



ETL Conversion Training

UC Systems Training – Session 4b

OMOP Team, IQVIA
March. 2021



Agenda

Time (03/26)

Section

1:00 – 1:15 PM	Introduction to ETL + Agile Methodology
1:15 – 1:35 PM	Source Data Analysis
1:35 – 2:00 PM	Exercise
2:00 – 2:20 PM	Vocabulary Mapping
2:20 – 2:35 PM	Exercise
2:35 – 2:50 PM	ETL Workflow
2:50 – 3:10 PM	Data Quality
3:10 – 3:30 PM	CDM Challenges
3:30 – 4:15 PM	Open Discussions



Introduction to ETL

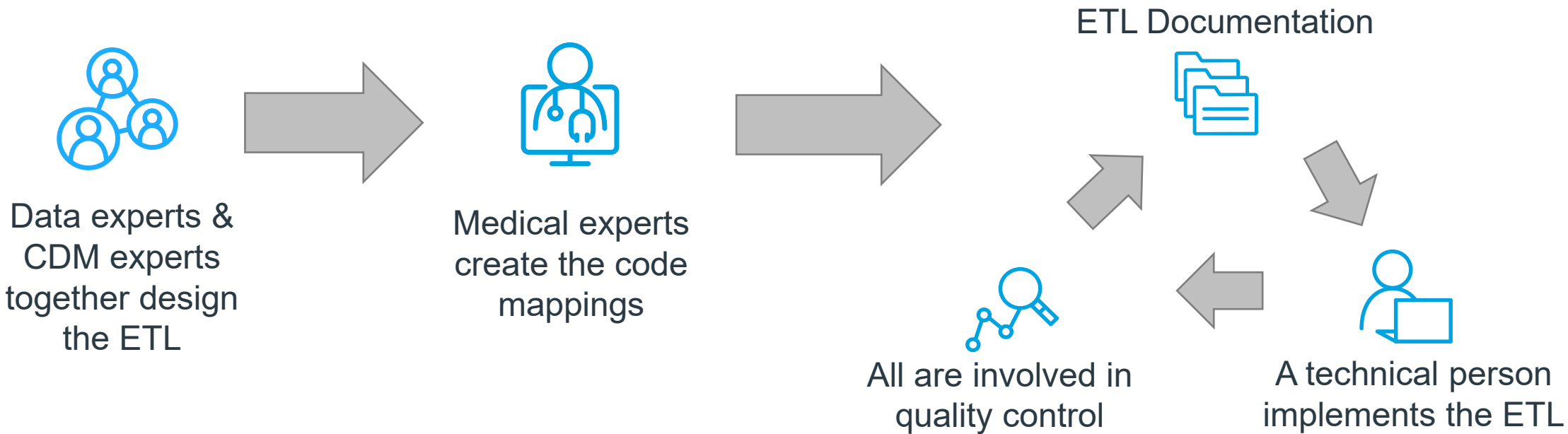
ETL

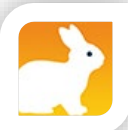
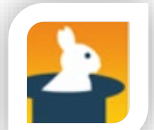






- Extract, Transform, Load 抽取、转换、加载
- In order to get from our native/raw data into the OMOP CDM we need to design and develop an ETL process



- Goal in ETLing is to standardize the format and terminology 标准化数据格式和术语
- This tutorial
 - Will teach you best practices around designing an ETL and CDM maintenance
 - Will not teach you how to program an ETL

OMOP conversion process flow



Tools	Analysis 源数据分析			Quality Control 数据质量监控			Development 数据开发	
	 White Rabbit	 Rabbit In a Hat	 Usagi	 Internal Quality Checks	 Achilles	 Data Quality Dashboard	 Jenkins	 Code Repository

Agile Methodology

What is Agile Scrum

1

Software development methodology



2

Iterative approach



3

Evolves through collaboration



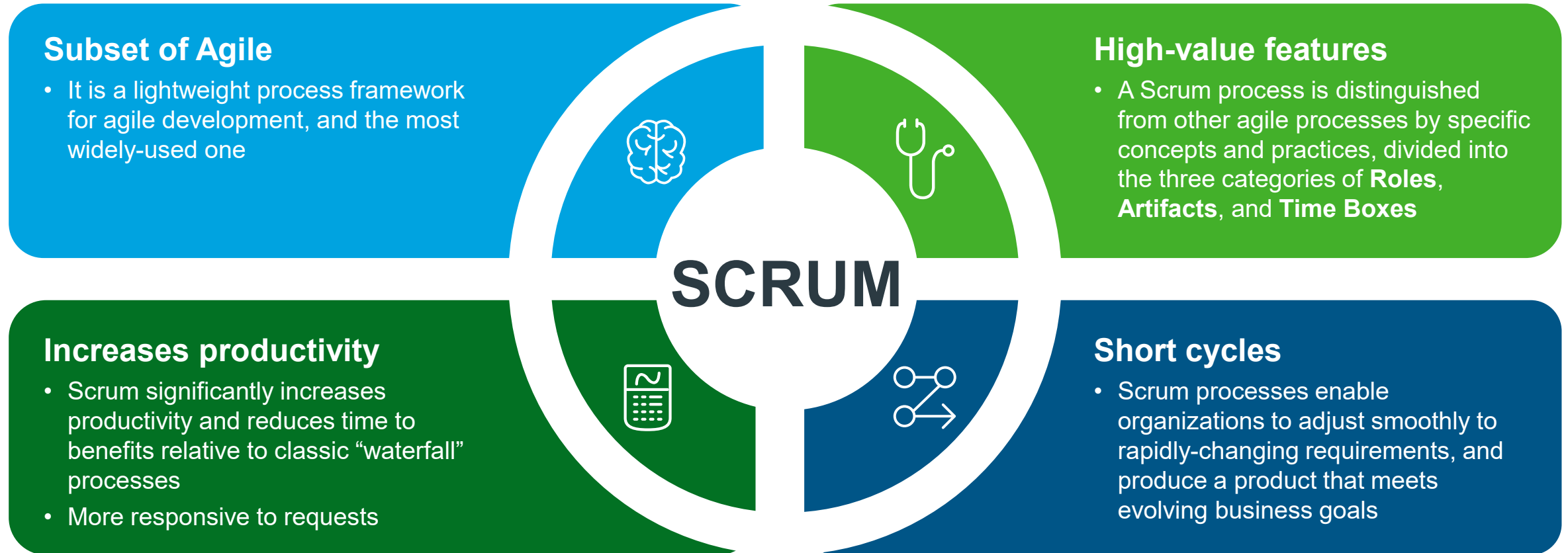
4

Self organizing cross functional team



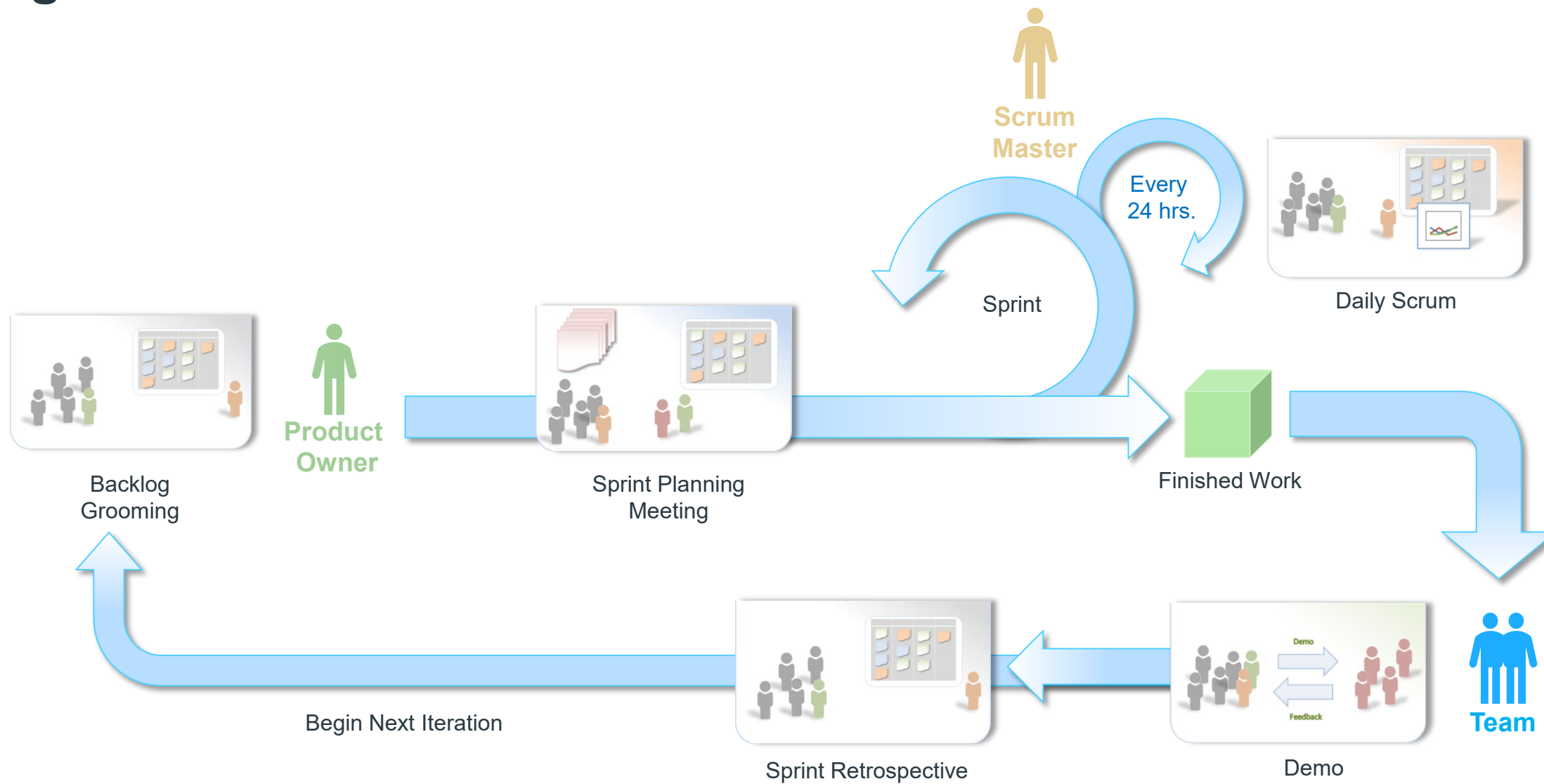
<https://www.cprime.com/resources/what-is-agile-what-is-scrum/>

Benefits of Agile Scrum



<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



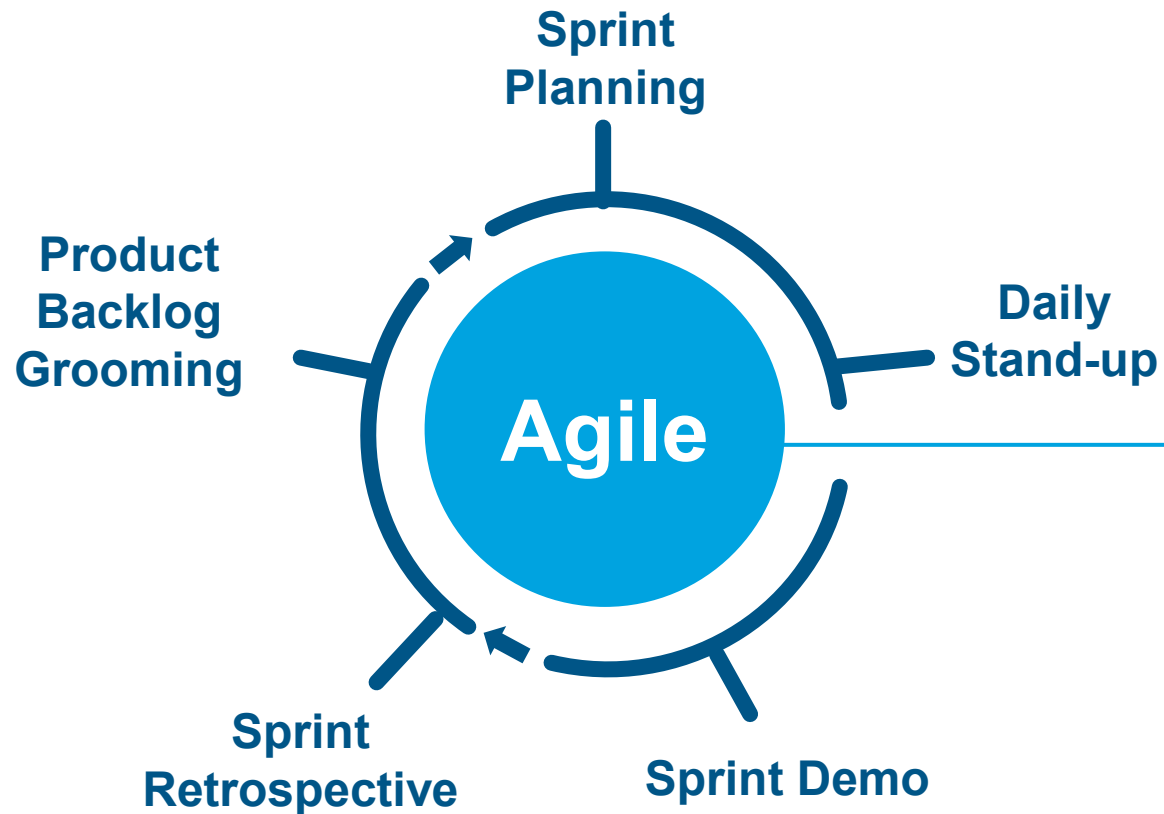
- Responsible for overall status of Sprint
- Help identify and remove impediments
- Blocks “noise” from team
- Ensures retrospective recommendations are executed
- Facilitate all ceremonies

Scrum Team



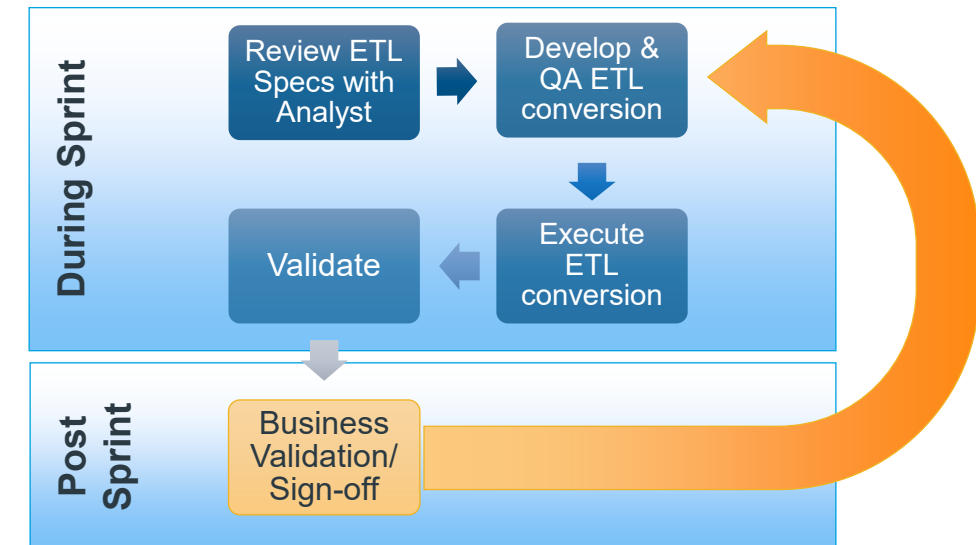
- “The Do-ers” consisting of 5 people, plus or minus 2
- Co-located - Cross-Functional - Dedicated
- Self-organizing / self-managing, without externally assigned roles
- Communicates commitments with the Product Owner, one Sprint at a time

OMOP Agile conversion methodology



What is Agile?

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



Cultural and behavioural changes

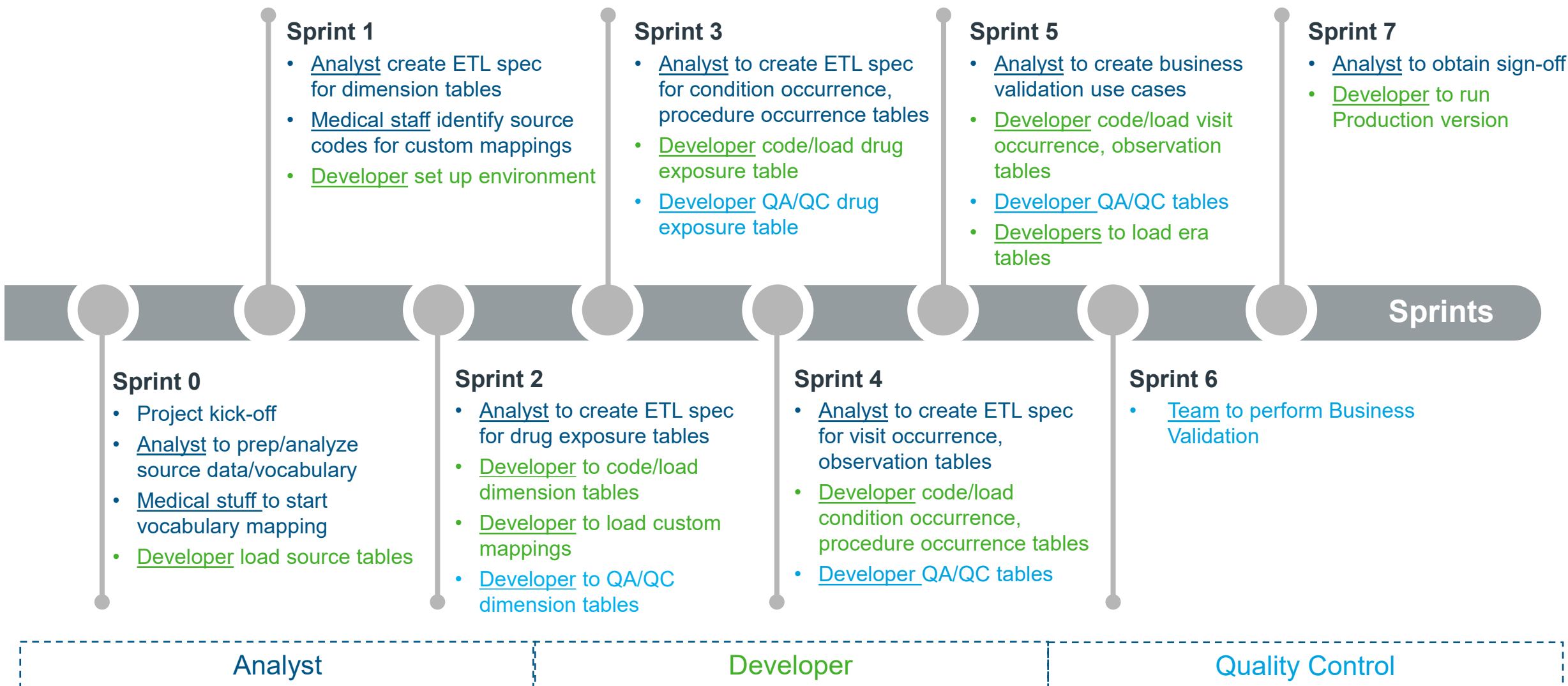
Waterfall

- ✗ 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

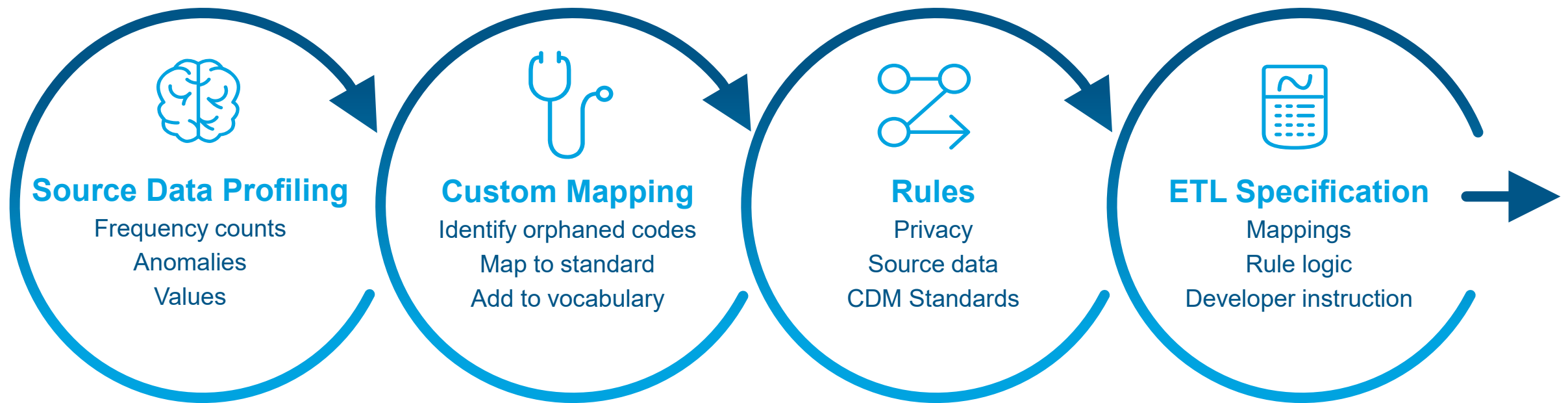
Conversion timeline in sprint – Example





Analysis of Source Data

Analysis process



Source Data - Synthea Database

SYNTHEA - Synthetic Patient Population Stimulator

What types of database? (EMR or Claims)

- Synthetic patient population simulator

Included Tables and data dictionary

- Included 12 tables. See next slide for included tables

Data Anomalies and Values

- Free of cost, privacy and security restrictions.



Synthea



Standard Health
Record



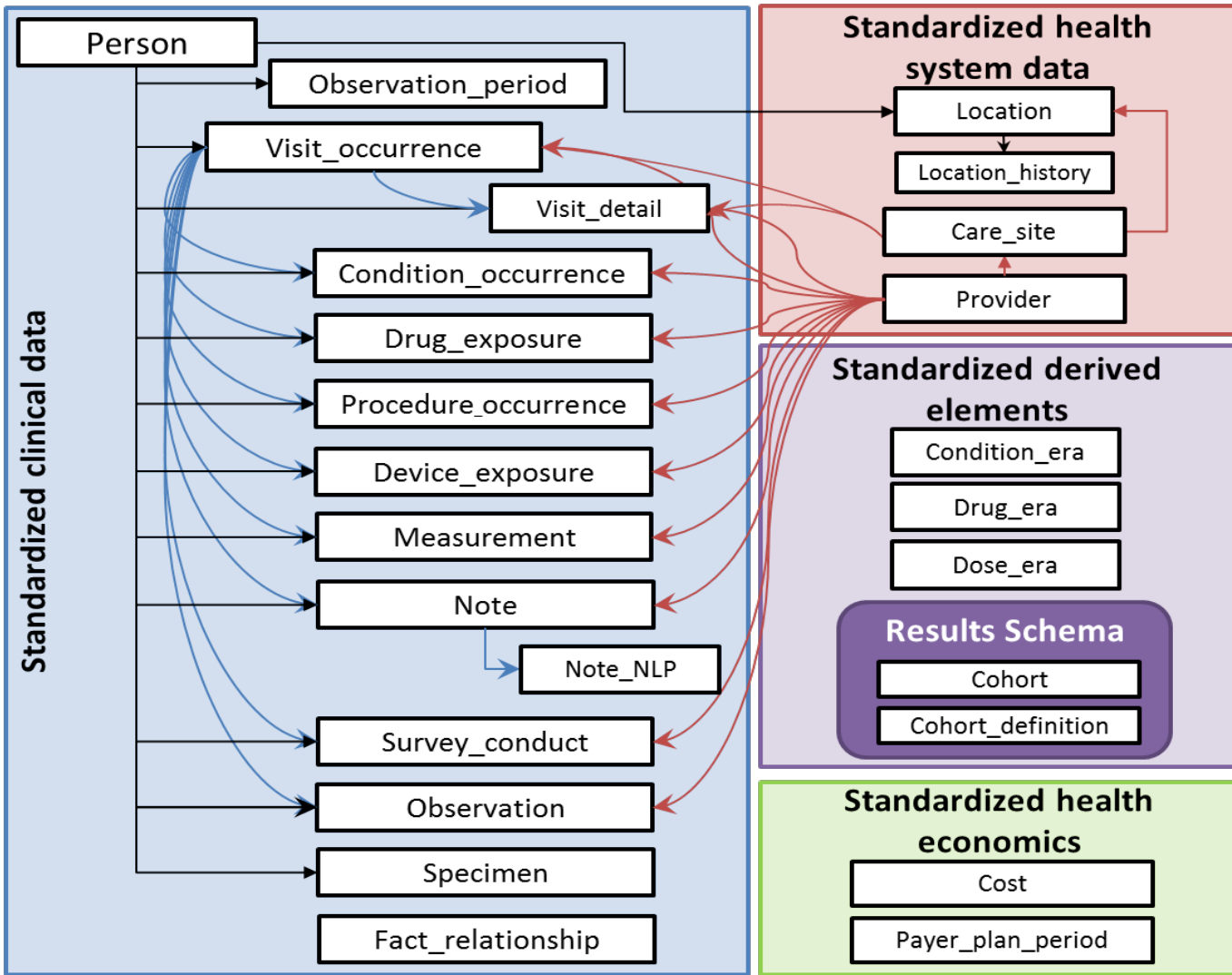
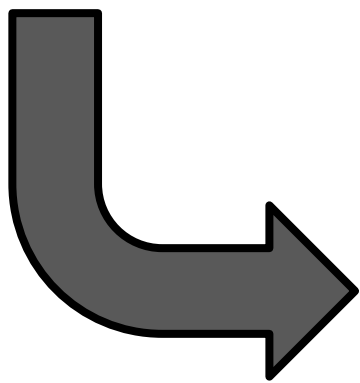
Synthetic Mass

Source Data - Synthea

Table	Description
Patients	Patient demographic data
Allergies	Patient allergy data
Careplans	Patient care plan data
Encounters	Patient encounter data
Imaging_studies	Patient imaging metadata
Immunizations	Patient immunization data
Medications	Patient medication data
Observations	Patient observations including vital signs
Organizations	Provider organizations including hospitals
Procedures	Patient procedure data including surgeries
Providers	Clinicians that provide patient care

Goal – Mapping Synthea to OMOP CDM

Synthea Source Data





Source Data Analysis with White Rabbit

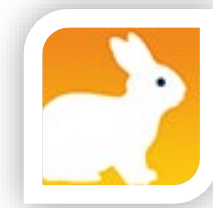
Source data profiling



- Used to analyze the structure and content of source data
- Assists with data types, values, frequency, anomalies
- Creates scan report of tables, columns, files
- 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_cvrgage_tot	integer	2	1031348	100000	0
11	beneficiary_summary	bene_smi_cvrgage_to	integer	2	1031348	100000	0
12	beneficiary_summary	bene_hmo_cvrgage_t	integer	2	1031348	100000	0
13	beneficiary_summary	plan_cvrg_mos_num	integer	2	1031348	100000	0
14	beneficiary_summary	sp_alzhmta	smallint	1	1031348	100000	0
15	beneficiary_summary	sp_chf	smallint	1	1031348	100000	0
16	beneficiary_summary	sp_chrmkidn	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_striktia	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
<div><div>< ></div><div>Overview</div><div>beneficiary_summary</div><div>carrier_claims</div><div>inpatient_claims</div><div>outpatient_claims</div><div>prescription_drug_events</div></div>							

White Rabbit – Location and Scan



White Rabbit

Help

Locations Scan Fake data generation

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

Source data location

Data type Delimited text files

Server location 127.0.0.1

User name

Password

Database name

Delimiter ,

Test connection

Console

White Rabbit

Help

Locations Scan Fake data generation

Tables to scan

Add all in DB

Add

Remove

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

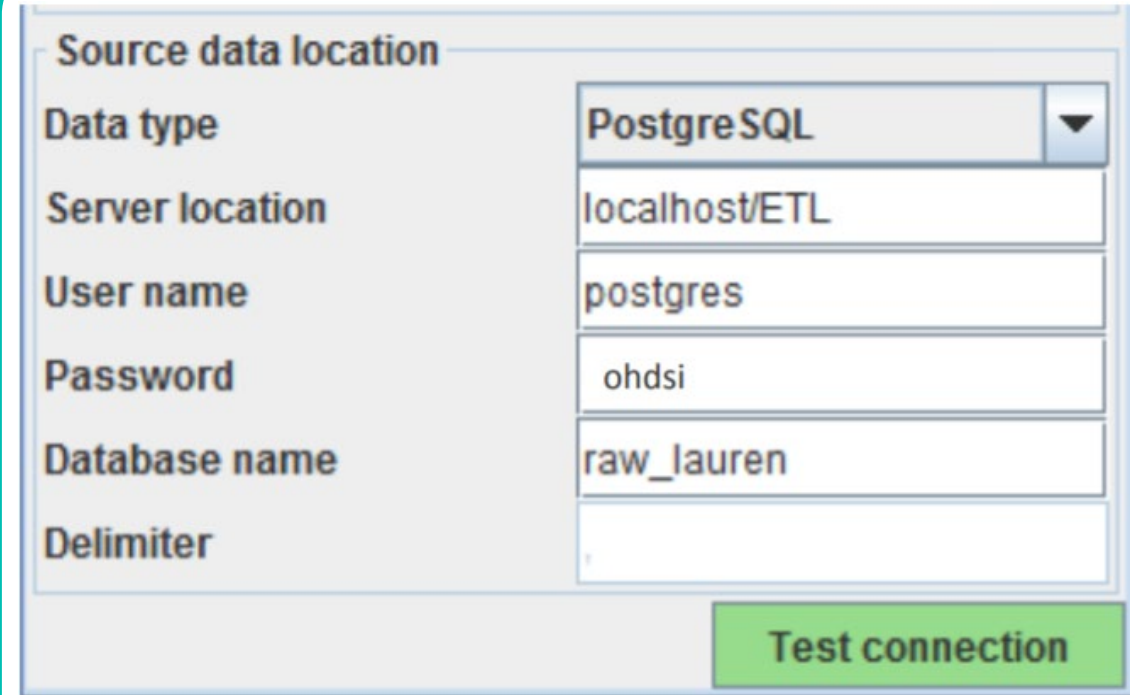
Scan tables

Console

Exercise – Scan Lauren’s Data

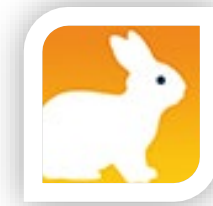


- Open OHDSI-in-a-box
- Click on WhiteRabbit shortcut
- Connect to Lauren’s Data
- Go to the “Scan” tab
- Press “Add all in DB” button, set “min cell count” to 0, and then “Scan tabs”
- Open ScanReport.xlsx

A screenshot of the WhiteRabbit configuration window. It has a title bar "Source data location". Below it are several fields: "Data type" is a dropdown menu showing "PostgreSQL"; "Server location" is a text box with "localhost/ETL"; "User name" is a text box with "postgres"; "Password" is a text box with "ohdsi"; "Database name" is a text box with "raw_lauren"; and "Delimiter" is a text box with a single quote character. At the bottom right is a green button labeled "Test connection".

Source data location	
Data type	PostgreSQL
Server location	localhost/ETL
User name	postgres
Password	ohdsi
Database name	raw_lauren
Delimiter	'
Test connection	

White Rabbit - Scan



White Rabbit

Help | Locations | **Scan** | Fake data generation

Tables to scan

Add all in DB

Add

Remove

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

Scan tables

Console

White Rabbit

Help | Locations | **Scan** | Fake data generation

Tables to scan

Add all in DB

Add

Remove

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

Scan tables

Console

Exercise – Create Agile timeline for your project

Background

You are the **scrum master** of converting your group's raw data to OMOP CDM and responsible for developing a project plan for the conversion. To successfully and effectively facilitate the conversion process, you need to apply **agile framework** to the conversion process.

Exercises

Create the agile timeline for your project.

Exercise – Using White Rabbit to Scan Lauren's Data

Background

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

Exercises

- When is Lauren's birth date?
- What is the most common condition Lauren has?
- What is the name and dose of the drug that Lauren has taken?
- How many outpatient visits does Lauren have?
- What is the date range of start date in condition table?
- What is the name of immunization Lauren has received?.



ETL Design

CDM Version Control

CDM version control resources: Working Group, Github, Wiki Resources

Working Group



+

CDM working group meets regularly to discuss proposed changes to the CDM

Github



+

All CDM documentation, versions, and proposals located on GitHub

Wiki Resources



+

Meeting information can be found on the working group [wiki page](#)

Three pillars of ETL design



Data Model

Determines the structure and data elements of clinically relevant domains (tables) model.



Vocabulary

Determines which domain (table) a clinical event goes to

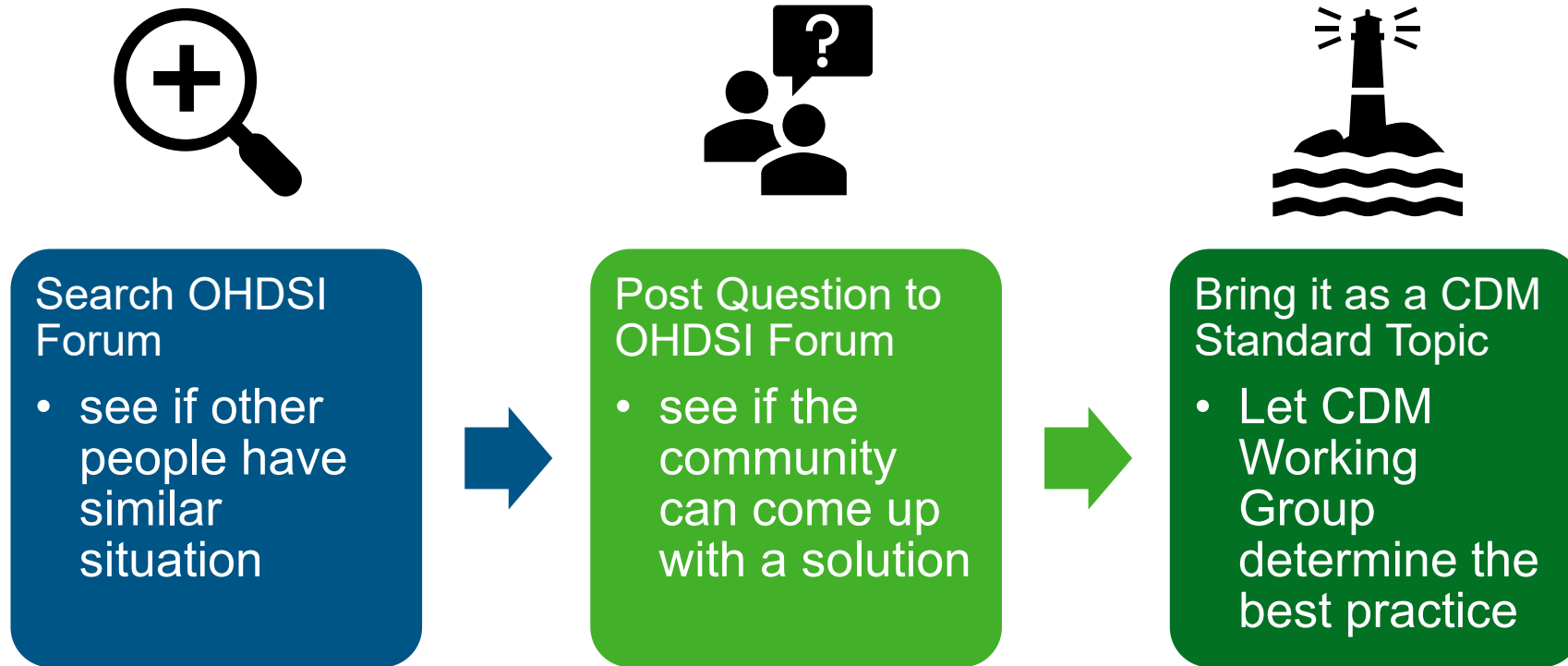


CDM Standards

Determines all the rules and conventions where data model and vocabulary do not govern

Steps to Follow When a New Situation is Encountered

To ensure abiding by OMOP standard/convention, steps below are recommended so that conversion can be done consistently across all data assets



CDM Standards and Conventions – Examples

(Procedure_Occurrence Table)

If a Procedure has a **quantity of '0'** in the source, this should default to '1' in the ETL. If there is a record in the source it can be assumed the exposure occurred at least once

When dealing with **duplicate records**, the ETL must determine whether to sum them up into one record or keep them separate. Things to consider are:

- Same Procedure
- Same PROCEDURE_DATETIME
- Same Visit Occurrence or Visit Detail
- Same Provider
- Same Modifier for Procedures
- Same COST_ID

Privacy considerations

Privacy manipulation can happen at 3 tiers: source data, OMOP data and client delivery

Source data tier ▶

+

Data elements are masked at the source level

Example: Clinical event dates are jittered in source tables

OMOP CDM tier ▶

+

Privacy manipulation happened at the OMOP CDM level

Example: Death dates are not allowed to be loaded into OMOP CDM

Organization tier ▶

+

Some privacy information are not available to all parties within an organization

Example: Psychological related clinical conditions are masked



Create ETL Spec

Creating ETL specification

1

Analyze Data

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

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'.

ETL Spec Table Writing Sequence

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

- Payer_Plan_Period
- Cost

ETL Spec Content – Common Data Elements to All Event Tables

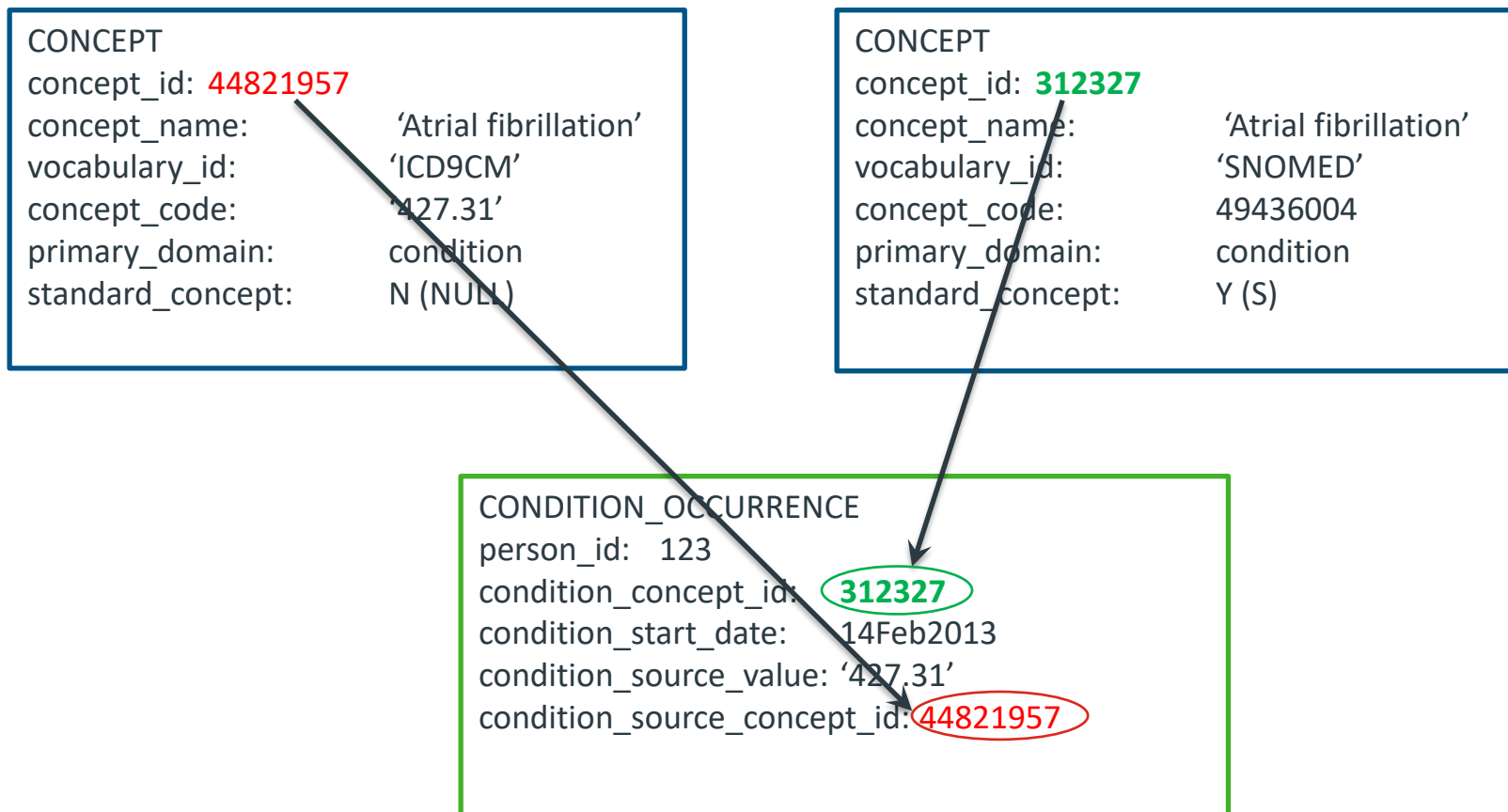
Common vocabulary related columns in clinical event tables

Field name	Purpose and example
<entity>_concept_id	Standard OMOP concept_id for source value condition_concept_id 313217 (SNOMED "Atrial Fibrillation")
<entity>_source_concept_id	OMOP concept_id for source value condition_source_concept_id 44821957 (ICD9CM "Atrial Fibrillation")
<entity>_source_value	Verbatim information from the source data, not to be used by any standard analytics condition_source_value 427.31 (ICD9CM "Atrial Fibrillation")
<entity>_type_concept_id	OMOP concept_id for the origin of the information Domain = 'Type concept', concept = 'Standard' in ATHENA (32817 'EHR')



Vocabulary Mapping

Integration of CDM and Vocabulary



Source code mapping to standards

Concept Code – 427.31

Concept Table – Source Concept

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
44821957	Atrial fibrillation	Condition	ICD9CM	5-dig billing code	NULL	427.31



Concept Relationship Table

concept_id_1	concept_id_2	relationship_id	valid_start_date	valid_end_date	invalid_reason
44821597	313217	Maps to	1/1/1970 0:00	12/31/2099 0:00	NULL



Concept Table – Standard Concept

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
313217	Atrial fibrillation	Condition	SNOMED	Clinical Finding	S	49436004

```
SELECT *
FROM concept c
LEFT JOIN concept_relationship cr ON c.concept_id = cr.concept_id_1 AND cr.relationship_id =
'Maps to'
LEFT JOIN concept c2 ON cr.concept_id_2 = c2.concept_id
WHERE c.concept_code = '427.31'
```

One source field can go to multiple CDM domains

An example showing source Diagnosis table (diagnosis_code) can be mapped to different domains

diagnosis_code (ICD9CM)	diagnosis_description
525.5	Partial Edentulism
V26.33	Genetic Counseling
V18.2	Family History of Anemia
790.2	Abnormal Glucose

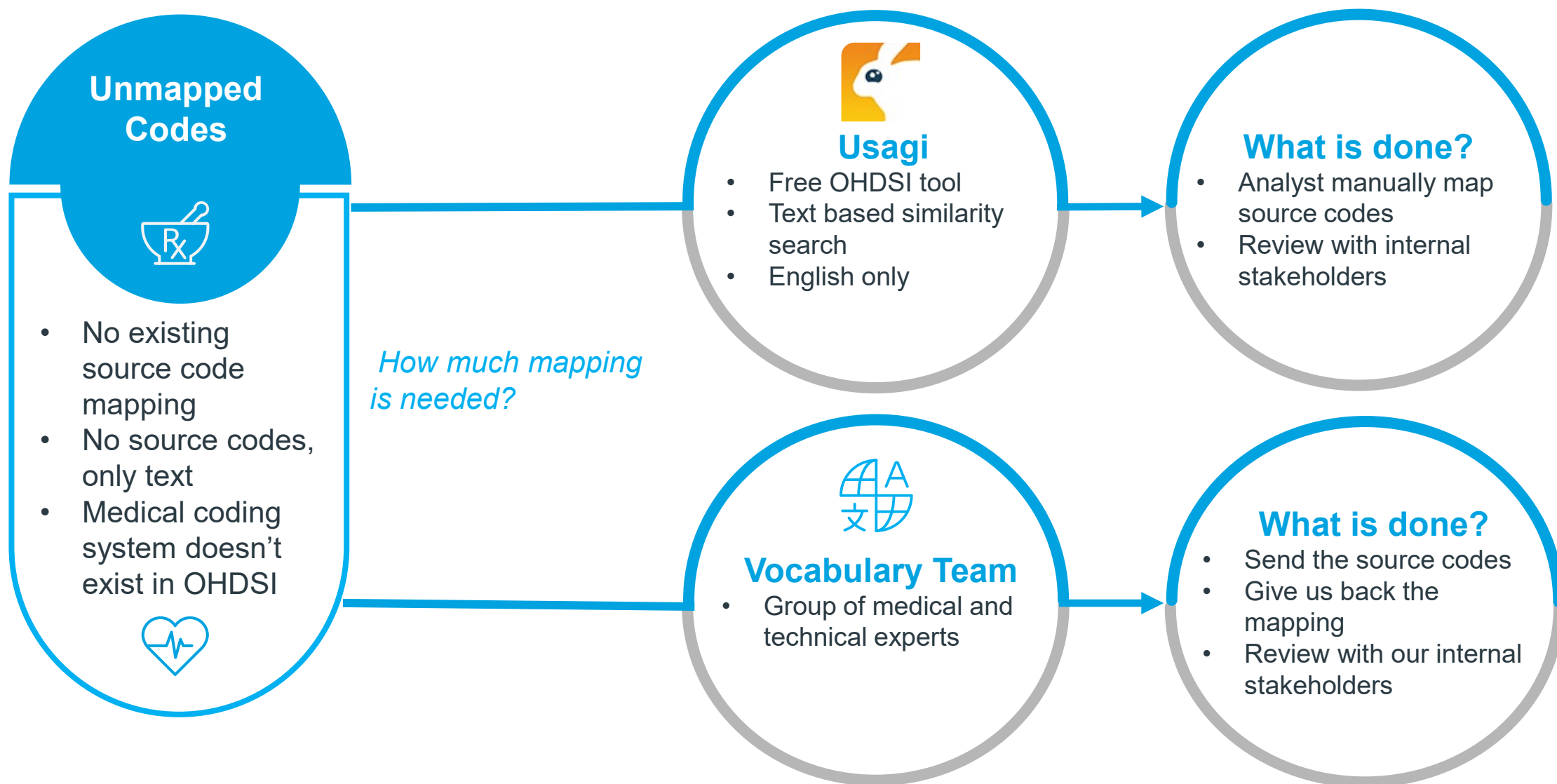


concept_id (standard)	concept_name (standard)	domain_id
40481091	Partial edentulism	Condition
4196362	Genetic counseling	Procedure
4167217	Family history of clinical finding	Observation
4149519	Glucose measurement	Measurement



Custom Mapping Process

Custom source code mapping



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



- 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%

Exercise – Review the Usagi Mapping

Background

You have some custom code that you must map but it only has text fields.

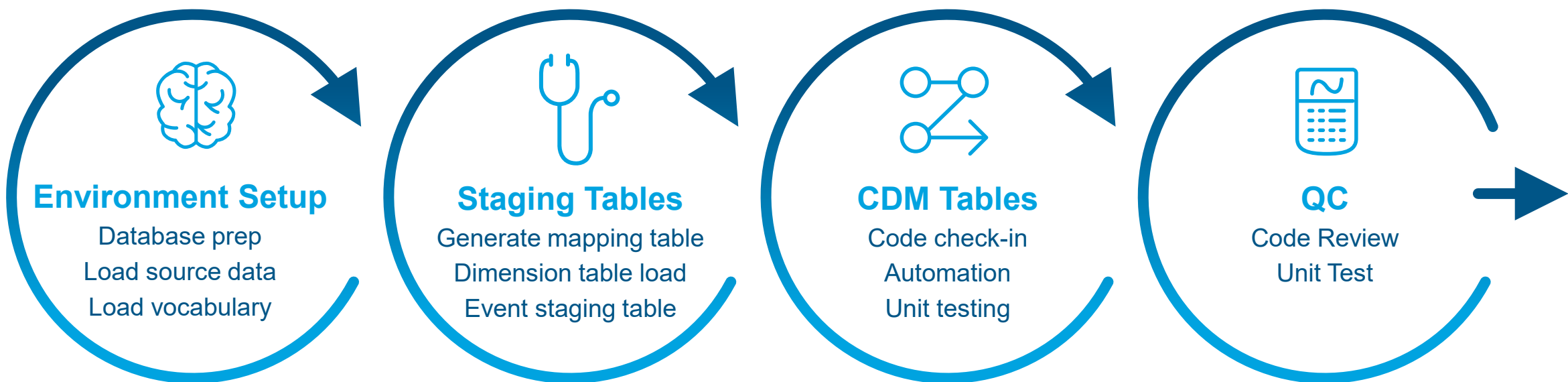
Exercises

Using the file we've loaded, review the Usagi mapping to see which one is the best choice



ETL Workflow

ETL Workflow

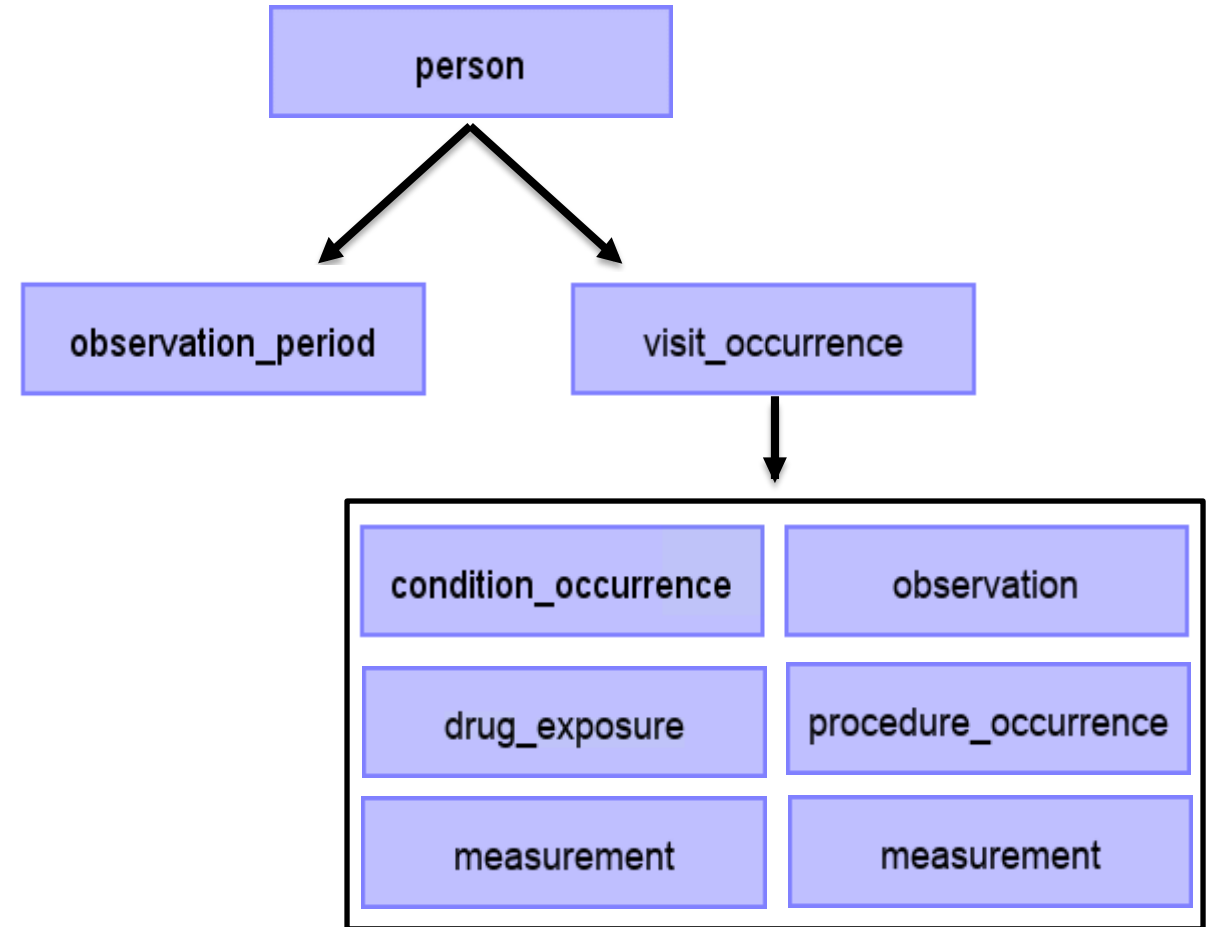


ETL Implementation

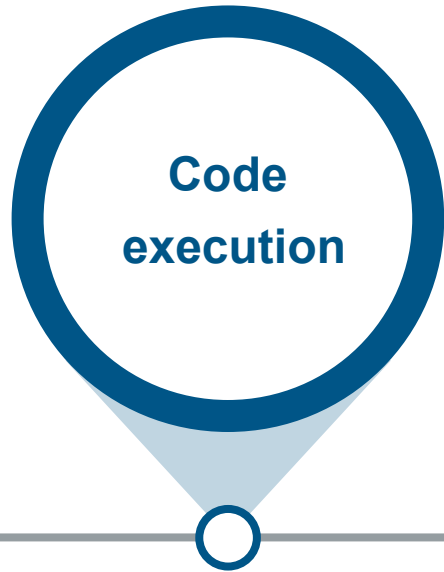
Step 1. Source Data to staging table

Step 2. Staging table to OMOP CDM tables

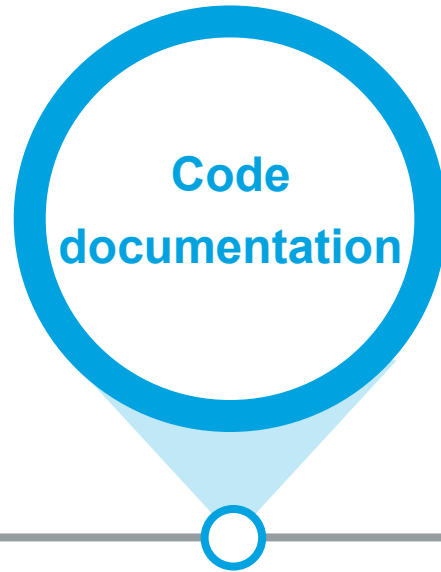
Step 3. QA and Validation



Coding Best Practices



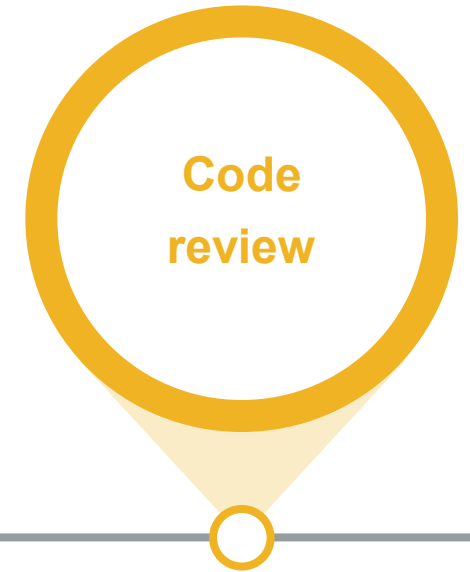
Does it work?
Does it follow the ETL rules?



Can it be interpreted?
Is a guideline or SOP?



Are there coding standards?
Is the code written in the most efficient way?



Does it conform to internal guidelines?
Does both developer understand the same ETL rules?

Code Review After ETL Development

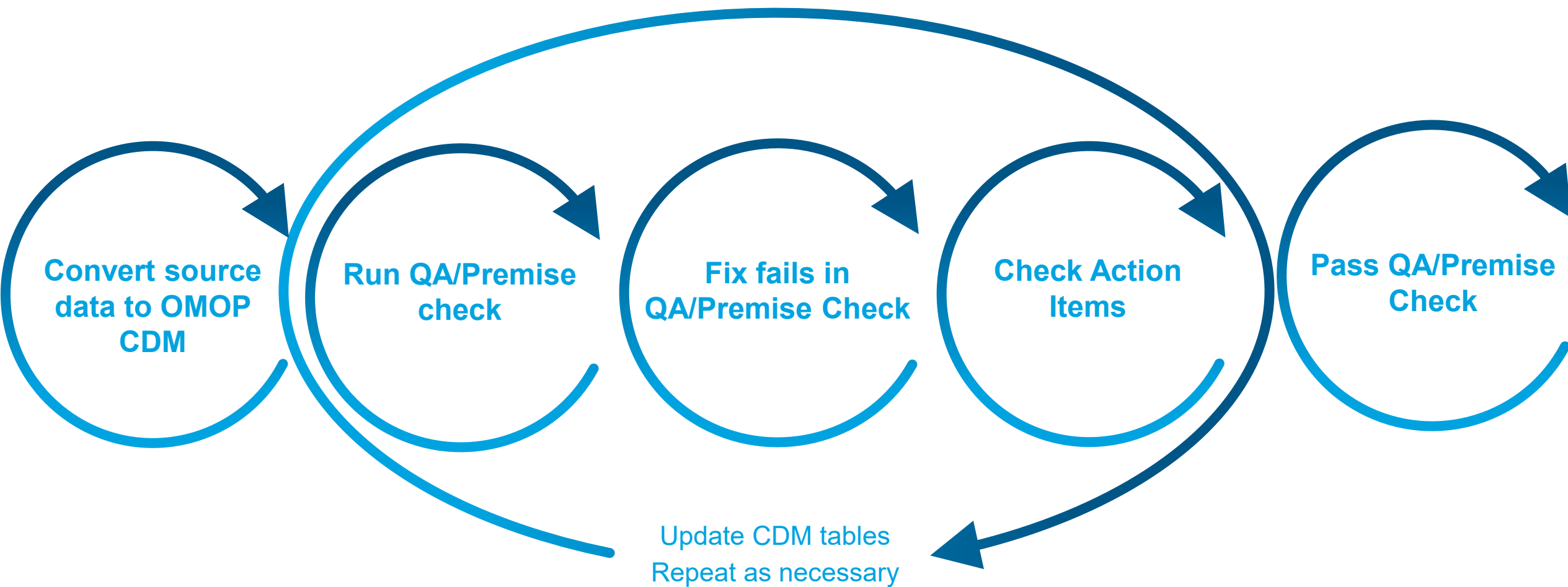
- Peer review of new/modified code
- Allows for “another set of eyes”
- Designed to catch bugs/errors
- Enforces standards
- Knowledge transfer/information sharing
- Reduce rework/troubleshooting in the future

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

    bool found = false;
    ItemToItem moved = null;
    ItemToItem previous = null;
    ItemToItem next = null;
    foreach (int orderedChildId in orderedChildIds)
    {
        ItemToItem current = ChildItems.FirstOrDefault(c => c.ChildId == orderedChildId);
        if (current != null)
        {
            if (current.ChildItem.ItemId == movedChildId)
            {
                moved = current;
                found = true;
            }
            else
            {
                // ...
            }
        }
    }
}
```

Data Quality Tools and Scripts

QA/QC Process



Data quality checks

Internal QC Checks

Over 230+ quality checks

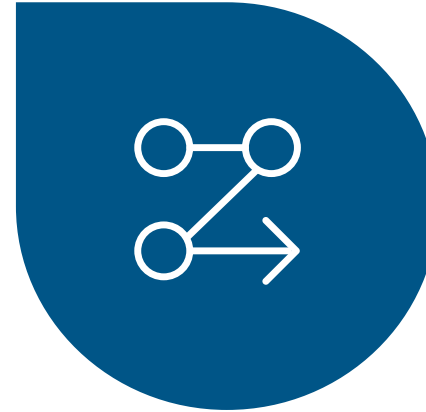
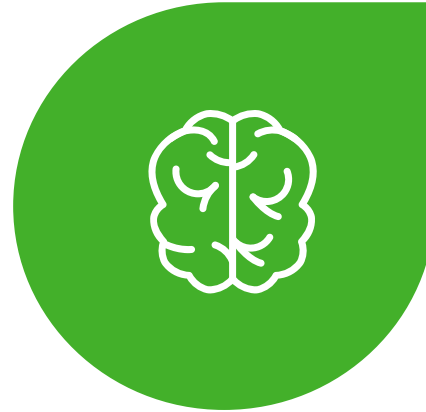


Data Quality Dashboard

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

Achilles

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



Demo Statistics

Vocabulary mapping statistics

Internal QC checks

Description

- Created by IQVIA OMOP Team
- Contains over 230+ checks
- Ensures CDM standard conventions are followed
- Documents pass/fails and record counts
- Checks executed during each sprint within the development process
- Checked during development and initial conversion

Deliverable

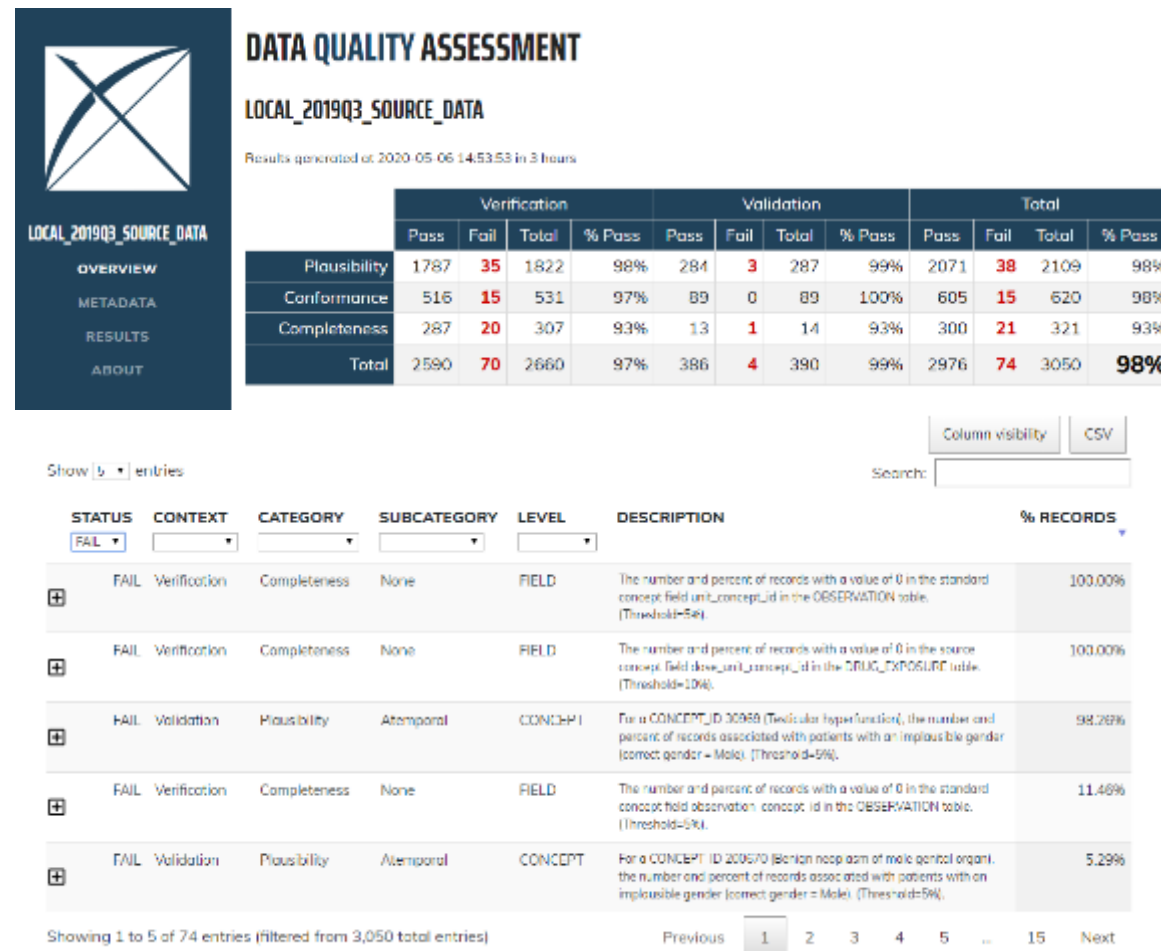
VALIDATIONACTION	COUNTDETAILS	VALIDATIONRESULT
A unique identifier for each Care Site		PASS
populated with a concept_id from concept_table where domain_id = Place of Service and standard_concept = S. If none is found then just populate this field with 0.	0	PASS
Primary key. A unique identifier for each row in the table.		PASS
This is a required field. A foreign key identifier to the Person in the person table.		PASS
This is a required field. It is populated with a concept_id from concept_table where domain_id = Condition and standard_concept = S. If none is found then just populate this field with 0.		PASS
Must be populated with a date after cdm.person.year_of_birth	389157	FAIL
Must be populated with a date before cdm.death.death_date + 60 days	0	PASS
Must be populated with a date after cdm.person.year_of_birth	0	PASS
Must be populated with a date before cdm.death.death_date + 60 days	0	PASS
If populated, must be between 00:00 and 23:59	0	PASS
Must be populated with a concept_id from domain_id = Type Concept, vocabulary_id = Condition Type, and standard_concept = S OR 0	0	PASS
If populated, must be with a provider_id from the provider table	0	PASS
If populated, must be with a visit_occurrence_id from the visit_occurrence table	0	PASS
This is not a required field. It is populated with a concept_id from concept_table		PASS
This is a required field. A foreign key identifier to the Person in the person table. Each person_id can only have 1 death_date.		PASS
Must be populated with a date after cdm.person.year_of_birth	0	PASS
This is a required field. It is populated with a concept_id from concept_table where domain_id = Type Concept, vocabulary_id = Death Type, and standard_concept = S. If none is found then just populate this field with 0.		PASS
This is a required field. It is populated with a concept_id from concept_table where domain_id = Condition. If none is found then just populate this field with 0.		PASS

Data quality dashboard

Description

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

Deliverable

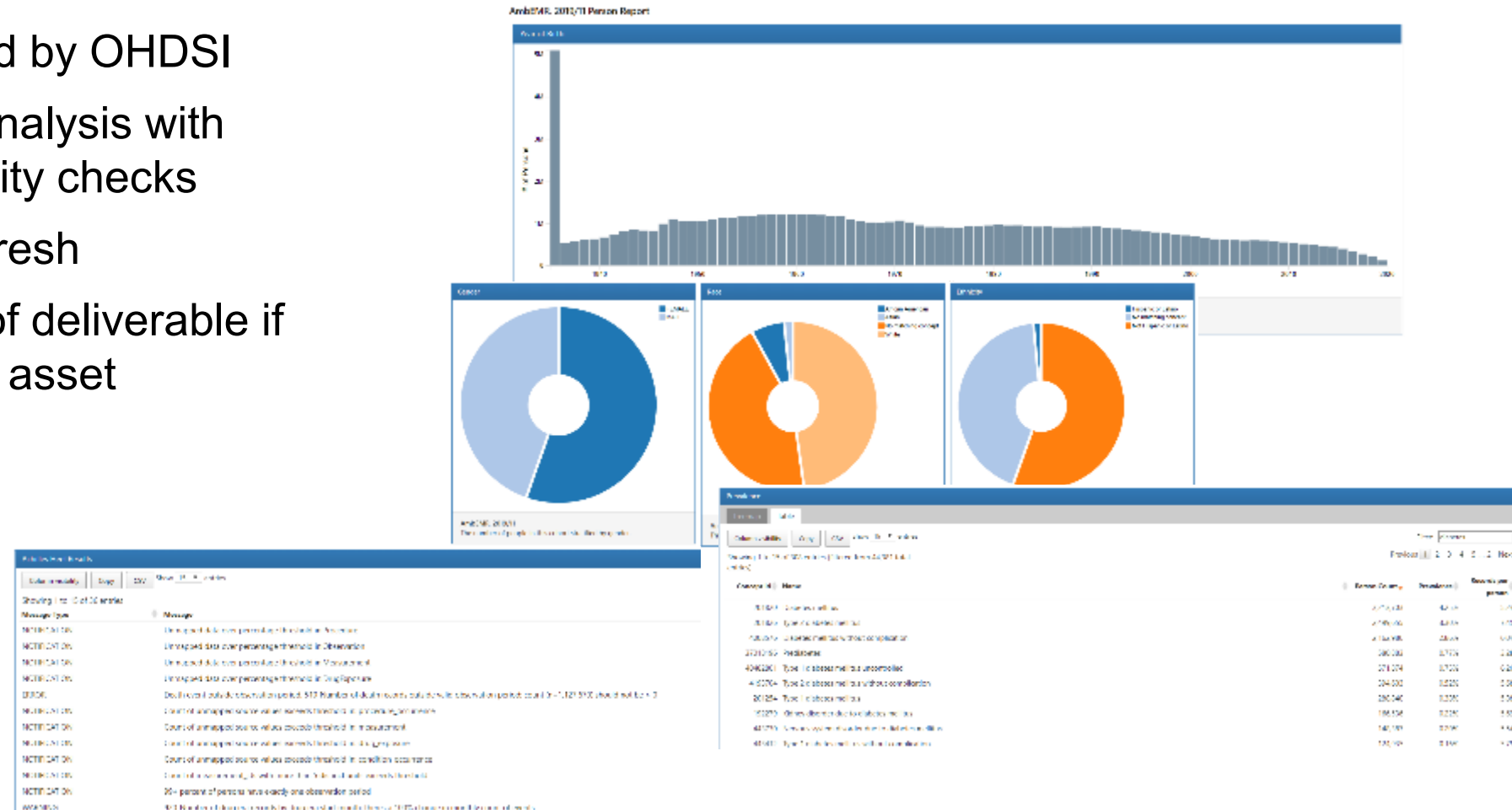


Achilles

Description

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

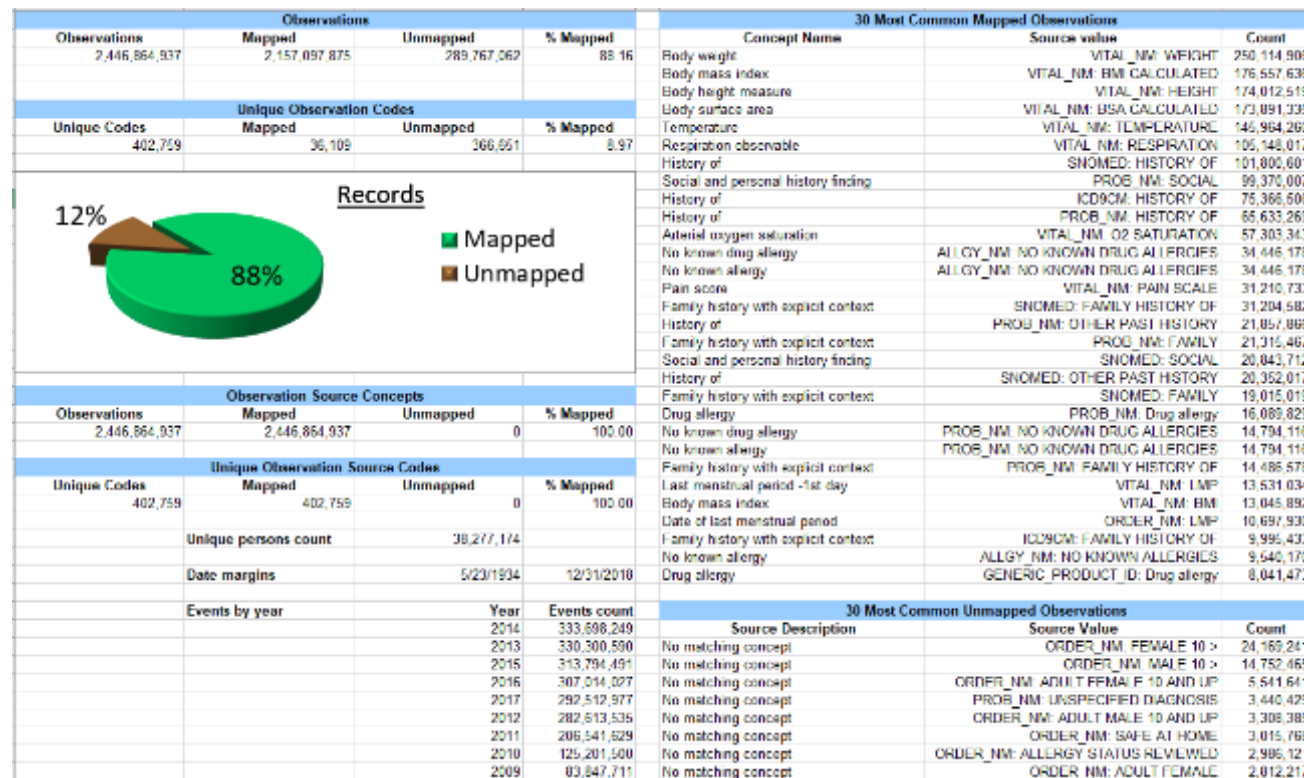
Deliverable



Description

- Created by IQVIA OMOP Team
- High level statistical counts for tables, mapping percentages, most common mappings
- Executed with each refresh
- Sent to clients as part of deliverable if purchased OMOP data asset

Deliverable

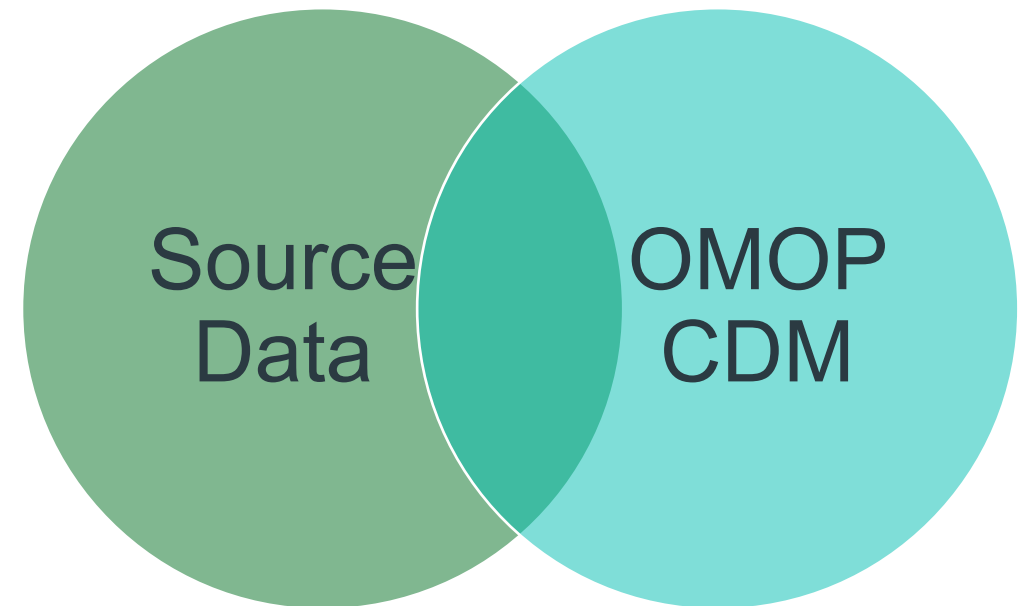


Business validation

Description

- Comparison of use cases between source data and OMOP data
- Completed in collaboration with the data owner(s)
- Executed once at the end of an OMOP conversion
- Tool for data owner(s) to sign off on conversion

Deliverable



80/20 Rule

Conclusions

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

Lessons Learned

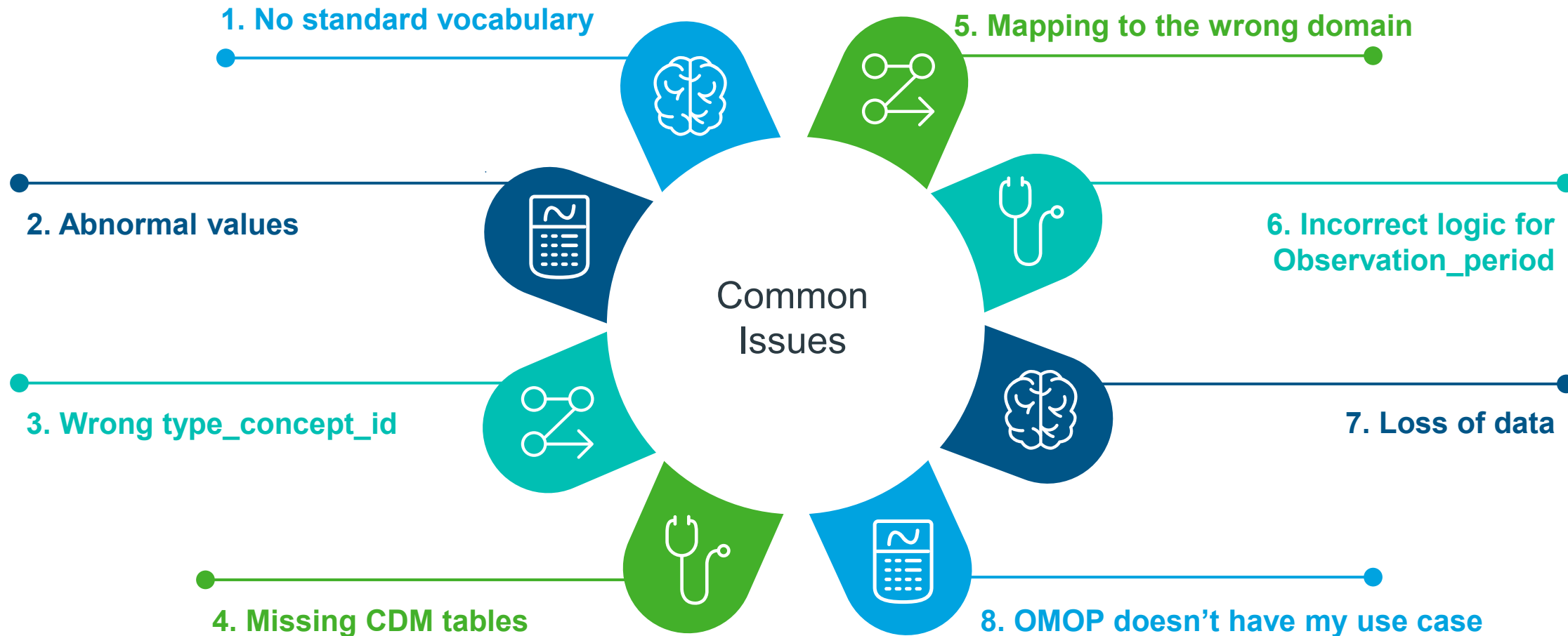
ETL helps standardize source data to research quality. The goal is to accurately transform the data into CDM format and standardized terms with acceptable information loss, and high-frequency source codes are mapped.



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

Conversion Challenges

Common OMOP CDM issues



1. No Standard Vocabulary

Issue

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

Source Value	Count
VITAL_NM: COMMENTS	3,057,155
VITAL_NM: ALLERGY STATUS REVIEWED	2,991,269
TEST_NM: HEMOGLOBIN A1C GLYH	1,470,687
VITAL_NM: ACCOMPANIED BY:	1,189,602
VITAL_NM: (YEARLY AND AS INDICATED) DO YOU EVER FEEL UNSAFE AT HOME?	1,174,079
VITAL_NM: HCC SCORE	1,102,620
TEST_NM: CHOLESTEROL-VLDL	1,020,947
TEST_NM: CHOLESTEROL-LDL	1,020,745
VITAL_NM: NURSING COMMENTS	1,008,800
VITAL_NM: (YEARLY) DO YOU HAVE ANY RELIGIOUS OR CULTURAL BELIEFS THAT MAY IMPACT YOUR CARE?	672,286
VITAL_NM: ***ARE YOU HAVING PAIN RELATED TO TODAY'S VISIT?	663,437
VITAL_NM: (EVERY VISIT) WOMEN ONLY: WOULD YOU LIKE TO HAVE A FEMALE CHAPERONE DURING THE EXAM PART OF YOUR VISIT?	619,900
TEST_NM: EGRFAA	616,850
VITAL_NM: *HAVE YOU RECENTLY HAD THOUGHTS OF HARMING YOURSELF OR OTHERS?	598,256
VITAL_NM: PAIN SITE	595,640
TEST_NM: LDL CALC	501,707
TEST_NM: GLUCOSE AU480	490,157
TEST_NM: UA - URINE SEDIMENT	460,089
TEST_NM: UA - REDUCING SUBSTANCE	440,470
TEST_NM: LYMPHOCYTES RELATIVE PERCENT	419,561
TEST_NM: MONOCYTES RELATIVE PERCENT	419,558
TEST_NM: NEUTROPHILS RELATIVE PERCENT	416,094
TEST_NM: HEMOGLOBIN CDS	364,652
TEST_NM: RBC DIS.WIDTH-SD	364,010
TEST_NM: RBC DIS.WIDTH-CV	363,712
TEST_NM: TSH (THYROID STIM HORMONE)	347,177
TEST_NM: ALT / GPT	343,109
TEST_NM: HEMATOCRIT CDS	341,776
VITAL_NM: COMMENT	338,204
TEST_NM: PLT CDS	335,977

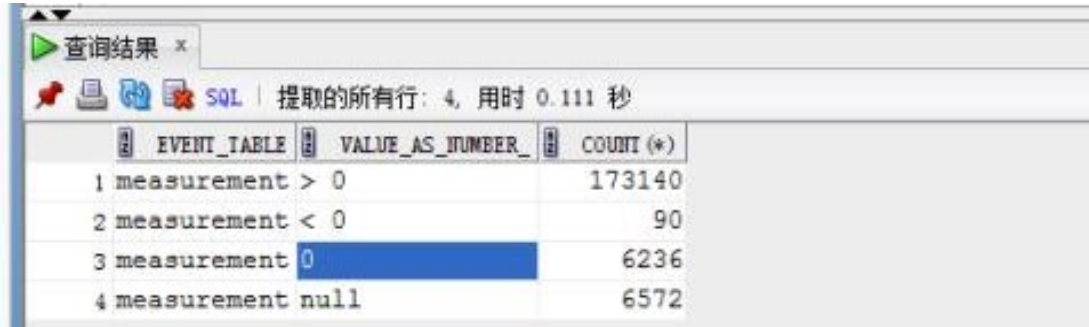
Solution

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

2. Abnormal Values

Issue

- Negative, 0, null values in measurement and drug_exposure tables



The screenshot shows a SQL query result window with the title '查询结果 x'. Below the title bar, there are icons for a red pin, a printer, a document, and a red 'X' icon, followed by the text 'SQL | 提取的所有行: 4, 用时 0.111 秒'. The main content is a table with three columns: 'EVENT_TABLE', 'VALUE_AS_NUMBER_', and 'COUNT (*)'. The table contains four rows of data.

	EVENT_TABLE	VALUE_AS_NUMBER_	COUNT (*)
1	measurement	> 0	173140
2	measurement	< 0	90
3	measurement	0	6236
4	measurement	null	6572

Solution

- Check source data for related domains and check if it's reasonable from medical perspective
- If valid, leave the values as they were. If not, remove the records as dirty data.

3. Wrong Type_concept_ids

Issue

- Used internal coding systems for type_concept_ids
- Wrong meanings were assigned to type_concept_ids

Solution

- Standardize all type_concept_ids in each table
- Find correct concept_id using ATHENA
- Guidelines:
 - use **condition_type_concept_id**= 32019 for EHR Billing Diagnosis, 42894222 for EHR Chief Complaint, 32030 for EHR Encounter Diagnosis, 45754805 for EHR Episode Entry, 38000245 for EHR Problem List Entry, 43542353 for Observation Recorded from EHR
 - use **drug_type_concept_id**= 38000180 for inpatient administration, 38000175 for prescription dispensed in pharmacy
 - use **device_type_concept_id** =44818707 for EHR detail, 32465 for inferred from claim, 44818705 for inferred from procedure, 44818706 for patient reported device

4. Missing CDM tables

Issue

- Incomplete OMOP CDM tables

Solution

- Check source data for related tables to see if available
- Provide mapping rules from source data to OMOP CDM, and populate the missing tables
- Apply Minimal Viable Product (MVP) – See next slide

4. Apply Minimal Viable Product (MVP)

Health System Tables	Clinical Data Tables	Derived Tables (Logic Provided)	Health Economic Tables
<ul style="list-style-type: none">• Location• Care_Site• Provider• Person• Death	<ul style="list-style-type: none">• Visit_Occurrence• Condition_Occurrence• Drug_Exposure• Procedure_Occurrence• Measurement• Observation• Observation_Period• Specimen• Device_Exposure• Fact_Relationship• Visit_Detail• Note• Note_NLP	<ul style="list-style-type: none">• Drug_Era• Dose_Era• Condition_Era	<ul style="list-style-type: none">• Payer_Plan_Period• Cost

5. Mapping to the wrong domain

Issue

- The source vocabulary domain may differ from its mapped standard vocabulary domain.
- Example: for ICD10CM Z82.49 – Family history of heart disease, it maps to concept 4148407 [FH: Cardiovascular disease](#), which is not in Condition domain, but Observation domain.

Solution

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

6. Incorrect logic for observation_period

Issue

- The observation_period for patients didn't cover the whole time period of all the events that patients had in the data.
- Example – patient A's observation_period was recorded as 1/1/2020 to 1/1/2021, but in the condition_occurrence table, there was a record of diabetes diagnosis on 2/1/2021.

Solution

- Use the observation_period logic and codes from OHDSI Github when doing OMOP conversion.
- Refresh the observation_period table in every data update.
- OHDSI Github resource
 - https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:observation_period

7. Loss of Data

Issue

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

Solution

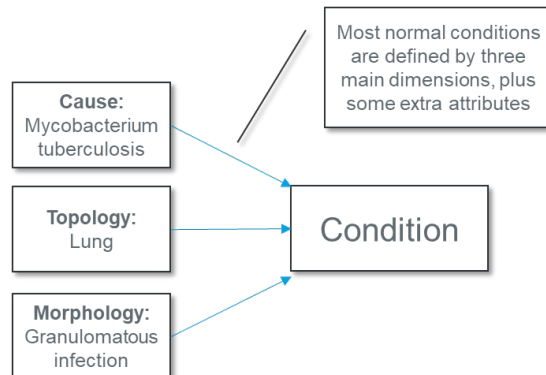
- Logic is introduced to ensure patients are valid
 - Test patients
 - Patients without birth year
 - Patient without any transaction
 - Depends on the data and scenario
- Some fields are used to derived the logic of the CDM field
 - For example: ICD Type helps determine if the code is an ICD9 or ICD10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

8. OMOP doesn't have my use case

Issue

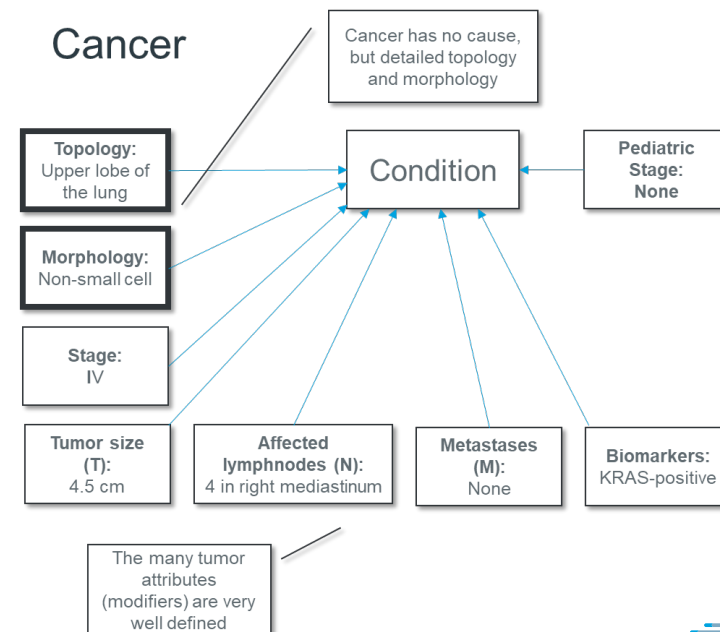
- OMOP cannot support oncology data.

Normal Conditions

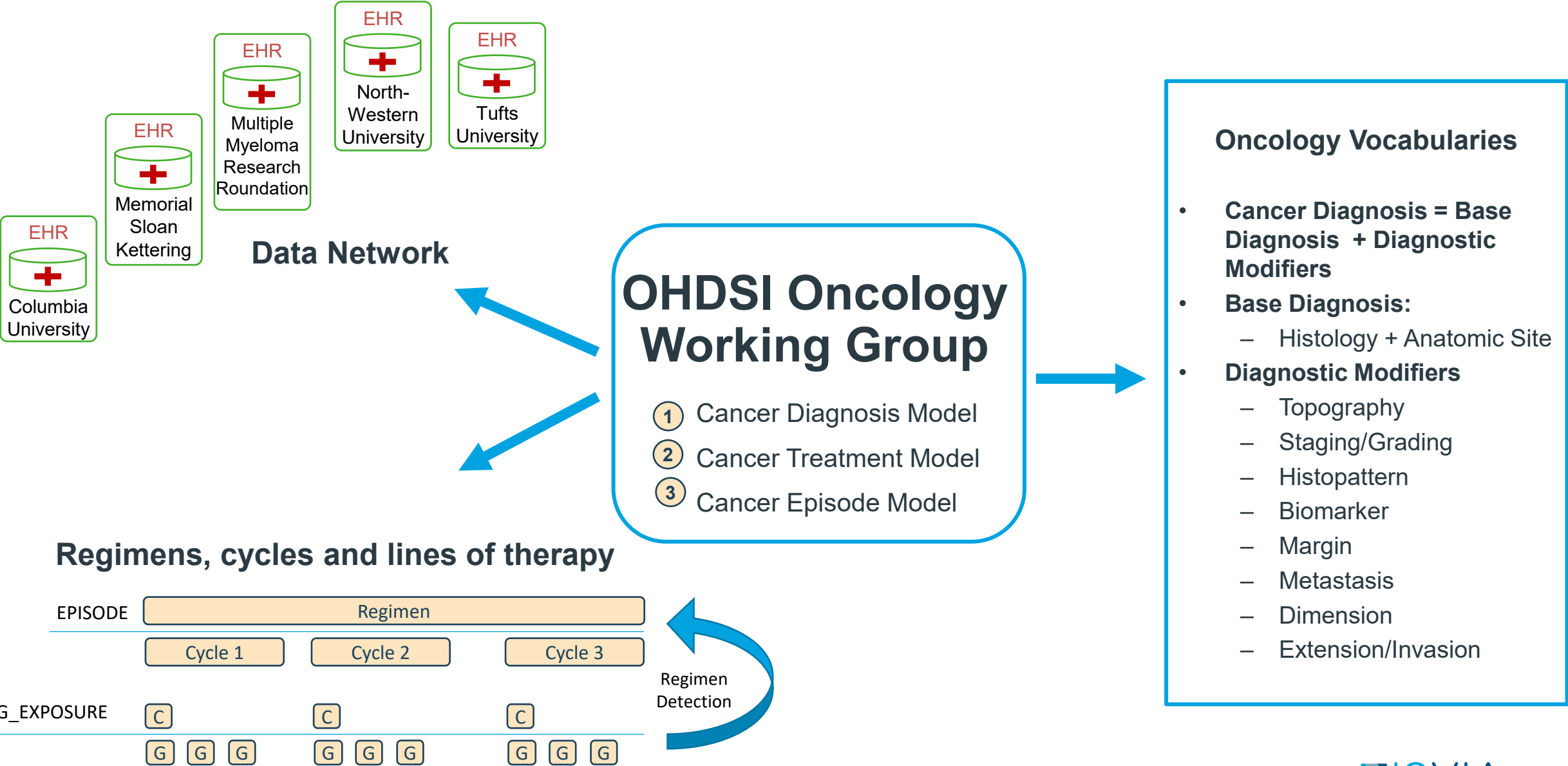


Solution

- OMOP Oncology experts created an Oncology working group.
- The Oncology working group designed an Oncology extension to house oncology-specific information in the OMOP CDM.



8. OMOP doesn't have my use case





Open Discussions

Question List

- **General Questions**

- The lead site who is writing a SQL does not have data with concept code used by other responding sites (e.g., Vitamin D). How can the lead site write a SQL without seeing the data?
- When multiple sites have populated differently with OMOP required fields (date) vs. optional fields (e.g. datetime), how do we harmonize it?
- Is there a guideline where to ETL data with different granularity? (e.g., To find ICU events/patients where do we look/store/ETL? visit_occurrence table vs. visit_detail tables)
- How can I know that I have mapped fields in a given SQL, BEFORE running it?

- **Dimensional tables**

- When the birthday falls in between visits, the age at the time of visit will be different and this patient will be counted twice when the unit of query/count is encounter level. What's the solution?
- How do we store/retrieve cause of death?

- **Clinical events tables**

- How to distinguish between admitting diagnosis and discharge diagnosis?
- How do we harmonize missing or different unit (need for unit conversion) for lab test measurements? Between UCUM and SNOMED codes, how do we map? Which one is more prevalent? e.g. we tried to provide a supplementary SQL of simple group-by count

CDM Planning Session – March 30th at 1 PM EST

- Release of CDM Version 5.4 later this year, the Common Data Model workgroup has scheduled a special two-hour planning session
 - Tuesday, March 30, at 1 pm
 - CDM Teams environment.
- Clair Blacketer has shared a proposed list of changes

https://ohdsiorg.sharepoint.com/sites/Workgroup-CommonDataModel/_layouts/15/Doc.aspx?sourcedoc={e6d1b920-83fd-43ad-a72c-5e76cfb81d3c}&action=edit&wd=target%28CDM%202021%20OKR.one%7Cd6360dad-b1cf-44c2-b694-c3c5d19bf227%2FKR1%20List%20of%20Changes%7C4632df59-d59f-7146-a77f-71f4f4edf46d%2F%29

Please join for some or all of the meeting if you are interested



Thank you

