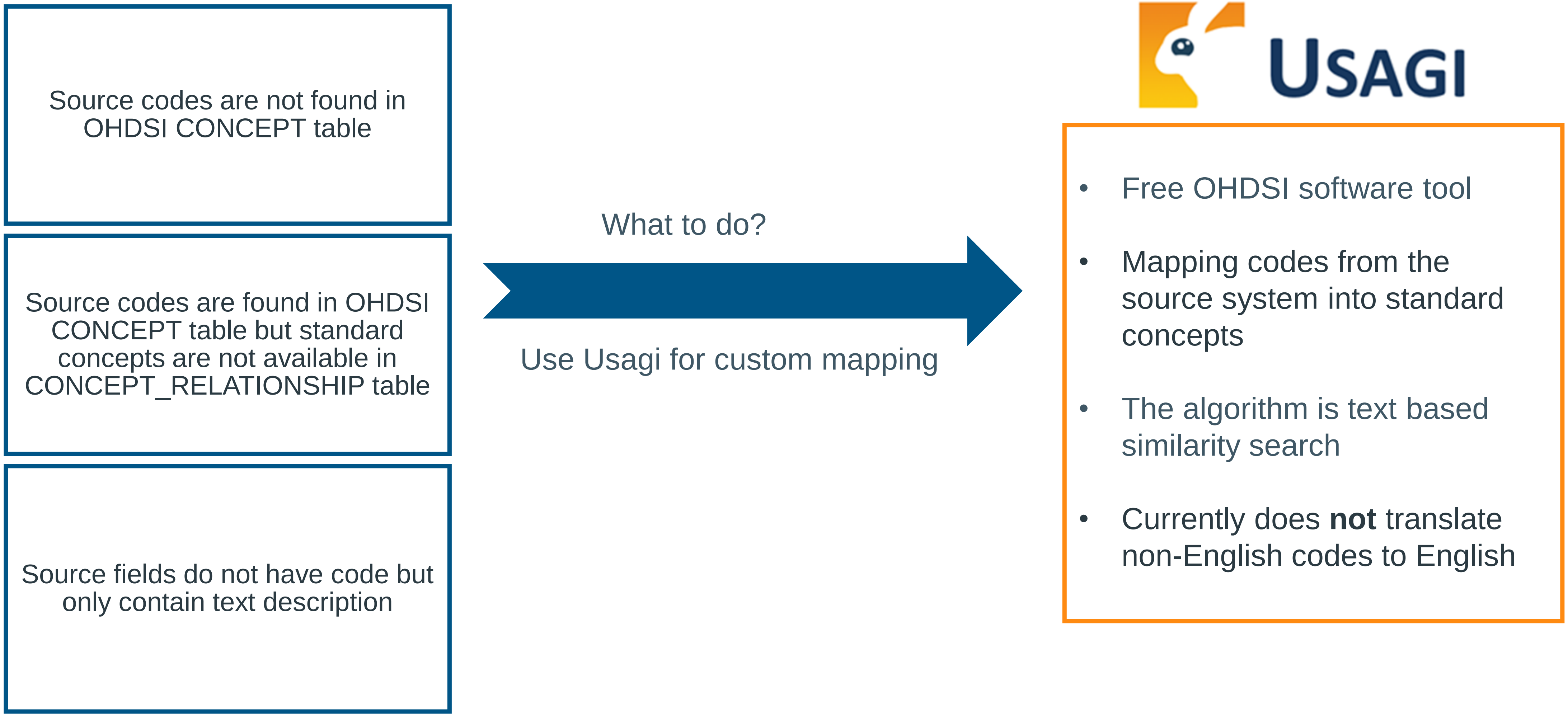
Vocabulary Team

- Group of medical and technical experts

What is done?

- Send the source codes
- Give us back the mapping
- Review with our internal stakeholders

Purpose of Usagi



What are unmapped codes?

Source codes are not found in OHDSI CONCEPT table

Source codes are found in OHDSI CONCEPT table but standard concepts are not available in CONCEPT_RELATIONSHIP table

Source fields do not have code but only contain text description

What to do?

Use Usagi for custom mapping

USAGI

- Free OHDSI software tool
- Mapping codes from the source system into standard concepts
- The algorithm is text based similarity search
- Currently does **not** translate non-English codes to English

Difficulties of custom mapping

Requires medical expertise

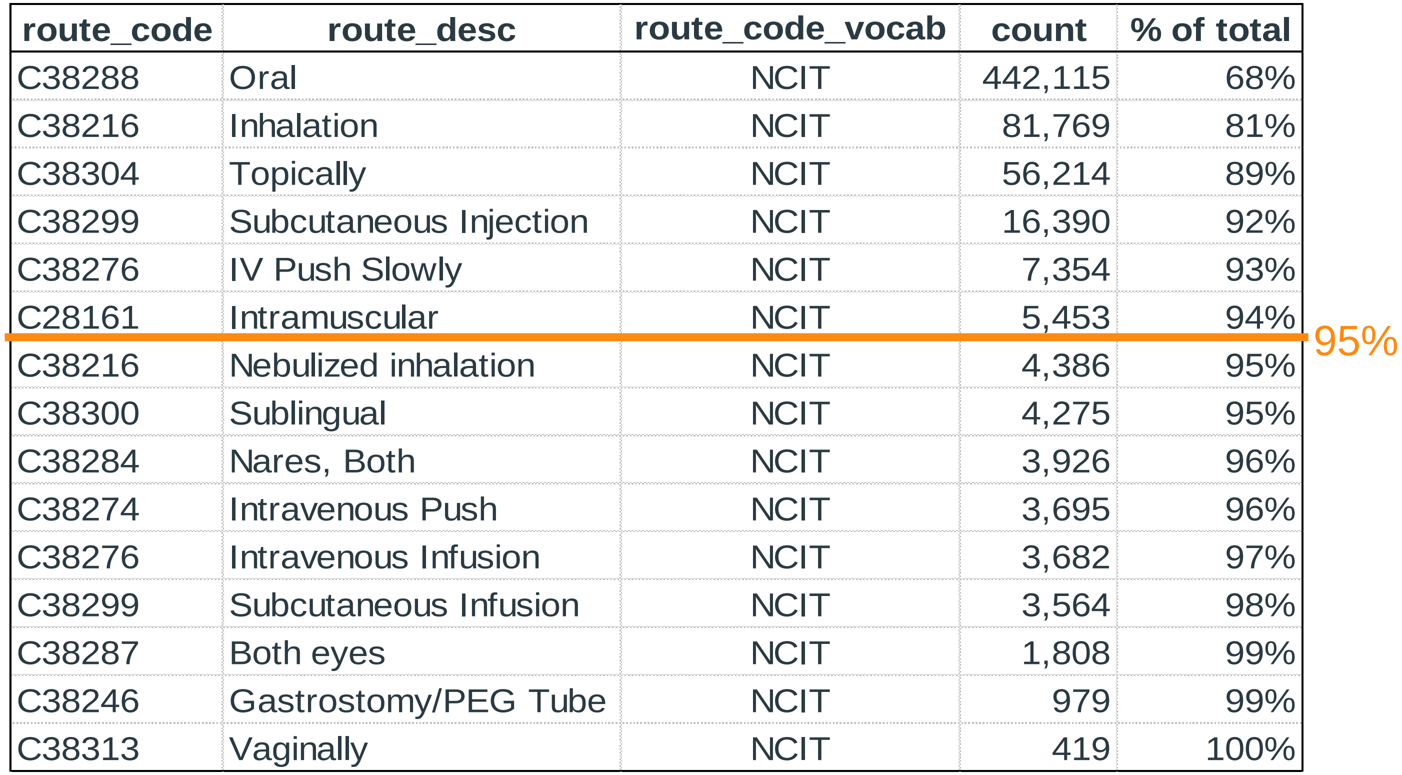
Non-English descriptions

Time consuming

- No capacity to custom map thousands of codes
- Instead focus on most frequent (95%)

Requires updating

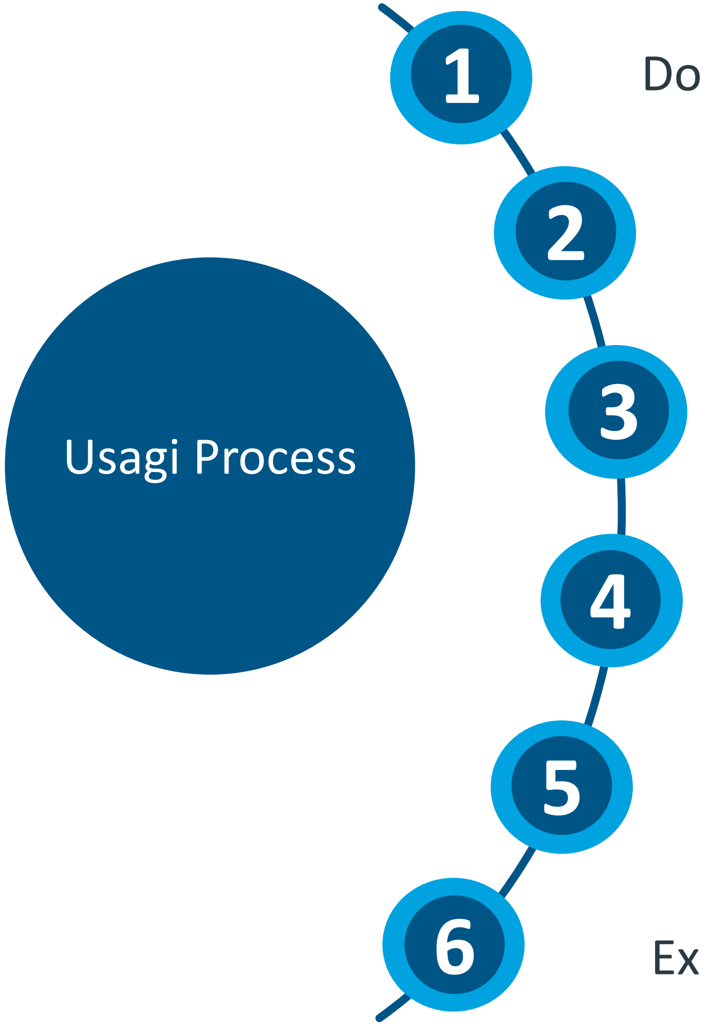
- A need to revisit custom mapping
- New codes added
- Old standard concepts become invalid



route_code	route_desc	route_code_vocab	count	% of total
C38288	Oral	NCIT	442,115	68%
C38216	Inhalation	NCIT	81,769	81%
C38304	Topically	NCIT	56,214	89%
C38299	Subcutaneous Injection	NCIT	16,390	92%
C38276	IV Push Slowly	NCIT	7,354	93%
C28161	Intramuscular	NCIT	5,453	94%
C38216	Nebulized inhalation	NCIT	4,386	95%
C38300	Sublingual	NCIT	4,275	95%
C38284	Nares, Both	NCIT	3,926	96%
C38274	Intravenous Push	NCIT	3,695	96%
C38276	Intravenous Infusion	NCIT	3,682	97%
C38299	Subcutaneous Infusion	NCIT	3,564	98%
C38287	Both eyes	NCIT	1,808	99%
C38246	Gastrostomy/PEG Tube	NCIT	979	99%
C38313	Vaginally	NCIT	419	100%

95%

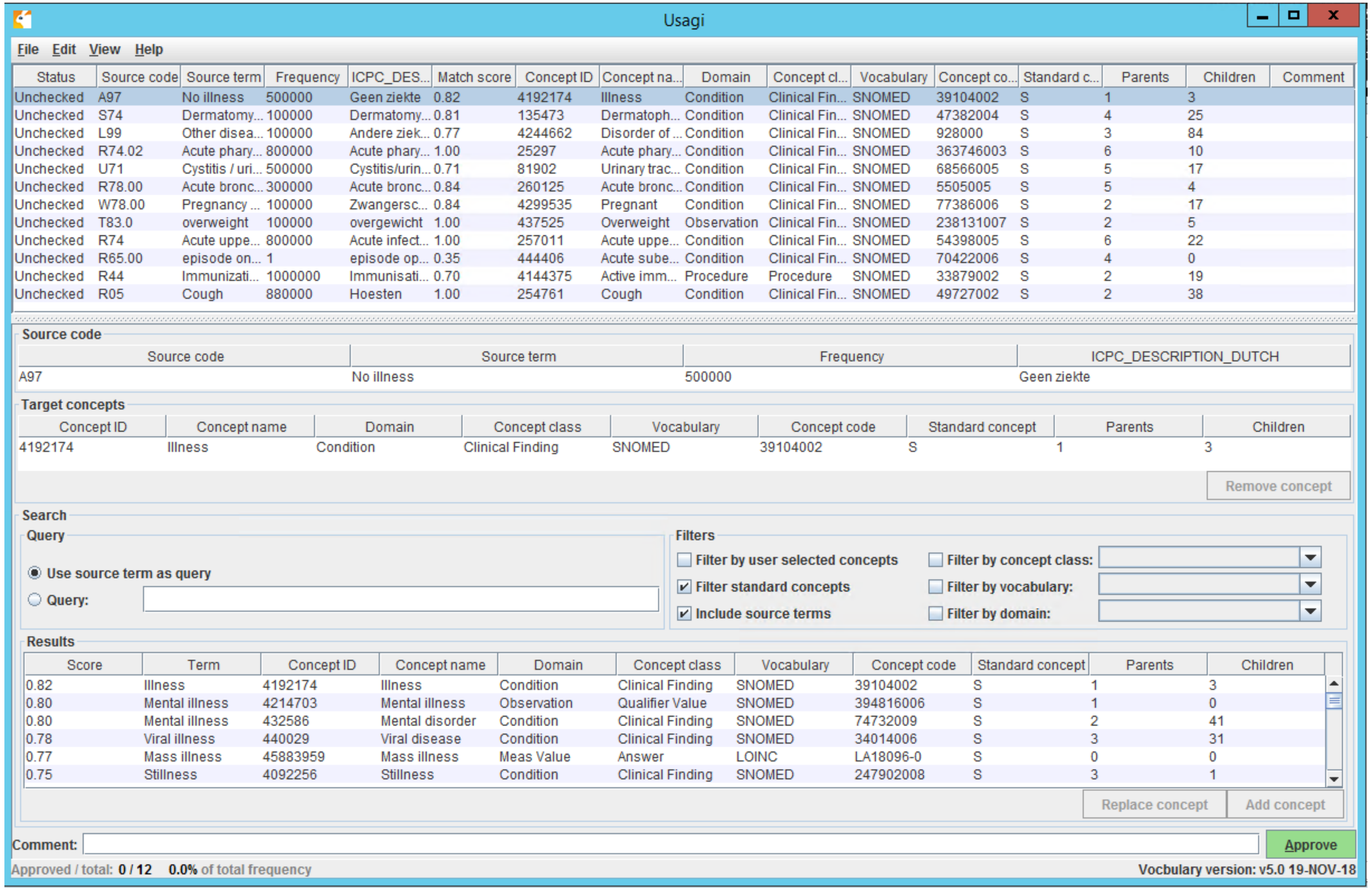
Usagi Process Overview



Usagi Process

1. Download Usagi – https://github.com/OHDSI/Usagi
2. Get a copy of the Vocabulary from ATHENA – http://athena.ohdsi.org
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 medical experts
6. Export codes into the SOURCE_TO_CONCEPT_MAP

Usagi Demo



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

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Break
– 15min

ETL Specification Writing

What is an ETL Specification

Document created by analysts

Roadmap for the development team

Used during QA process

- Cooperate with Data Owner

- Tells exactly which fields to map into the OMOP model
- Applies rules to the data
- Specifies what records to deduplicate or filter out completely

- Cross reference ETL Spec to ensure rules were applied

Creating ETL Specification

1 Analyze Data

- Review the source data table by table, field by field
- Study the data dictionary
- Study any other supporting documents

2 Work with Data Owners

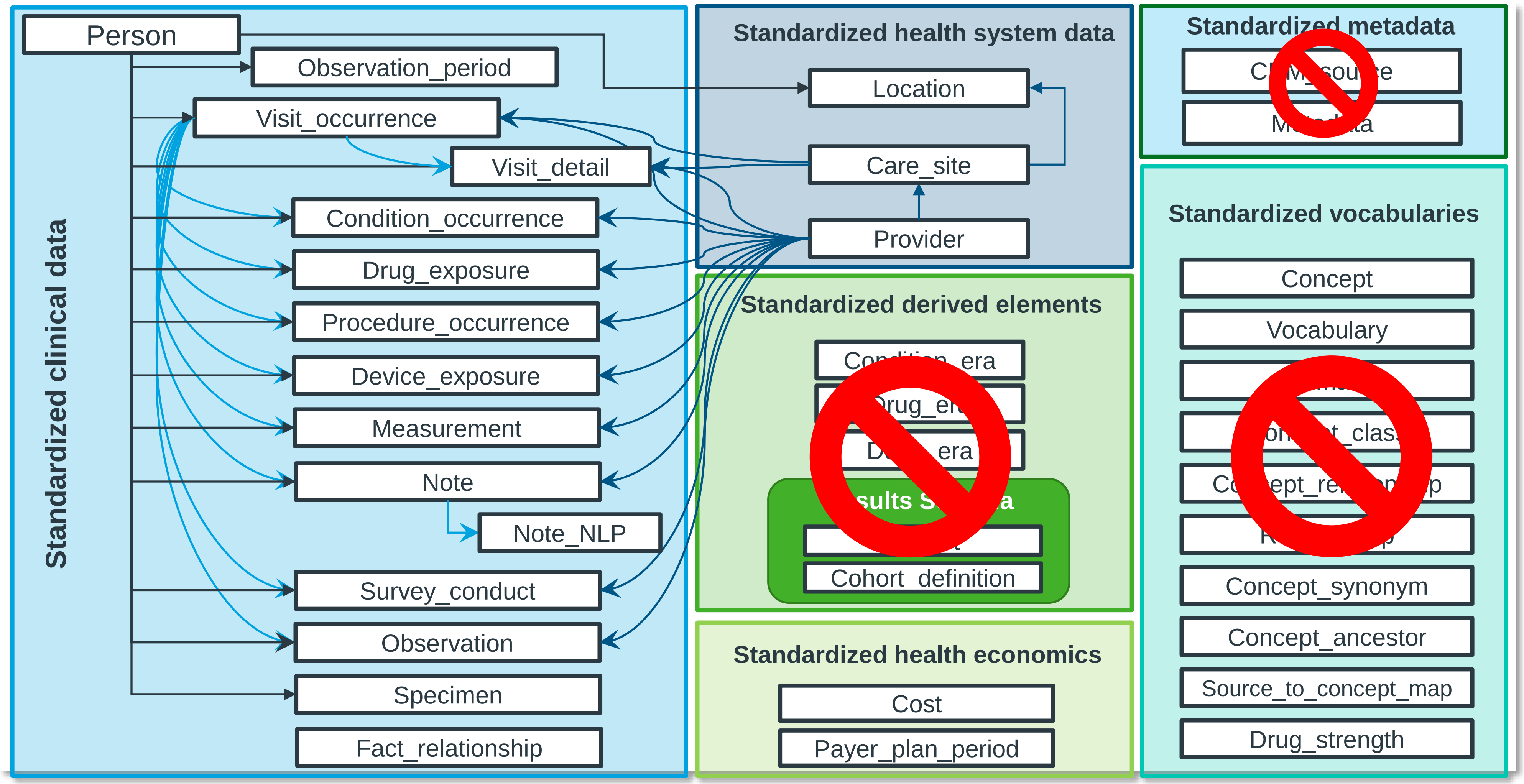
- Confirm your understanding of the data
- Ask questions on things that are not clear

3 Continued Project Review

- Review with team
- Review with data owners

Destination Field	Source Field	Applied Rule
Person_Id		System generated id based on unique source identifier
Gender_concept_id	Bene_sex_ident_cd	If 1 then '8507' If 2 then '8532' All else/unknown = 0
Year_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in 'YYYY'. Exclude patients with NULL or invalid year of birth
Month_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in 'MM'.
Day_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in 'DD'.

Tables in ETL Specification



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

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

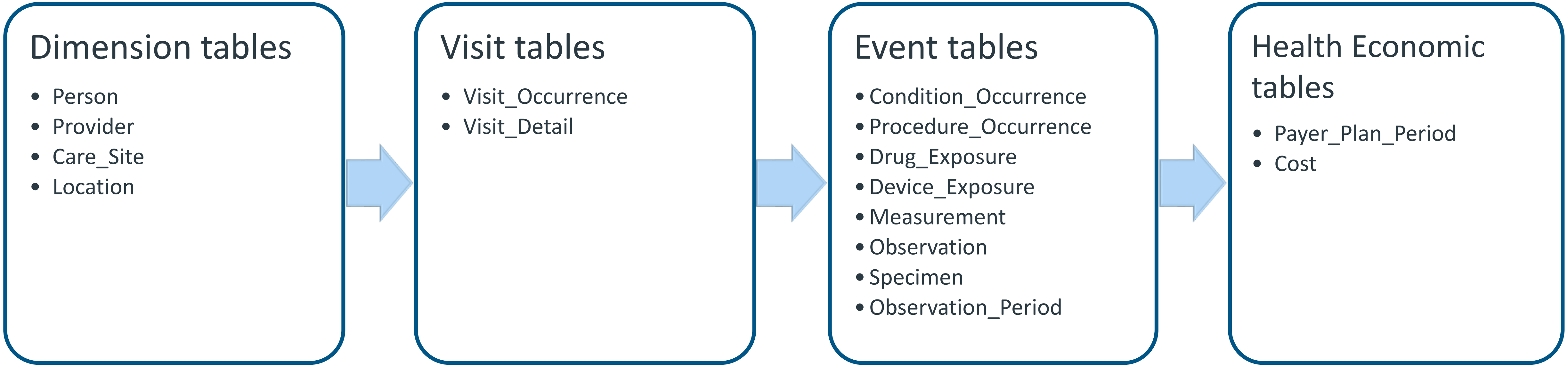
Concept_synonym

Concept_ancestor

Source_to_concept_map

Drug_strength

ETL Spec Table Writing Sequence (Recommended)



Dimension tables

- Person
- Provider
- Care_Site
- Location

Visit tables

- Visit_Occurrence
- Visit_Detail

Event tables

- Condition_Occurrence
- Procedure_Occurrence
- Drug_Exposure
- Device_Exposure
- Measurement
- Observation
- Specimen
- Observation_Period

Health Economic tables

- Payer_Plan_Period
- Cost

ETL Spec Content – Common Data Elements to All Event Tables

Clinical event tables

- Condition_Occurrence
- Procedure_Occurrence
- Drug_Exposure
- Device_Exposure
- Measurement
- Observation
- Specimen

Common primary key and foreign key columns in clinical event tables

Field name	Purpose and example
<entity>_id	Primary key for the entity
Person_id	Foreign key to the Person table
Provider_id	Foreign key to the Provider table
Visit_occurrence_id	Foreign key to the Visit_occurrence table

Common vocabulary related columns in clinical event tables

Field name	Purpose and example
<entity>_concept_id	**Standard** OMOP concept_id for source value condition_concept_id 4068155 (SNOMED "Atrial arrhythmia")
<entity>_source_concept_id	OMOP concept_id for source value condition_source_concept_id 45596206 (ICD10 "Atrial fibrillation and flutter")
<entity>_source_value	Verbatim information from the source data, **not to be used** by any standard analytics condition_source_value I48 (ICD10 "Atrial fibrillation and flutter")
<entity>_type_concept_id	OMOP concept_id for the **origin of the information** condition_type_concept_id 32817 (“EHR") Domain = ‘Type Concept’, Concept = ‘Standard’ in ATHENA

ETL Spec – Written in a Template

Destination Field	Source Field	Applied Rule	Comment
Person_id			
Gender_concept_id			
Year_of_birth			
Month_of_birth			
Day_of_birth			

- Destination Field = OMOP field being referenced
- Source Field = field from source data that will be mapped into the Destination Field
- Applied Rule = any rules that are being applied to the data as it is mapped in
- Comment = additional notes that are relevant

ETL Spec – Written in a Template

Destination Field	Source Field	Applied Rule	Comment
Person_Id		System generated id based on desynpuf_id	
Gender_concept_id	Bene_sex_ident_cd	If 1 then ‘8507’ If 2 then ‘8532’ All else/unknown = 0	8507 is Male 8532 is Female
Year_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘YYYY’. Exclude patients with NULL or invalid year of birth	
Month_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘MM’.	
Day_of_birth	Bene_birth_dt	Format is YYYY-MM-DD. Map in ‘DD’.	

ETL Spec – Source and Target Tables Relationship

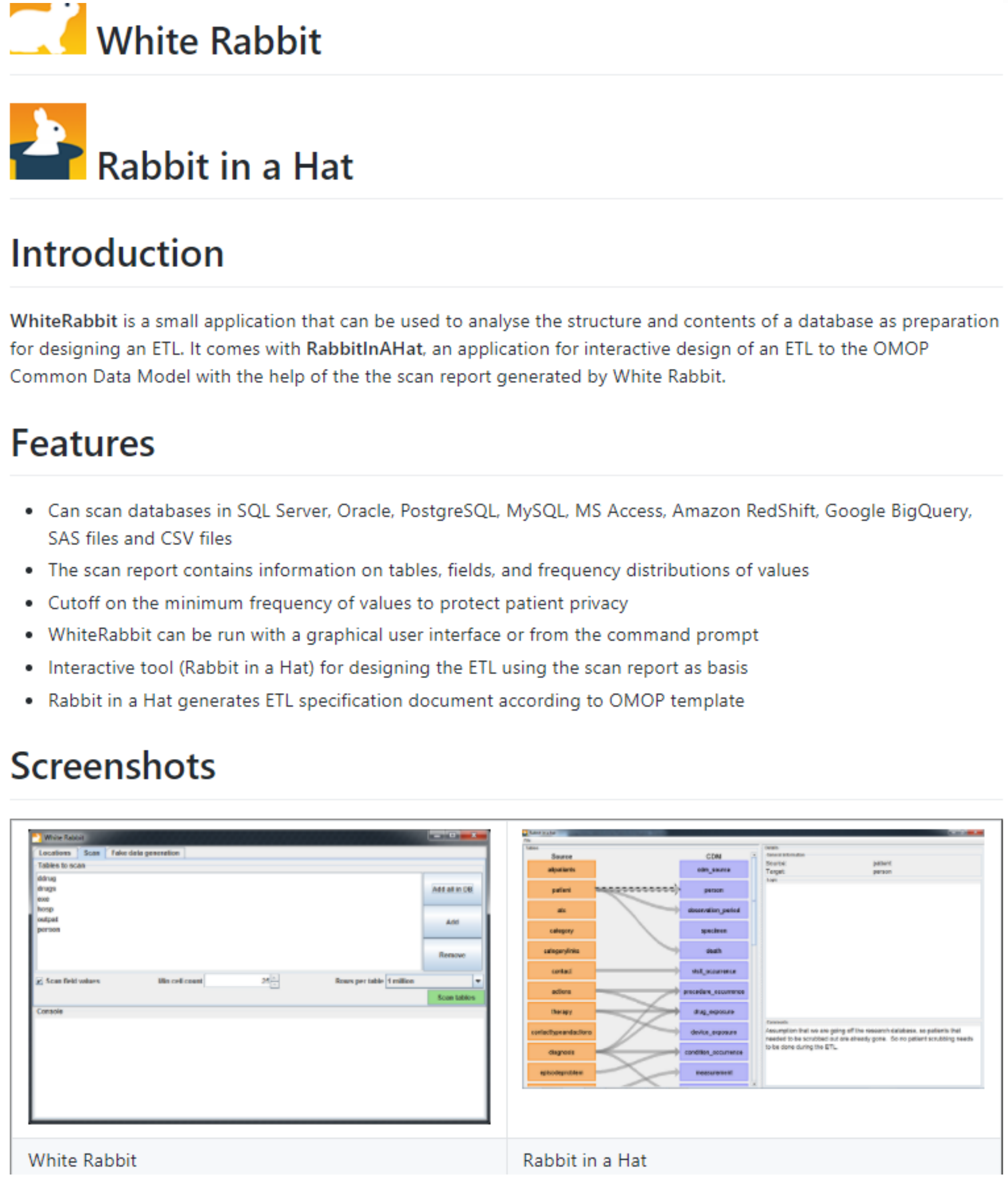
- Multiple source tables can be mapped to the same OMOP CDM table
- Multiple fields within one source table can be mapped to the same OMOP CDM table
- Example: If a table has three fields which hold an ICD10 code, these three fields can all be used to create three different records in omop

Destination Field	Source Field	Applied Rule	Comment
Condition_occurrence_id	A unique, system generated identifier		
Person_id	Cdm.person_id		
Condition_concept_id	icd10_dgns_cd_1 OR icd10_dgns_cd_2 OR icd10_dgns_cd_3	Create one condition occurrence record for each ICD10 diagnosis code on source record	
Condition_start_date	Clm_from_dt		
Condition_start_datetime	NULL		Information is not available in the source data

Writing ETL Spec with Rabbit in a Hat

Rabbit in a Hat

- Is also part of the White Rabbit Download https://github.com/OHDSI/WhiteRabbit
- Allows users to map source fields in OMOP fields
- Can read and display a White Rabbit scan document
- Provides a graphical interface to allow a user to connect source data to tables
- Generates ETL Spec document, does not generate code



White Rabbit

Rabbit in a Hat

Introduction

WhiteRabbit is a small application that can be used to analyse the structure and contents of a database as preparation for designing an ETL. It comes with **RabbitInAHat**, an application for interactive design of an ETL to the OMOP Common Data Model with the help of the the scan report generated by White Rabbit.

Features

- Can scan databases in SQL Server, Oracle, PostgreSQL, MySQL, MS Access, Amazon RedShift, Google BigQuery, SAS files and CSV files
- The scan report contains information on tables, fields, and frequency distributions of values
- Cutoff on the minimum frequency of values to protect patient privacy
- WhiteRabbit can be run with a graphical user interface or from the command prompt
- Interactive tool (Rabbit in a Hat) for designing the ETL using the scan report as basis
- Rabbit in a Hat generates ETL specification document according to OMOP template

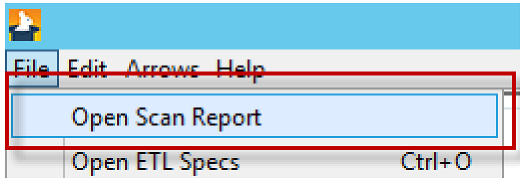
Screenshots

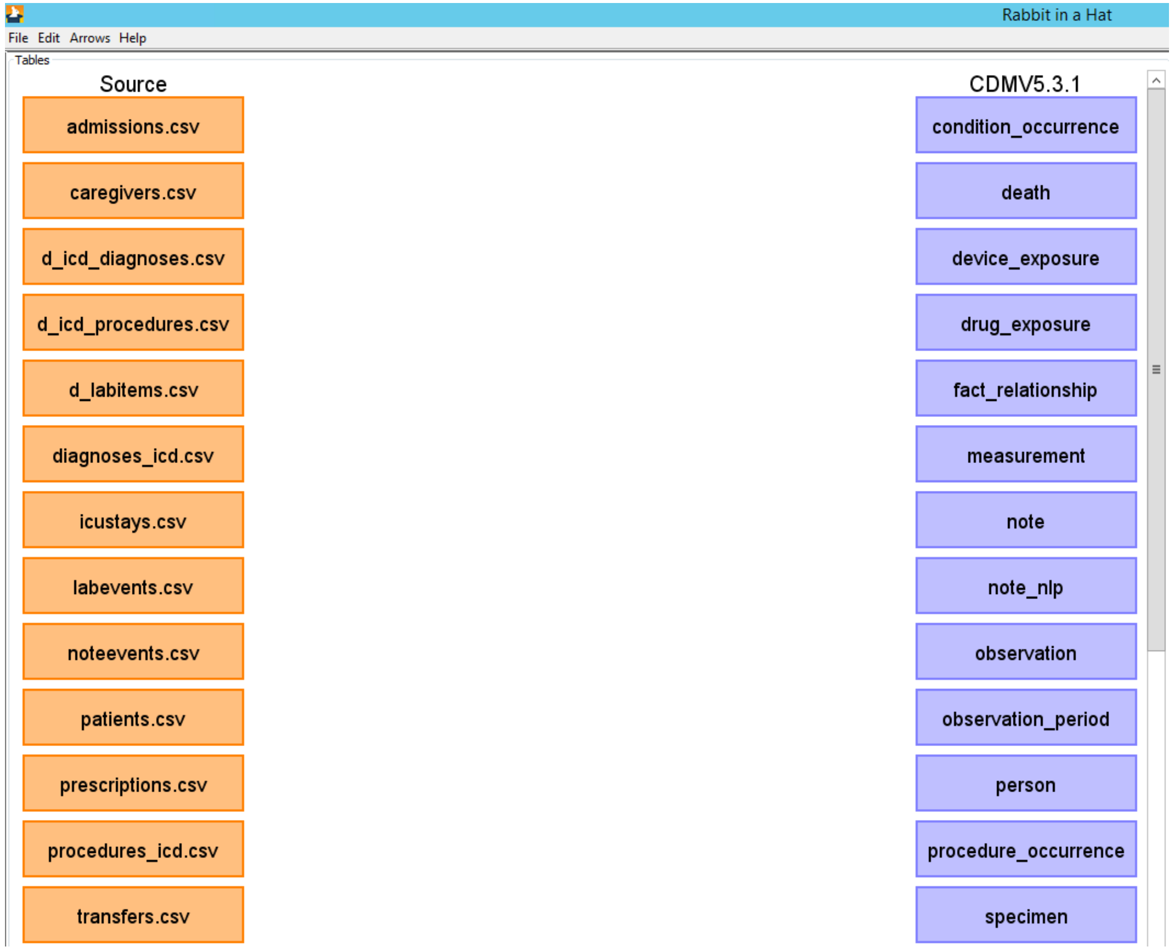
White Rabbit

Rabbit in a Hat

Rabbit in a Hat – Start

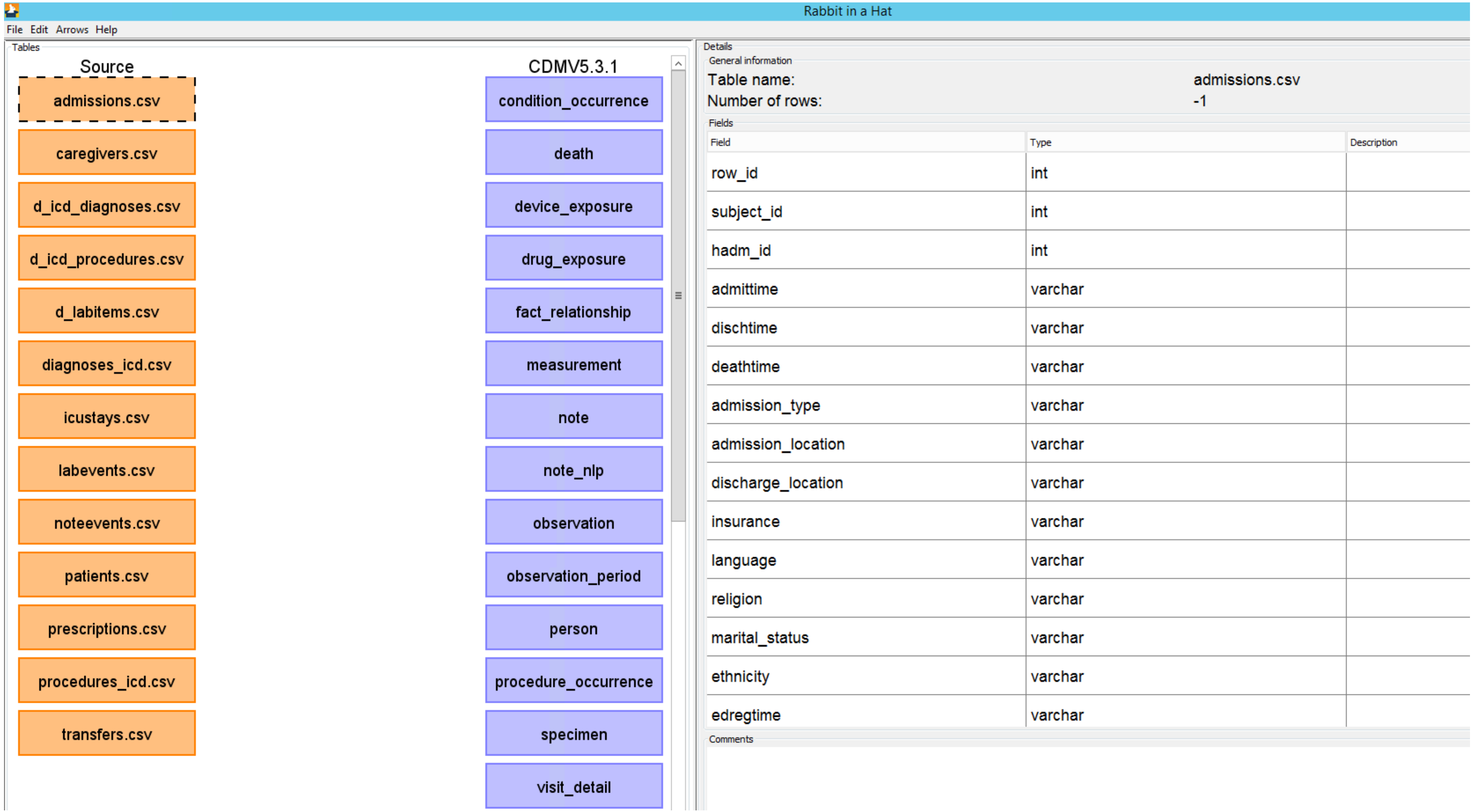
- Double click on Rabbit in a Hat from it’s stored location to start the application
- Select File, Open Scan Report. Use the Scan Report we recently generated with White Rabbit





Rabbit in a Hat – Learn source tables

- Select source table to learn more about the fields.



Source	CDMV5.3.1
admissions.csv	condition_occurrence
caregivers.csv	death
d_icd_diagnoses.csv	device_exposure
d_icd_procedures.csv	drug_exposure
d_labitems.csv	fact_relationship
diagnoses_icd.csv	measurement
icustays.csv	note
labevents.csv	note_nlp
noteevents.csv	observation
patients.csv	observation_period
prescriptions.csv	person
procedures_icd.csv	procedure_occurrence
transfers.csv	specimen
	visit_detail

General information

Table name:	admissions.csv
Number of rows:	-1

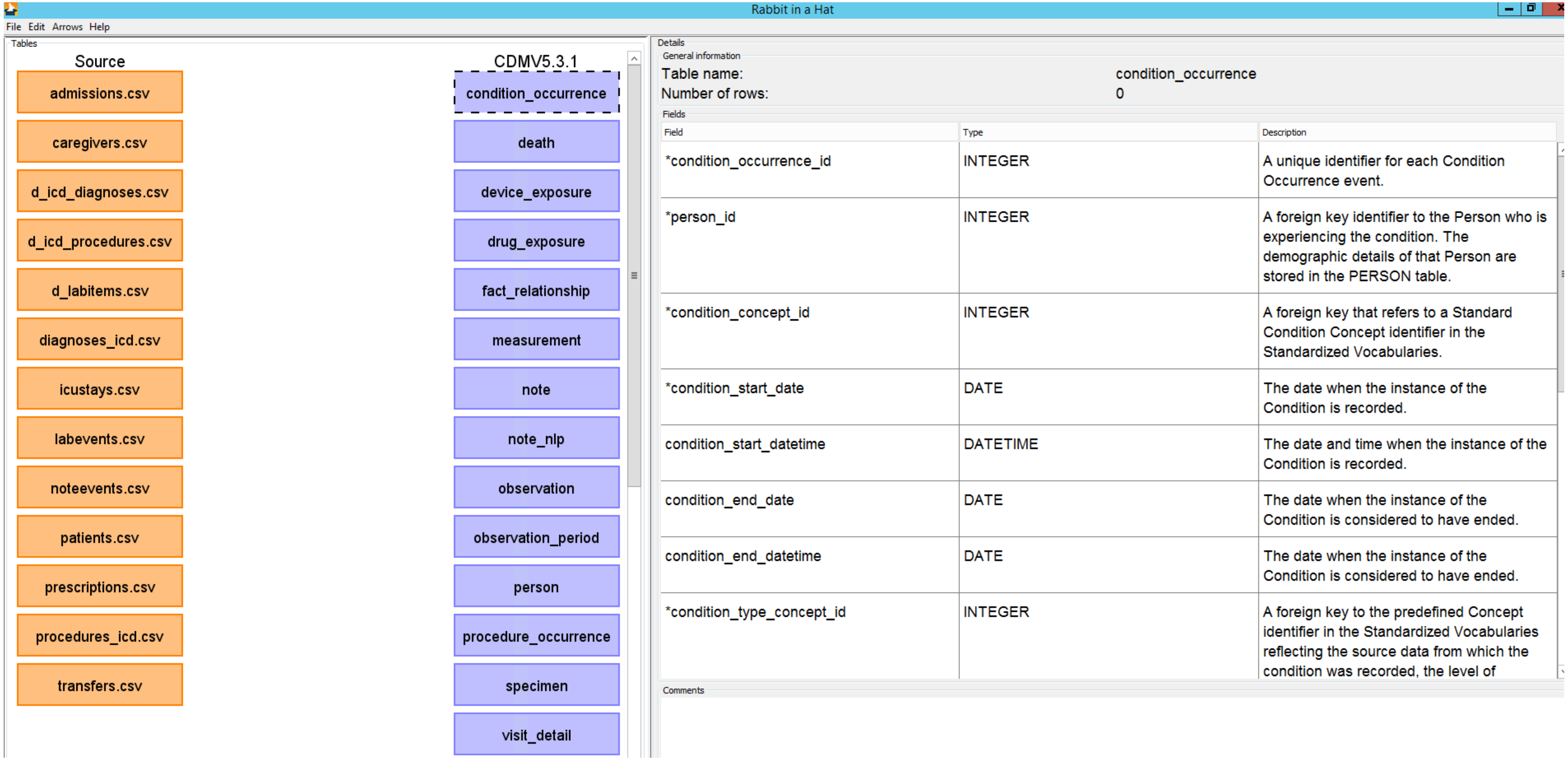
Fields

Field	Type	Description
row_id	int	
subject_id	int	
hadm_id	int	
admittime	varchar	
dischtime	varchar	
deathtime	varchar	
admission_type	varchar	
admission_location	varchar	
discharge_location	varchar	
insurance	varchar	
language	varchar	
religion	varchar	
marital_status	varchar	
ethnicity	varchar	
edregtime	varchar	

Comments

Rabbit in a Hat – Learn CDM tables

- Select OMOP table to learn more about the fields.



Source	CDMV5.3.1
admissions.csv	condition_occurrence
caregivers.csv	death
d_icd_diagnoses.csv	device_exposure
d_icd_procedures.csv	drug_exposure
d_labitems.csv	fact_relationship
diagnoses_icd.csv	measurement
icustays.csv	note
labevents.csv	note_nlp
noteevents.csv	observation
patients.csv	observation_period
prescriptions.csv	person
procedures_icd.csv	procedure_occurrence
transfers.csv	specimen
	visit_detail

Table name: condition_occurrence
Number of rows: 0

Field	Type	Description
*condition_occurrence_id	INTEGER	A unique identifier for each Condition Occurrence event.
*person_id	INTEGER	A foreign key identifier to the Person who is experiencing the condition. The demographic details of that Person are stored in the PERSON table.
*condition_concept_id	INTEGER	A foreign key that refers to a Standard Condition Concept identifier in the Standardized Vocabularies.
*condition_start_date	DATE	The date when the instance of the Condition is recorded.
condition_start_datetime	DATETIME	The date and time when the instance of the Condition is recorded.
condition_end_date	DATE	The date when the instance of the Condition is considered to have ended.
condition_end_datetime	DATE	The date when the instance of the Condition is considered to have ended.
*condition_type_concept_id	INTEGER	A foreign key to the predefined Concept identifier in the Standardized Vocabularies reflecting the source data from which the condition was recorded, the level of

ETL Specification Writing – Exercise & Homework

Exercise & Homework

- Map the Mimic data using Rabbit in a Hat

RabbitInAHat
.jar - Shortcut

- Data Dictionary can be found here:
  http://pi.cs.oswego.edu/~jmiles3/mimic/Miles-MIMIC-Project_report.pdf

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Thank You!