ETL development & updating your ETL

OHDSI Europe Symposium

April 18th 2026

Maxim Moinat, Liam Glueck, Anne van Winzum

Liam Glueck

Anne van Winzum

Maxim Moinat

Extract Load Transform (ETL) journey

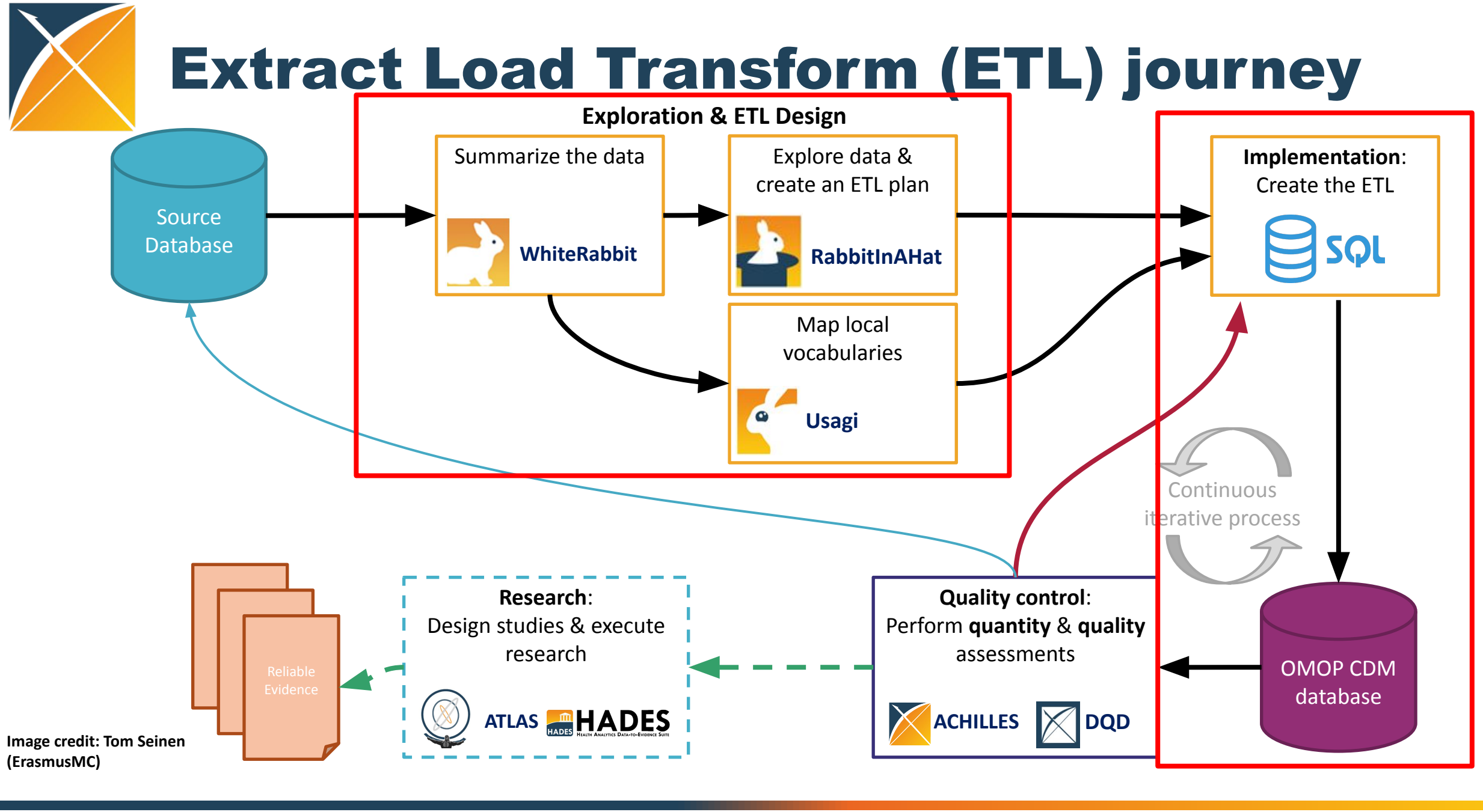


Image credit: Tom Seinen (ErasmusMC)

Your Journey

- Who has followed the EHDEN Academy course for the OMOP CDM?
- Who is in the process of converting their data to the OMOP CDM?
- Who already has an OMOP CDM instance completed?
- Who is already running studies on their OMOP CDM instance?

Agenda

Introduction	Maxim Moinat
ETL Implementation demo (part 1)	Liam Glueck
ETL Implementation demo (part 2)	Anne van Winzum
Q&A / Break	
ETL Execution Considerations	Anne van Winzum
Loading new OMOP vocabulary	Liam Glueck
Release comparison	Maxim Moinat
Q&A / Conventions Quiz	

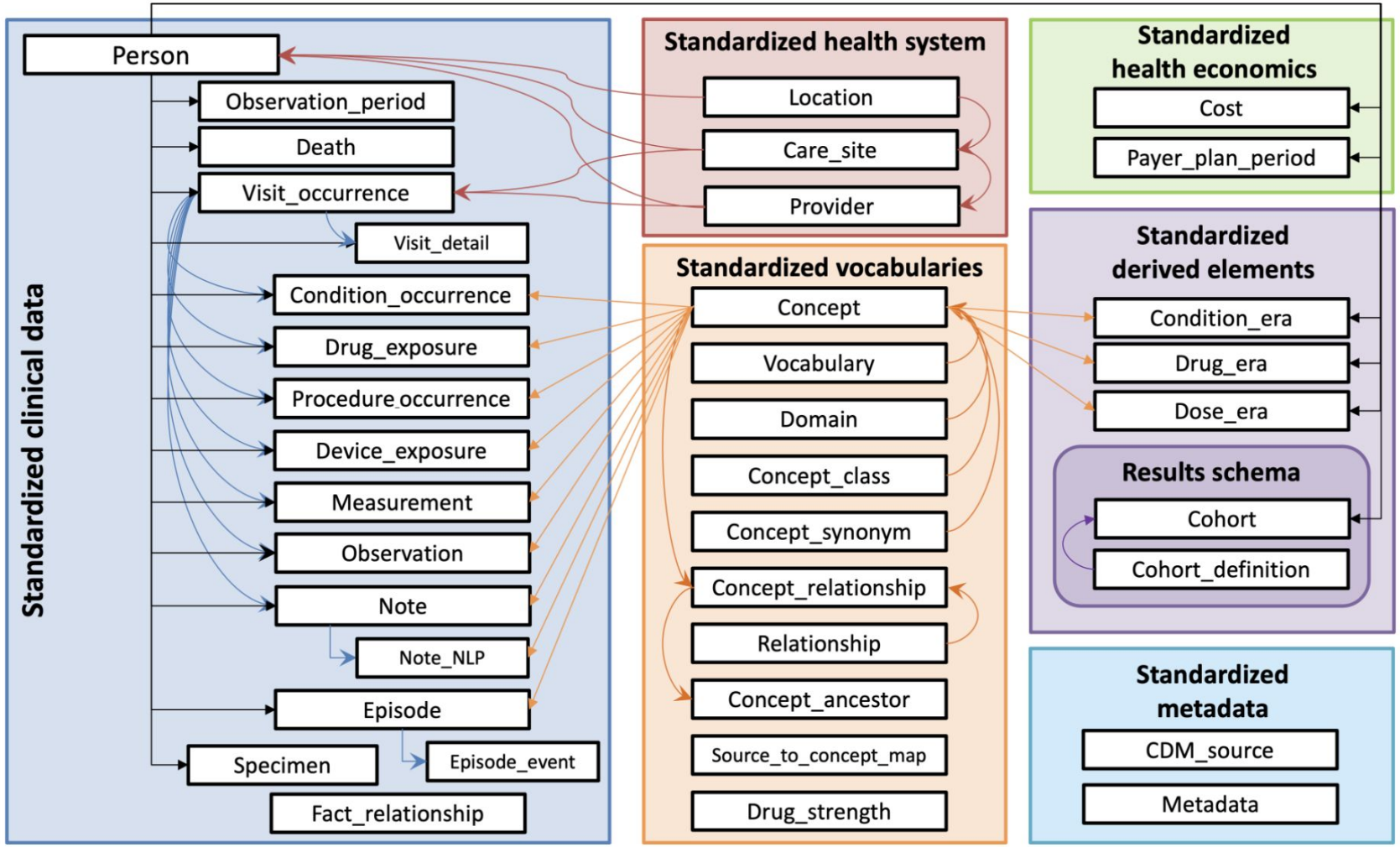
Introduction

Maxim Moinat

The OMOP CDM

Structure and Conventions

OMOP Common Data Model 5.4



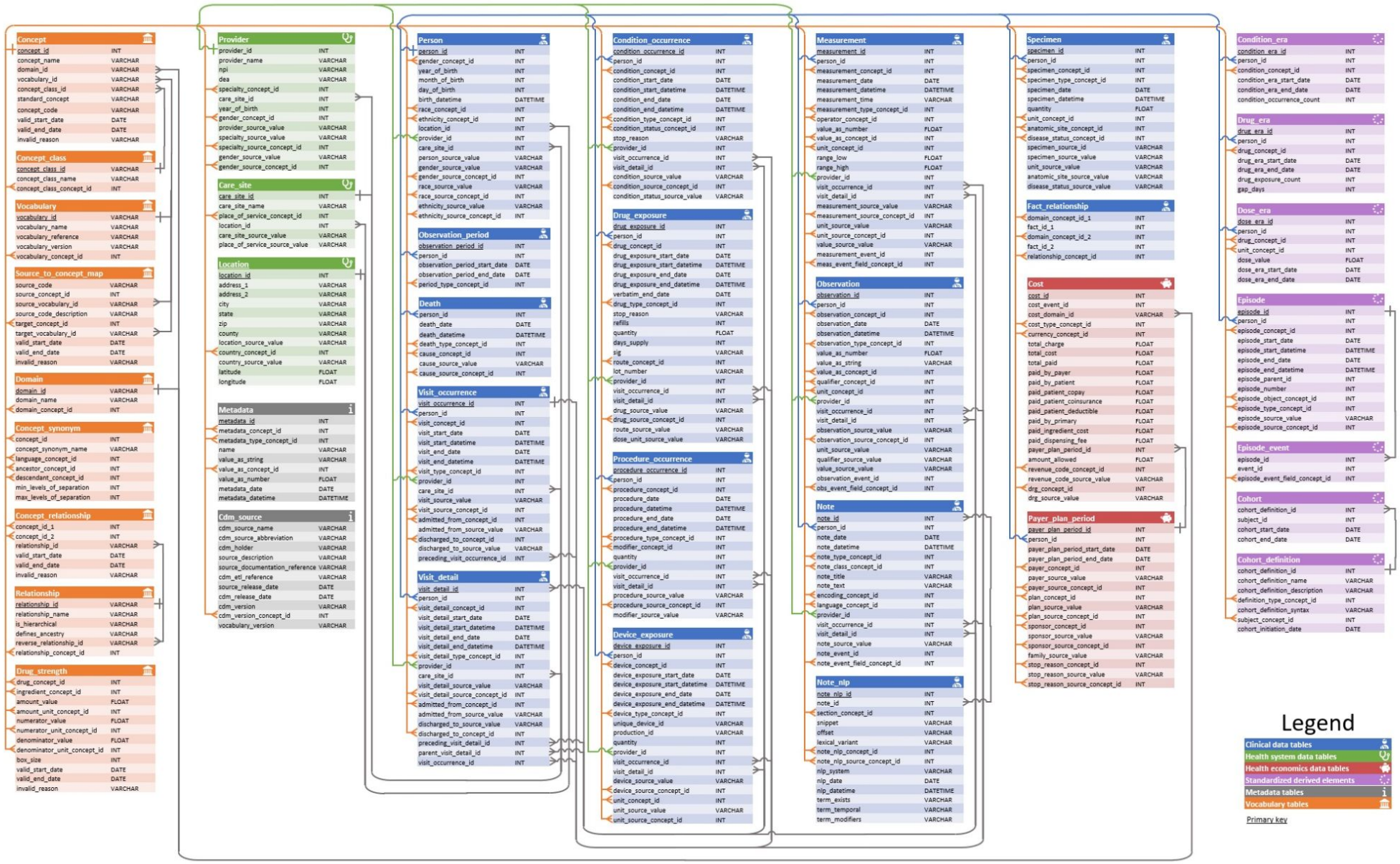
OMOP Principles (1/2)

- Patient centric
- Domain-oriented clinical event tables
  - Condition Occurrence
  - Drug Exposure
  - Measurement
  - …
- Designed for routinely collected longitudinal health data
- Accomodates data from various sources (Primary Care, Hospital Care, Insurers, Population registries, Disease registries)

OMOP Principles (2/2)

- Standard structure
  - Adding local tables/fields is allowed*
- Standard OMOP vocabularies
  - Diagnosis: SNOMED
  - Medication: RxNorm (Extension)
  - Laboratory results: LOINC
  - Procedures: SNOMED
- Preserves provenance and source codes
- Database platform independent

OMOP Common Data Model 5.4



Core OMOP Tables - Mapping order

1

Person

Date of birth, gender, unique (source) key

2

Visit Occurrence

Healthcare encounters (inpatient/outpatient) and specialty involved

?

Observation Period

For what period of time was a person observed?

3

Event tables

Condition, Drug, Procedures
(more details on next slide)

OMOP Event tables

Condition Occurrence

Diagnoses, state

Procedure Occurrence

Imaging, Surgery,

Drug Exposure

Medications prescribed, dispensed and/or administered

Device Exposure

Monitors, Implants, Instruments

Measurement

Vital signs, Laboratory results

Observation

Lifestyle, socio-economic factors

Note

Free text

Episode

Oncology episode, treatment regimen

General clinical event table structure

Field name	Purpose	Example	
<entity>_id	Primary Key	condition_occurrence_id	1234567
person_id	Who - Foreign key to person table	person_id	123
<entity>_start_date <entity>_end_date (	When the event started	condition_start_date	2026-04-18
<entity>_concept_id	What - Standard concept	condition_concept_id	320128 - Essential Hypertension (SNOMED)
<entity>_source_concept_id	What - Source concept	condition_source_concept_id	45591453 - Essential Hypertension (ICD10)
<entity>_source_value	What - Verbatim code	condition_source_value	“I10”
<entity>_type_concept_id	Provenance	condition_type_concept_id	32585 - EHR
visit_occurrence_id	Foreign key to visit_occurrence table	visit_occurrence_id	987654

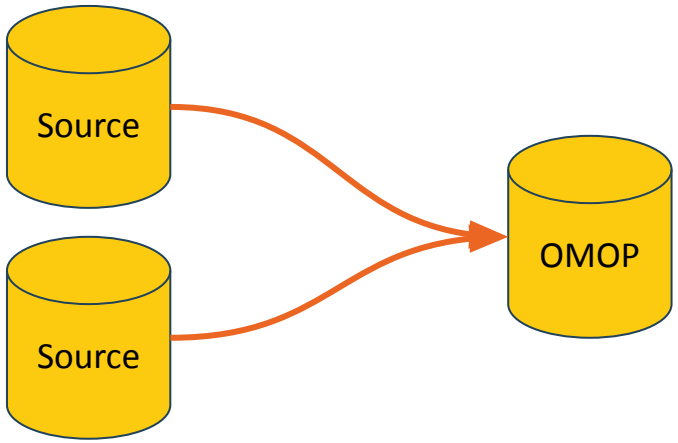
Mapping Principles (1/2)

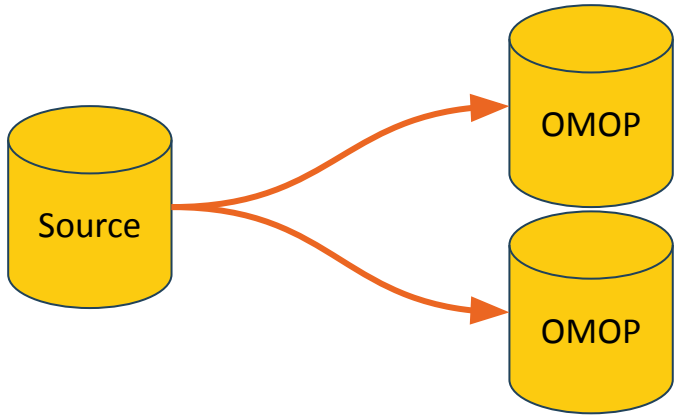
- Required Tables:
  - Person, Observation Period, Cdm Source
- Gender and Year of Birth are required
- Clinical events require at least:
  - person_id
  - *event*_start_date
  - *event*_concept_id.
- Keep original codes in source fields
  - *event*_source_value
  - *event*_source_concept_id

Mapping Principles (2/2)

- All *start_dates* must be <= death_date + 60 days
- If end_date given, star_date must be <= end_date (positive duration)
- All start_dates must be >= birth date
- Historic events or negative findings as observation

Be aware: sources might be combined or split





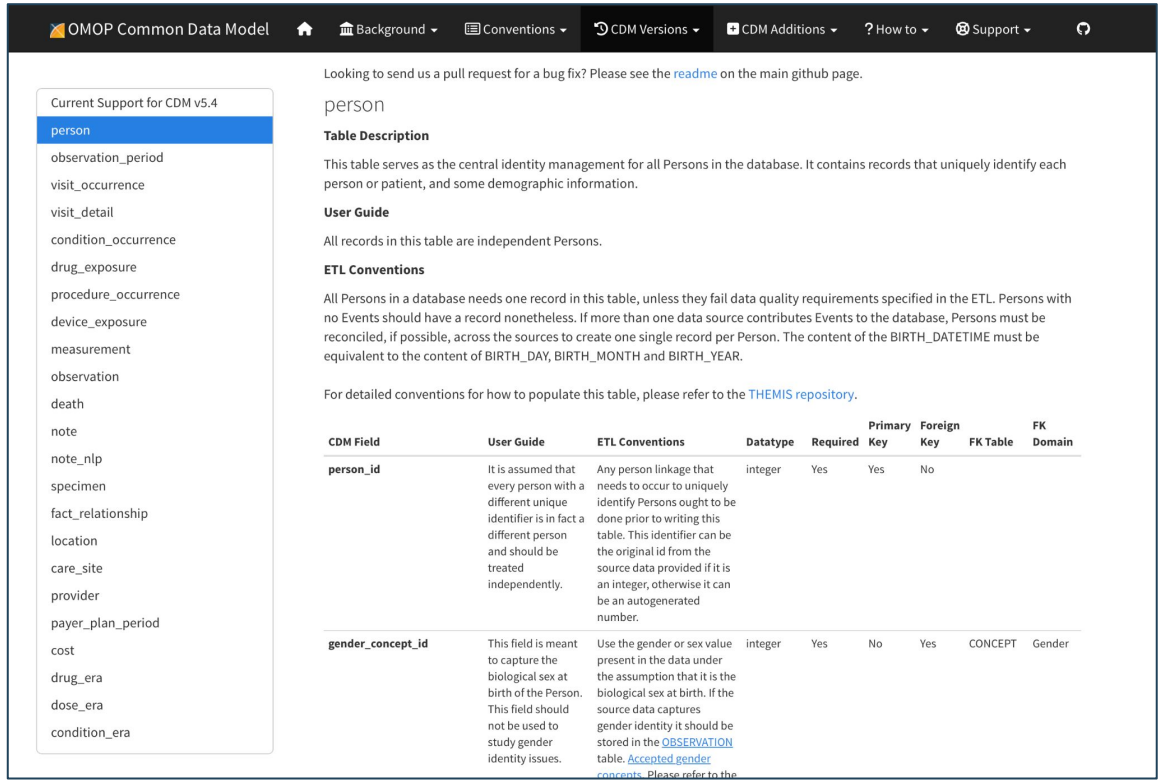
OMOP tables combine data from multiple sources

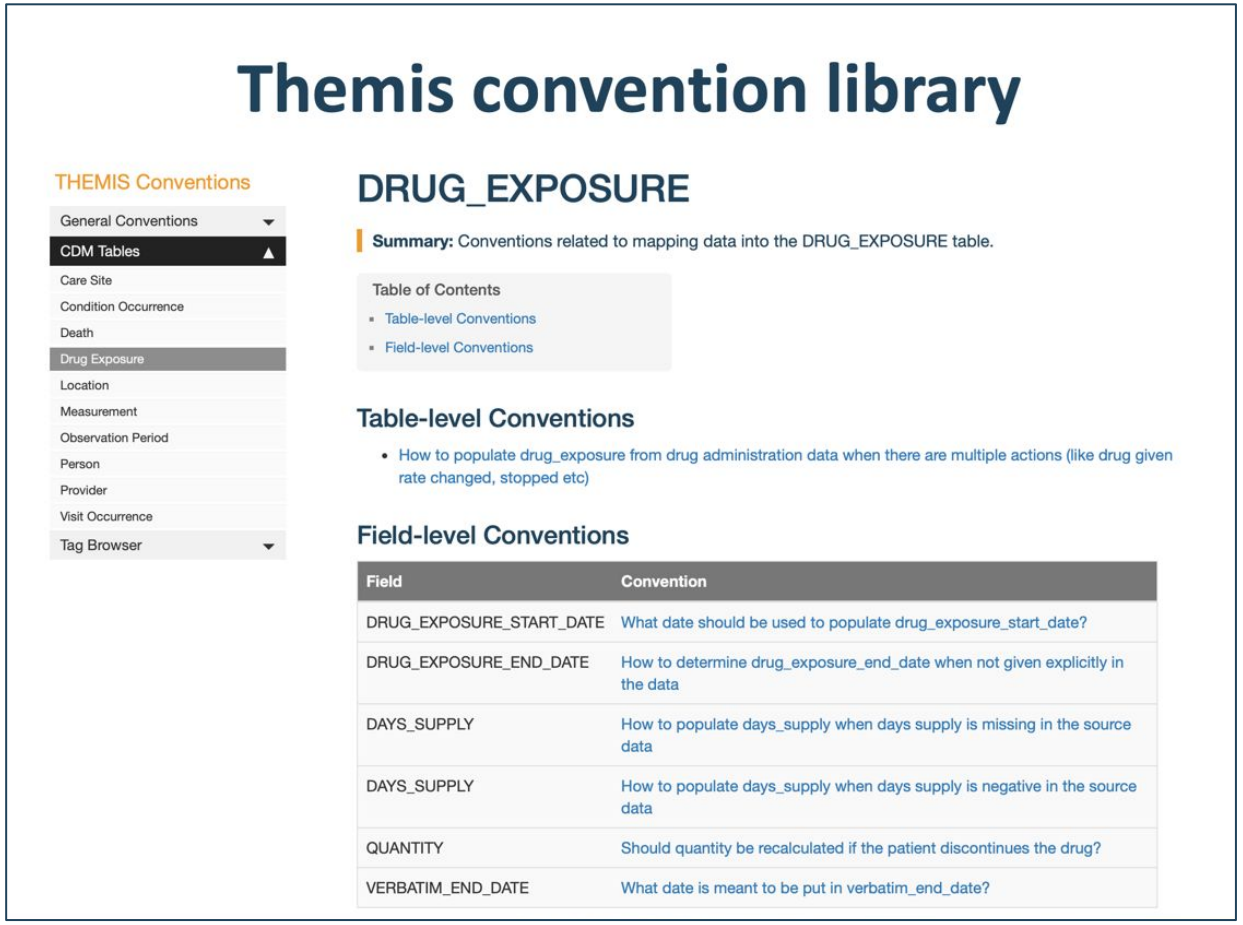
e.g. prescriptions and vaccines all go to drug exposure

One source table might map to multiple OMOP tables.

e.g. ‘treatment’ table maps to drug exposure and procedure occurrence

Where to find the mapping conventions?





https://ohdsi.github.io/CommonDataModel/cdm54.html

https://ohdsi.github.io/Themis/

OMOP CDM v5.5

Workgroup Spotlight: Common Data Model (Mar. 17 Community Call)

- **New tables (vocabulary)**
- Pack_content
- Concept_metadata

- **New columns**
- visit_occurrence_id and visit_detail_id to Specimen table
- value_as_date to observation table
- Add value_source_concept_id to observation and measurement table

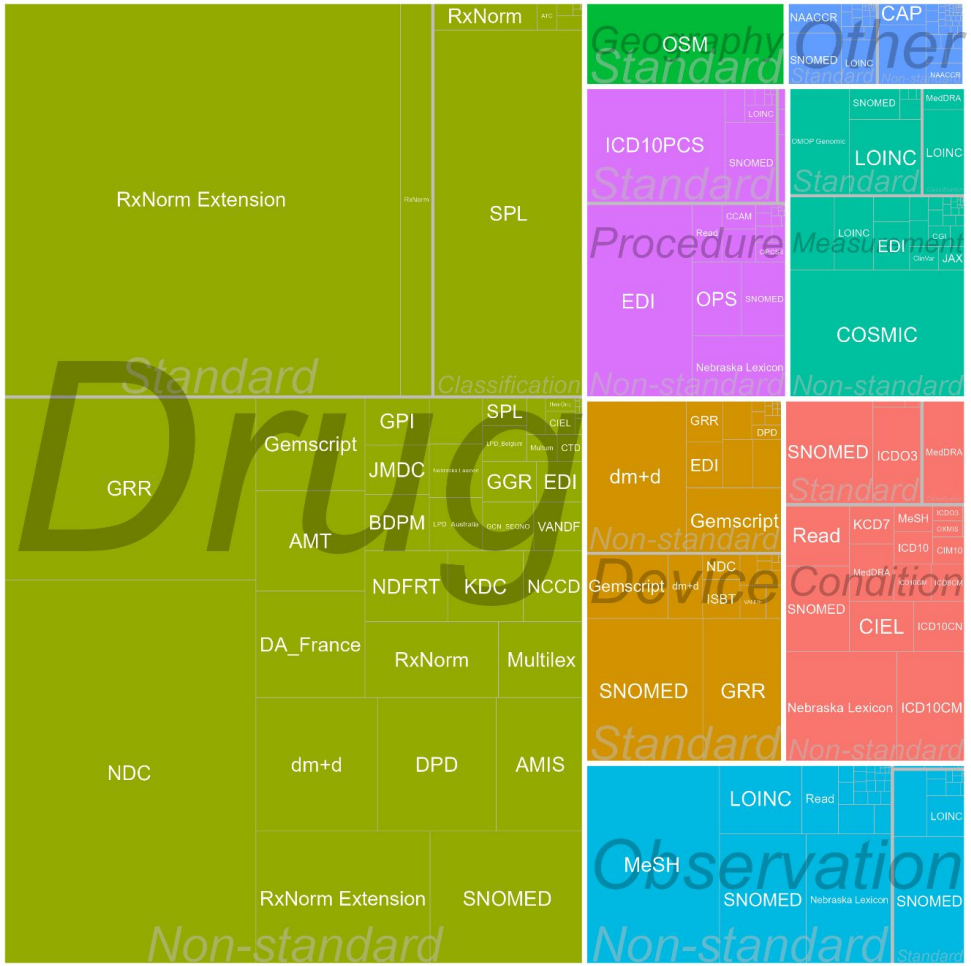
Currently under review, goal mid-2026

The OMOP Vocabularies

The Semantics

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

OMOP Standard Vocabularies

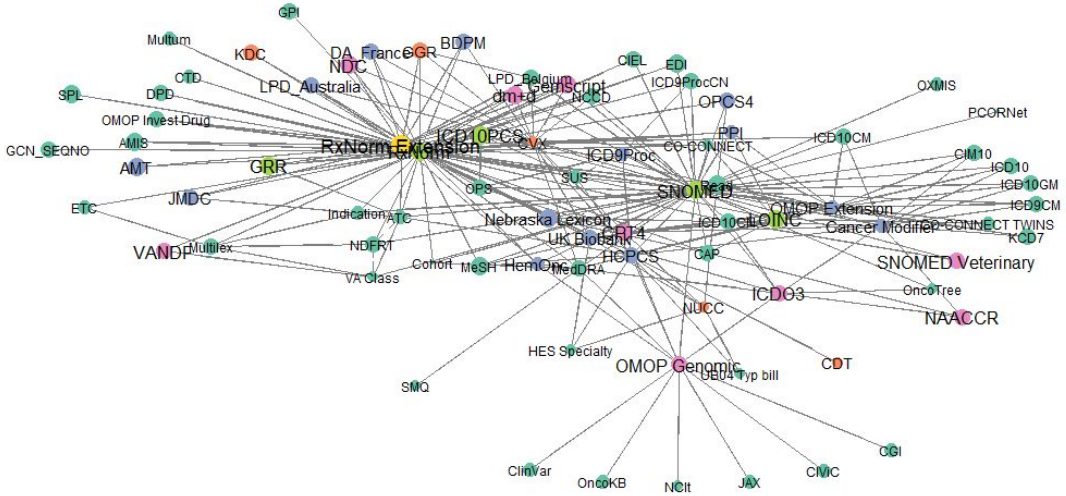


OHDSI Vocabularies By The Numbers

as of August 2023 release

- 11,027,290 concepts
  - 3,598,454 standard concepts
  - 847,008 classification concepts
- 142 vocabularies
- 44 domains
- 82,142,038 concept relationships
- 87,967,689 ancestral relationships
- 4,673,156 concept synonyms

1 Shared Resource to Enable Data Standards



The Source for Source Codes

1. May come from international terminology or code system
   - **SNOMED**
2. May come from a country specific terminology or code system
   - **Read, BDPM, ICD10CN, CVX**
3. May be free text strings
   - **Centimeter, Intravenous, Cigarette Smoker**
4. May come from an EHR specific code system
   - **Epic procedure codes: ‘L111’**

Three categories of OMOP concepts

Non-standard Concept

Function: Unique representation of a source code

Use in:

event_source_concept_id

Standard Concept

Function: Used for standardized analytics and by OHDSI tools

Use in:

event_concept_id

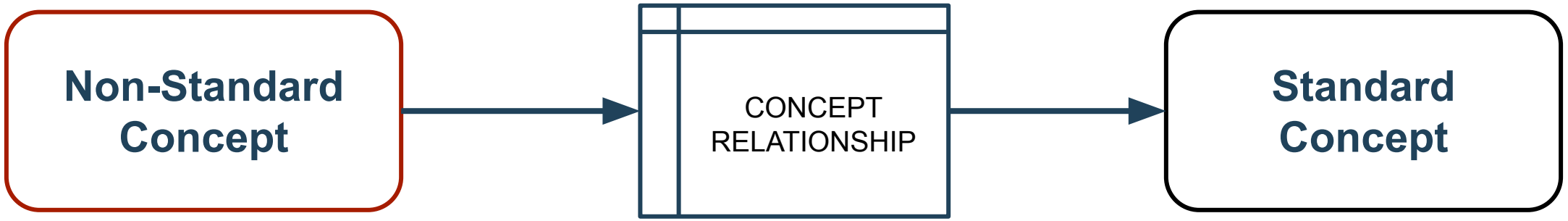
Classification Concept

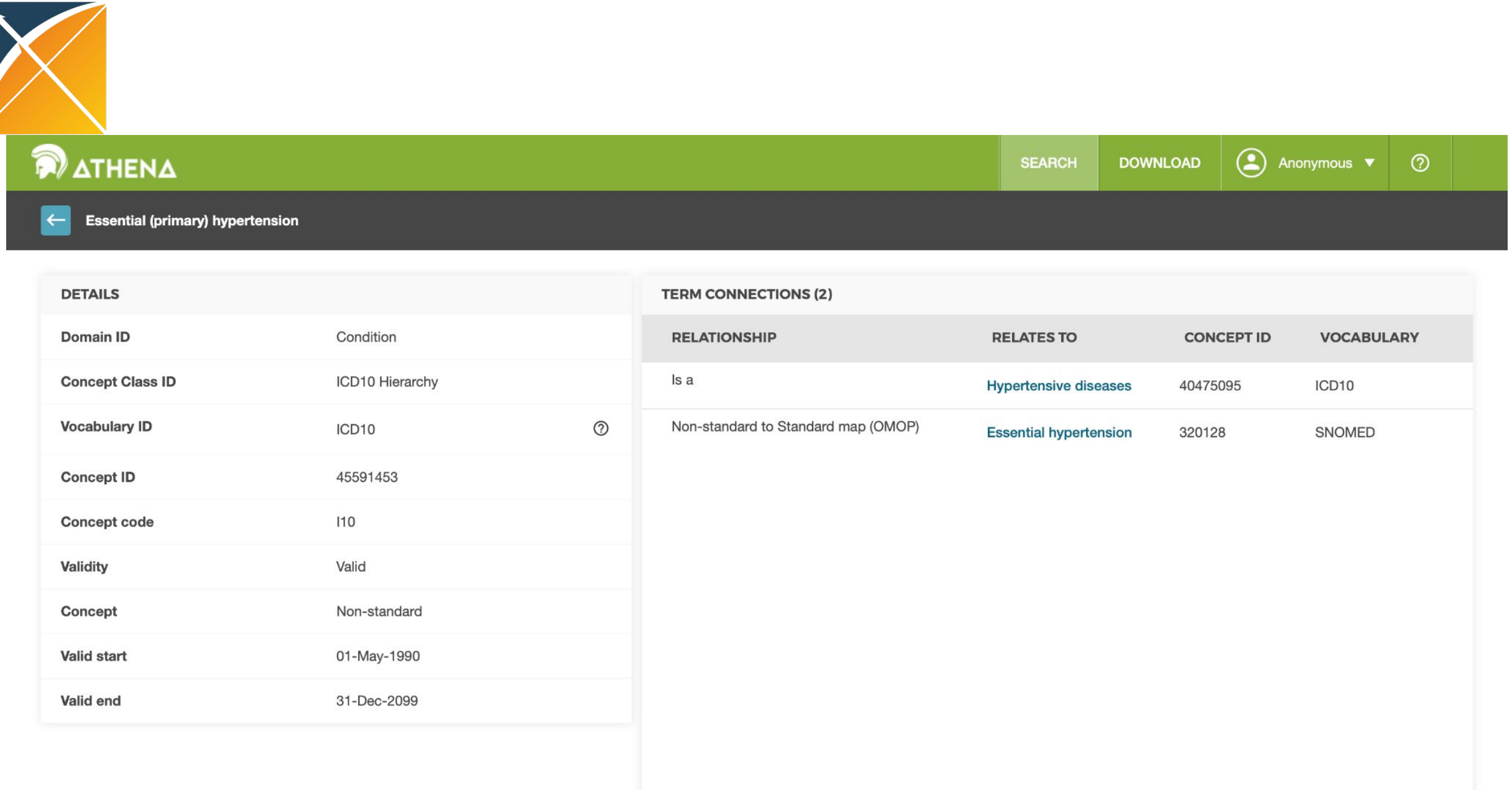
Function: Used to perform hierarchical queries

Use in:

study concept sets

Mapping non-standard concept_ids to standard concept_ids





https://athena.ohdsi.org/search-terms/terms/45591453

Mapping Scenarios

1. Source code available in OMOP Vocabularies
   - Lookup concept_id by concept_code
   - Possibly follow the ‘Maps to’ relationship

```
where <source code> =
CONCEPT.concept_code and
<source vocabulary> =
CONCEPT.vocabulary_id
```

2. Source code is a text string
   - Lookup standard concept by name
   - Bulk mappings and review with tooling, e.g. Usagi

3. Source data does not map to standard OMOP Concept
   - Request new concept(s) through community contribution https://github.com/OHDSI/Vocabulary-v5.0/issues


OHDSI
Observational Health Data Sciences and Informatics

Questions?

Agenda

Introduction	Maxim Moinat
ETL Implementation demo (part 1)	**Liam Glueck**
ETL Implementation demo (part 2)	Anne van Winzum
Q&A / Break	
ETL Execution Considerations	Anne van Winzum
Loading new OMOP vocabulary	Liam Glueck
Release comparison	Maxim Moinat
Q&A / Conventions Quiz	

ETL implementation: Person & Visit Occurrence

Liam Glueck

The example data we will be using:

Synthea™ Synthetic Patient Generator

What is Synthea?

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

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

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

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

Synthea Tables

Tablename	Description
allergies	Patient allergy data.
careplans	Patient care plan data, including goals.
claims	Patient claim data.
claims_transactions	Transactions per line item per claim
conditions	**Patient conditions or diagnoses.**
devices	Patient-affixed permanent and semi-permanent devices.
encounters	**Patient encounter data.**
imaging_studies	Patient imaging metadata.
immunizations	Patient immunization data.
medications	Patient medication data.
observations	**Patient observations including vital signs and lab reports.**
organizations	Provider organizations including hospitals.
patients	**Patient demographic data.**
payer_transitions	Payer Transition data (i.e. changes in health insurance).
payers	Payer organization data.
procedures	Patient procedure data including surgeries.

Our generated Synthea Population

1163 persons

- 616 Female
- 547 Male

Top 10 Conditions	
Gingivitis	Anemia
Viral sinusitis	Impaired glucose tolerance
Acute Bronchitis	Essential Hypertension
Acute Viral Pharyngitis	Chronic Pain
Dental caries	Fracture of Bone

Uses SNOMED
and LOINC as source vocabulary

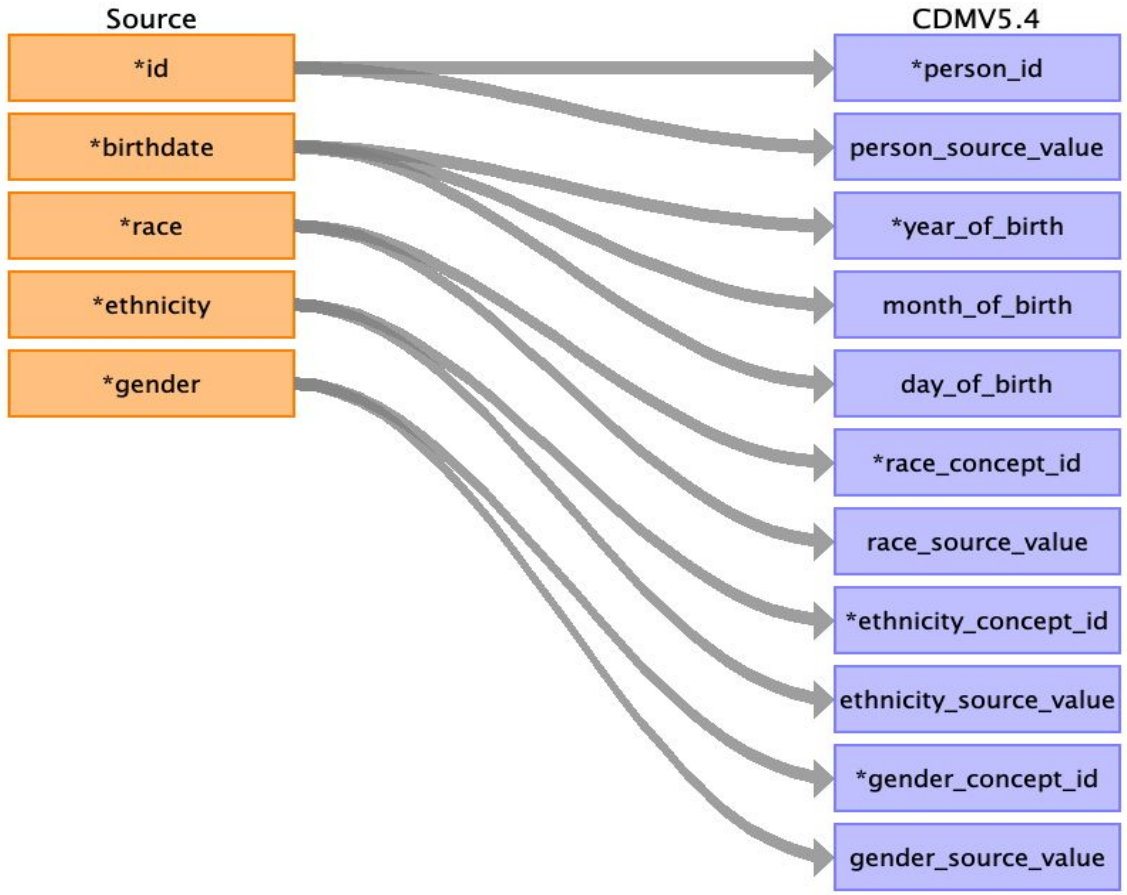
Source table - patients

id uuid	birthdate date	deathdate date	gender text	marital text	race text	ethnicity text	city text
b9c610cd-28a6-4636-ccb6-c7a0d2a…	2019-02-17	[null]	M	[null]	white	nonhispa…	Springfield
c1f1fcaa-82fd-d5b7-3544-c8f9708b0…	2005-07-04	[null]	F	[null]	white	nonhispa…	Bellingham
339144f8-50e1-633e-a013-f361391c…	1998-05-11	[null]	M	[null]	white	nonhispa…	Boston
d488232e-bf14-4bed-08c0-a82f34b6…	2003-01-28	[null]	F	[null]	white	nonhispa…	Hingham
217f95a3-4e10-bd5d-fb67-0cfb5e8b…	1993-12-23	[null]	M	M	black	nonhispa…	Revere
faac724a-a9e9-be66-fe1e-3044dc0b…	2020-12-04	[null]	F	[null]	white	nonhispa…	New Marlborough
23d16ee3-8cd4-eeb8-e77e-1e5fbf4c…	1990-12-15	[null]	M	M	black	hispanic	Revere
aade3c61-92bd-d079-9d28-0b2b7fde…	1985-06-05	[null]	M	M	white	nonhispa…	Somerville
0288c42c-43a1-9878-4a9d-6b96caa…	1979-12-17	[null]	F	M	white	nonhispa…	Cambridge
dc6c06d0-a7d8-100f-c08b-46c93700…	2006-07-11	[null]	M	[null]	white	nonhispa…	Springfield
61a2fcc0-d679-764c-7d86-b885b2c4…	1993-01-22	[null]	F	M	white	nonhispa…	Holbrook
c0219ca9-576f-f7c2-9c44-de030e94…	1971-12-06	[null]	F	M	white	hispanic	Somerville
cb8c092a-99aa-6c9c-1ec1-660f0b00…	1990-12-28	[null]	F	M	white	nonhispa…	Wellesley
97a20cf9-630d-939c-2f50-f13c434a…	1984-03-23	[null]	F	M	white	nonhispa…	Boston
55a6a46e-a1a4-0298-e58f-c0cb27cb…	1958-09-18	2009-11-13	F	M	white	nonhispa…	Quincy
961f61f8-ed32-f113-8450-192064b4…	2002-06-13	[null]	F	[null]	white	nonhispa…	Abington

Source to Concept Map: Race mappings

source_code	source_concept_id	source_vocabulary_id	source_code_description	target_concept_id	target_vocabulary_id	valid_start_date	valid_end_date	invalid_reason
other	0	RACE_MAPPING	other	0	None	1970-01-01	2099-12-31	
hawaiian	0	RACE_MAPPING	hawaiian	1546513	Race	1970-01-01	2099-12-31	
white	0	RACE_MAPPING	white	8527	Race	1970-01-01	2099-12-31	
native	0	RACE_MAPPING	native	1546878	Race	1970-01-01	2099-12-31	
black	0	RACE_MAPPING	black	38003598	Race	1970-01-01	2099-12-31	
asian	0	RACE_MAPPING	asian	8515	Race	1970-01-01	2099-12-31	

Mapping patients to person



Implement with SQL

```sql
INSERT INTO cdm.person (
    person_source_value,
    gender_concept_id,
    gender_source_value,
    year_of_birth,
    month_of_birth,
    day_of_birth,
    race_concept_id,
    race_source_value,
    ethnicity_concept_id
)

SELECT
    ID AS person_source_value,
    CASE
        WHEN GENDER = 'M' THEN 8507
        WHEN GENDER = 'F' THEN 8532
        END AS gender_concept_id,
    GENDER AS gender_source_value,
    EXTRACT(YEAR FROM BIRTHDATE) AS year_of_birth,
    EXTRACT(MONTH FROM BIRTHDATE) AS month_of_birth,
    EXTRACT(DAY FROM BIRTHDATE) AS day_of_birth,
    COALESCE(stcm_race.target_concept_id, 0) AS race_concept_id,
    RACE AS race_source_value,
    COALESCE(stcm_eth.target_concept_id, 0) AS ethnicity_concept_id

FROM native.patients
LEFT JOIN cdm.source_to_concept_map stcm_race
    ON patients.race = stcm.source_code
    AND stcm.source_vocabulary_id = 'RACE_MAPPING'
LEFT JOIN cdm.source_to_concept_map stcm_eth
    ON patients.ethnicity = stcm.source_code
    AND stcm.source_vocabulary_id = 'ETHNICITY_MAPPING'
```

OMOP Person output

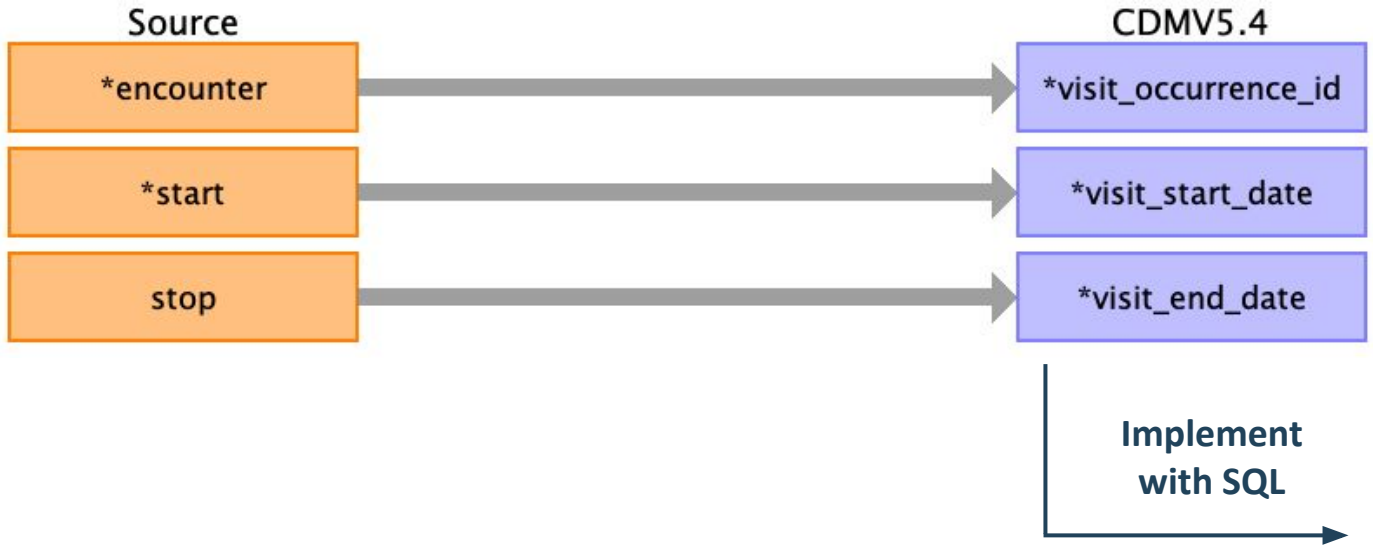
What should the results look like?

person_id [PK] integer	gender_concept_id integer	year_of_birth integer	month_of_birth integer	day_of_birth integer	birth_datetime timestamp without time zone	race_concept_id integer	ethnicity_concept_id integer
2	8507	2019	2	17	[null]	8527	0
3	8532	2005	7	4	[null]	8527	0
4	8507	1998	5	11	[null]	8527	0
5	8532	2003	1	28	[null]	8527	0
6	8507	1993	12	23	[null]	38003598	0
7	8532	2020	12	4	[null]	8527	0
8	8507	1990	12	15	[null]	38003598	0
9	8507	1985	6	5	[null]	8527	0
10	8532	1979	12	17	[null]	8527	0
11	8507	2006	7	11	[null]	8527	0
12	8532	1993	1	22	[null]	8527	0
13	8532	1971	12	6	[null]	8527	0
14	8532	1990	12	28	[null]	8527	0
15	8532	1984	3	23	[null]	8527	0
16	8532	1958	9	18	[null]	8527	0
17	8532	2002	6	13	[null]	8527	0
18	8532	1985	3	24	[null]	8527	0
19	8532	1960	1	21	[null]	8527	0
20	8532	2008	11	9	[null]	8515	0

Source table - encounters

id uuid	start timestamp without time zone	stop timestamp without time zone	patient text	organization text
748f8357-6cc7-551d-f31a-32fa2cf841…	2019-02-17 05:07:38	2019-02-17 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
5a4735ae-423f-6563-28ab-b3d11b49b…	2019-03-24 05:07:38	2019-03-24 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
0bee1ce6-3e2c-5506-f71c-a7ba8f64a3…	2019-05-26 05:07:38	2019-05-26 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
6e93bcf9-45a4-8528-0120-1c1eaa930…	2019-07-28 05:07:38	2019-07-28 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
8b6787c3-4316-a0cb-899d-4746525c3…	2019-10-27 05:07:38	2019-10-27 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
8f424287-ee3a-c144-bc1d-3ba926e93…	2020-01-26 05:07:38	2020-01-26 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
fb15e123-fea7-cae8-6d49-ee9d2a85fc…	2020-02-05 06:07:38	2020-02-05 06:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	5103c940-0c08-392f-95c
01efcc52-15d6-51e9-faa2-bee069fcbe…	2020-02-17 10:07:38	2020-02-17 10:40:32	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	5103c940-0c08-392f-95c
1a7debfc-9582-7f23-a109-4f154a182e…	2020-04-26 05:07:38	2020-04-26 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
bf38c711-941f-7509-f9ec-b864d6929f3f	2020-07-26 05:07:38	2020-07-26 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
7bb78da2-31b8-497d-bc08-25eca8a90…	2021-01-24 05:07:38	2021-01-24 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
53c2bae0-f0ff-7eac-4ca1-3dce0ecebb…	2021-07-25 05:07:38	2021-07-25 05:22:38	b9c610cd-28a6-4636-ccb6-c7a0d2a4cb…	f7ae497d-8dc6-3721-940
c7b935e1-02ce-8ec2-bc48-69671e986…	2012-06-25 08:20:52	2012-06-25 08:35:52	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	620d4a8a-ef3b-30c4-9c5
0b2794bd-ec2b-d34f-0610-2523b3b7f…	2013-06-24 06:20:52	2013-06-24 06:39:19	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	24cb4eab-6166-3530-bd
7d6494ff-bf75-6120-dcc9-e73db03479…	2013-07-01 08:20:52	2013-07-01 08:35:52	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	620d4a8a-ef3b-30c4-9c5
6bdd1328-49b4-b73c-1495-db6e46e0a…	2013-07-08 08:20:52	2013-07-08 08:35:52	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	b06770b5-b695-3c1a-968
34894bce-59ce-59c6-7f0e-96dc4100df…	2014-07-07 08:20:52	2014-07-07 08:35:52	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	620d4a8a-ef3b-30c4-9c5
f4357834-ba77-f8a7-6b64-9b1809bf13…	2015-07-13 08:20:52	2015-07-13 08:35:52	c1f1fcaa-82fd-d5b7-3544-c8f9708b06a8	620d4a8a-ef3b-30c4-9c5

Mapping encounters to visit_occurrence




```sql
INSERT INTO cdm.visit_occurrence (
    person_id,
    visit_concept_id,
    visit_start_date,
    visit_end_date,
    visit_type_concept_id,
)
SELECT
    p.person_id AS person_id,

    9201 AS visit_concept_id, -- Concept for 'Inpatient visit'
    visit_start_date e.start::date AS visit_start_date,
    visit_end_date COALESCE(e.stop, e.start)::date AS visit_end_date,

    -- Always map as "EHR"
    32817 AS visit_type_concept_id,

FROM source_data.encounters e
LEFT JOIN cdm.person p
    ON e.patient = p.person_source_value
WHERE e.start IS NOT NULL;
```

Visit Occurrence Output

visit_occurrence_id [PK] integer	person_id integer	visit_concept_id integer	visit_start_date date	visit_end_date date	visit_type_concept_id integer
563132	2328	9201	2019-02-17	2019-02-17	32817
563133	2328	9201	2019-03-24	2019-03-24	32817
563134	2328	9201	2019-05-26	2019-05-26	32817
563135	2328	9201	2019-07-28	2019-07-28	32817
563136	2328	9201	2019-10-27	2019-10-27	32817
563137	2328	9201	2020-01-26	2020-01-26	32817
563138	2328	9201	2020-02-05	2020-02-05	32817
563139	2328	9201	2020-02-17	2020-02-17	32817
563140	2328	9201	2020-04-26	2020-04-26	32817
563141	2328	9201	2020-07-26	2020-07-26	32817
563142	2328	9201	2021-01-24	2021-01-24	32817
563143	2328	9201	2021-07-25	2021-07-25	32817
563144	2329	9201	2012-06-25	2012-06-25	32817
563145	2329	9201	2013-06-24	2013-06-24	32817
563146	2329	9201	2013-07-01	2013-07-01	32817
563147	2329	9201	2013-07-08	2013-07-08	32817

Missing visit occurrence end date

"...For Inpatient Visits ongoing at the date of ETL, put date of processing the data into visit_end_datetime and visit_type_concept_id with 32220 "Still patient" to identify the visit as incomplete. - All other Visits: visit_end_datetime = visit_start_datetime. If this is a one-day visit the end date should match the start date."

https://ohdsi.github.io/CommonDataModel/cdm54.html#visit_occurrence -> visit_end_date ETL convention

ETL implementation: Condition Occurrence, Stem Table, CDM Source

Anne van Winzum

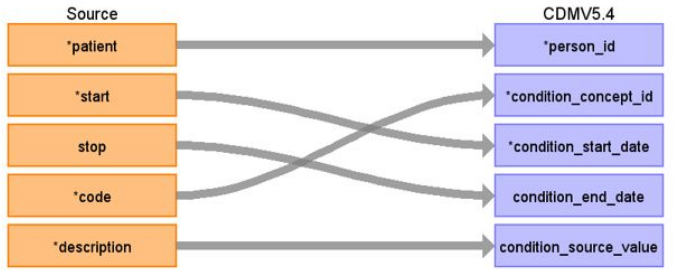
Source table - Conditions

patient	start	stop	encounter	code	description	system
39e43e57-0074-9010-687a-72feab636f44	20/06/2015	27/06/2015	da3e20ac-726c-6c2f-0b98-071c8790eac7	444814009	Viral sinusitis (disorder)	http://snomed.info/sct
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	15/08/2015	29/08/2015	9e6fb82d-27c3-47df-59c3-f728971c76a8	66383009	Gingivitis (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	12/10/2016	28/12/2016	04c17818-3c57-164a-84c8-24b576ca8eee	125605004	Fracture of bone (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	12/10/2016	28/12/2016	04c17818-3c57-164a-84c8-24b576ca8eee	58150001	Fracture of clavicle (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	17/06/2015	17/06/2015	e1e70381-7f74-0d0d-adde-934fc19ff17f	109570002	Primary dental caries (disorder)	http://snomed.info/sct
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	02/03/2015	02/03/2015	bcd4ff75-1a31-d255-a334-41a101f69f7b	109570002	Primary dental caries (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	11/08/2017	11/08/2017	a1ab0bc1-1a6f-536e-8202-2e480509756a	307426000	Acute infective cystitis (disorder)	http://snomed.info/sct
6faea64b-7045-77b3-f9fe-3710368f8525	01/02/2017	24/07/2017	ba20487d-80d0-bde6-93b8-6a3fee39982d	65363002	Otitis media (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	27/06/2018	04/07/2018	4e4e51b1-0bdf-6237-bbfe-883923006e4f	66383009	Gingivitis (disorder)	http://snomed.info/sct
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	09/09/2018	21/09/2018	cc947428-ab9e-f3d9-a9e1-d0f51954855c	195662009	Acute viral pharyngitis (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	22/10/2019	09/11/2019	389714a0-304e-8853-e9e8-4732c1be6c63	312608009	Laceration - injury (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	22/10/2019	09/11/2019	389714a0-304e-8853-e9e8-4732c1be6c63	284549007	Laceration of hand (disorder)	http://snomed.info/sct
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	07/03/2016	07/03/2016	81c15bde-d8a6-6859-b8e6-a58ff71aa15c	427898007	Infection of tooth (disorder)	http://snomed.info/sct
6faea64b-7045-77b3-f9fe-3710368f8525	20/11/2017	09/12/2017	f1d3b484-e842-db44-ae10-cc325b70fa69	10509002	Acute bronchitis (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	03/07/2019	10/07/2019	932f1044-3be8-8720-5880-821a69329d3f	66383009	Gingivitis (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	08/09/2022	22/09/2022	434be880-15e6-03c6-4d6e-7ddbab270e5c	195662009	Acute viral pharyngitis (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	10/07/2019	10/07/2019	59f2364d-4c20-149e-7e4f-1a1609be95de	18718003	Gingival disease (disorder)	http://snomed.info/sct
6faea64b-7045-77b3-f9fe-3710368f8525	20/05/2018	28/01/2019	7e8c5e4e-1131-e5b2-1e19-fc6bd7f19706	65363002	Otitis media (disorder)	http://snomed.info/sct
39e43e57-0074-9010-687a-72feab636f44	19/03/2024	02/04/2024	0702e6e2-0f67-a656-96f1-075ef77c5d9a	444814009	Viral sinusitis (disorder)	http://snomed.info/sct
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	26/09/2020	26/09/2020	9d6f94b1-81e3-0129-e28d-c44d5dac715a	427898007	Infection of tooth (disorder)	http://snomed.info/sct
6faea64b-7045-77b3-f9fe-3710368f8525	10/09/2021	21/09/2021	1be8a2d9-aa2e-188d-c93f-fc062de81dcb	43878008	Streptococcal sore throat (disorder)	http://snomed.info/sct
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	29/03/2021	12/04/2021	3cc01e8a-5d83-b1f2-b689-6d62e65151f9	66383009	Gingivitis (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	03/08/2022	03/08/2022	cb83e20b-f02d-1e46-9b87-895b271c6663	109570002	Primary dental caries (disorder)	http://snomed.info/sct
6faea64b-7045-77b3-f9fe-3710368f8525	20/02/2023	27/02/2023	d2f6cc87-9c93-45a9-50f7-3547659161ef	66383009	Gingivitis (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	16/11/2022	28/11/2022	5e53994c-403f-882d-e79f-b73ab842c3eb	195662009	Acute viral pharyngitis (disorder)	http://snomed.info/sct
3adf9f94-9499-d221-c715-11bf40ec757e	26/07/2023	09/08/2023	366f613a-fb7b-4f6e-3aac-828fe184af5e	66383009	Gingivitis (disorder)	http://snomed.info/sct

Mapping implementation

Table name: condition_occurrence

Reading from conditions



Destination Field	Source Field	Logic	Comment
condition_occurrence_id			autoincrement
person_id	patient		
condition_concept_id	code		snomed concept code, map to concept_id
condition_start_date	start		
condition_start_datetime			
condition_end_date	stop		if stop is not empty
condition_end_datetime			
condition_type_concept_id			32817 - EHR
condition_status_concept_id			
stop_reason			
provider_id			
visit_occurrence_id			visit_occurrence.visit_occurrence_id
visit_detail_id			
condition_source_value	description		
condition_source_concept_id			

```
INSERT INTO cdm.condition_occurrence
(
    person_id,
    condition_concept_id,
    condition_start_date,
    condition_end_date,
    condition_type_concept_id,
    condition_source_value
)

SELECT
    person.person_id,
    concept.concept_id AS condition_concept_id,
    start AS condition_start_date,
    stop AS condition_end_date,
    32817 AS condition_type_concept_id,
    description AS condition_source_value

FROM native.conditions AS conditions
LEFT JOIN cdm.concept AS concept
    ON code = concept.concept_code
LEFT JOIN cdm.person AS person
    ON conditions.patient = person.person_source_value
WHERE vocabulary_id = 'SNOMED';
```

```
cr.concept_id_2 a
concept_id as con
```

```
JOIN cdm.concept_re
    ON concept_id =
    AND relationshi
```

OMOP table - Condition

condition_occurrence_id [PK] integer	person_id integer	condition_concept_id integer	condition_start_date date	condition_end_date date	condition_type_concept_id integer	condition_source_value character varying (250)
547	1	4144035	2017-08-11	2017-08-11	32817	Acute infective cystitis (disorder)
106	1	40481087	2024-03-19	2024-04-02	32817	Viral sinusitis (disorder)
554	1	443419	2019-10-22	2019-11-09	32817	Laceration - injury (disorder)
85	1	4112343	2022-09-08	2022-09-22	32817	Acute viral pharyngitis (disorder)
1	1	40481087	2015-06-20	2015-06-27	32817	Viral sinusitis (disorder)
589	1	4113008	2019-10-22	2019-11-09	32817	Laceration of hand (disorder)
15	1	75053	2016-10-12	2016-12-28	32817	Fracture of bone (disorder)
22	1	4237458	2016-10-12	2016-12-28	32817	Fracture of clavicle (disorder)
162	2	40274283	2024-08-14	2024-08-14	32817	Primary dental caries (disorder)
141	2	4112343	2022-11-16	2022-11-28	32817	Acute viral pharyngitis (disorder)
92	2	4090111	2019-07-10	2019-07-10	32817	Gingival disease (disorder)
29	2	40274283	2015-06-17	2015-06-17	32817	Primary dental caries (disorder)
127	2	40274283	2022-08-03	2022-08-03	32817	Primary dental caries (disorder)
148	2	4281516	2023-07-26	2023-08-09	32817	Gingivitis (disorder)
78	2	4281516	2019-07-03	2019-07-10	32817	Gingivitis (disorder)
50	2	4281516	2018-06-27	2018-07-04	32817	Gingivitis (disorder)

Source table - Observations

patient	date	encounter	code	description	value	units	type	category
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	8302-2	Body Height	115.7	cm	numeric	vital-signs
6faea64b-7045-77b3-f9fe-3710368f8525	26/01/2015	47378e02-76f0-2bd8-56c6-15f330f50ce0	8302-2	Body Height	76.6	cm	numeric	vital-signs
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	15/08/2015	9e6fb82d-27c3-47df-59c3-f728971c76a8	8302-2	Body Height	141.4	cm	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	72514-3	Pain severity - 0-10 verbal numeric rating [Score] - Reported	4.0	{score}	numeric	vital-signs
3adf9f94-9499-d221-c715-11bf40ec757e	10/06/2015	427fa12d-35a0-53e5-247a-5e9bd78561c0	8302-2	Body Height	123.0	cm	numeric	vital-signs
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	15/08/2015	9e6fb82d-27c3-47df-59c3-f728971c76a8	72514-3	Pain severity - 0-10 verbal numeric rating [Score] - Reported	2.0	{score}	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	29463-7	Body Weight	22.2	kg	numeric	vital-signs
6faea64b-7045-77b3-f9fe-3710368f8525	26/01/2015	47378e02-76f0-2bd8-56c6-15f330f50ce0	72514-3	Pain severity - 0-10 verbal numeric rating [Score] - Reported	1.0	{score}	numeric	vital-signs
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	15/08/2015	9e6fb82d-27c3-47df-59c3-f728971c76a8	29463-7	Body Weight	48.2	kg	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	39156-5	Body mass index (BMI) [Ratio]	16.6	kg/m2	numeric	vital-signs
3adf9f94-9499-d221-c715-11bf40ec757e	10/06/2015	427fa12d-35a0-53e5-247a-5e9bd78561c0	72514-3	Pain severity - 0-10 verbal numeric rating [Score] - Reported	4.0	{score}	numeric	vital-signs
3c3a542d-8574-b9c0-8ef5-fe1a0c43dd25	15/08/2015	9e6fb82d-27c3-47df-59c3-f728971c76a8	72166-2	Tobacco smoking status	Never smoked tobacco (finding)	NULL	text	social-history
6faea64b-7045-77b3-f9fe-3710368f8525	26/01/2015	47378e02-76f0-2bd8-56c6-15f330f50ce0	8480-6	Systolic Blood Pressure	107.0	mm[Hg]	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	9279-1	Respiratory rate	13.0	/min	numeric	vital-signs
3adf9f94-9499-d221-c715-11bf40ec757e	10/06/2015	427fa12d-35a0-53e5-247a-5e9bd78561c0	8867-4	Heart rate	70.0	/min	numeric	vital-signs
39e43e57-0074-9010-687a-72feab636f44	30/03/2017	ed42eb48-bd04-38ef-572c-bfbe812713e5	8867-4	Heart rate	68.0	/min	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	6690-2	Leukocytes [#/volume] in Blood by Automated count	4.2	10*3/uL	numeric	laboratory
6faea64b-7045-77b3-f9fe-3710368f8525	26/01/2015	47378e02-76f0-2bd8-56c6-15f330f50ce0	8867-4	Heart rate	83.0	/min	numeric	vital-signs
3adf9f94-9499-d221-c715-11bf40ec757e	10/06/2015	427fa12d-35a0-53e5-247a-5e9bd78561c0	9279-1	Respiratory rate	15.0	/min	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	789-8	Erythrocytes [#/volume] in Blood by Automated count	4.0	10*6/uL	numeric	laboratory
3adf9f94-9499-d221-c715-11bf40ec757e	10/06/2015	427fa12d-35a0-53e5-247a-5e9bd78561c0	72166-2	Tobacco smoking status	Never smoked tobacco (finding)	NULL	text	social-history
6faea64b-7045-77b3-f9fe-3710368f8525	26/01/2015	47378e02-76f0-2bd8-56c6-15f330f50ce0	9279-1	Respiratory rate	15.0	/min	numeric	vital-signs
39e43e57-0074-9010-687a-72feab636f44	30/03/2017	ed42eb48-bd04-38ef-572c-bfbe812713e5	9279-1	Respiratory rate	15.0	/min	numeric	vital-signs
d671420e-faa1-eb25-5a1f-4b05df7c2cf6	23/02/2015	6e2af486-7114-033a-017b-b52bbca187e9	718-7	Hemoglobin [Mass/volume] in Blood	15.0	g/dL	numeric	laboratory

Mix of clinical OMOP domains

Body height

DETAILS

Domain ID	Measurement
Concept Class ID	Clinical Observation
Vocabulary ID	LOINC
Concept ID	3036277
Concept code	8302-2
Validity	Valid
Concept	Standard

LANGUAGE	SYNONYM CONCEPT
English	AOEObservation; Axial length; bod; Bodies; BODY HEIGHT(LENGTH).ATOM; Body length; Length; Point in time; QNT; Quan; Quant; Quantitative; Random
Chinese	身高:长度:时间点:^患者:定量型

Valid start	06-Sep-1996
Valid end	31-Dec-2099

Smoking Status

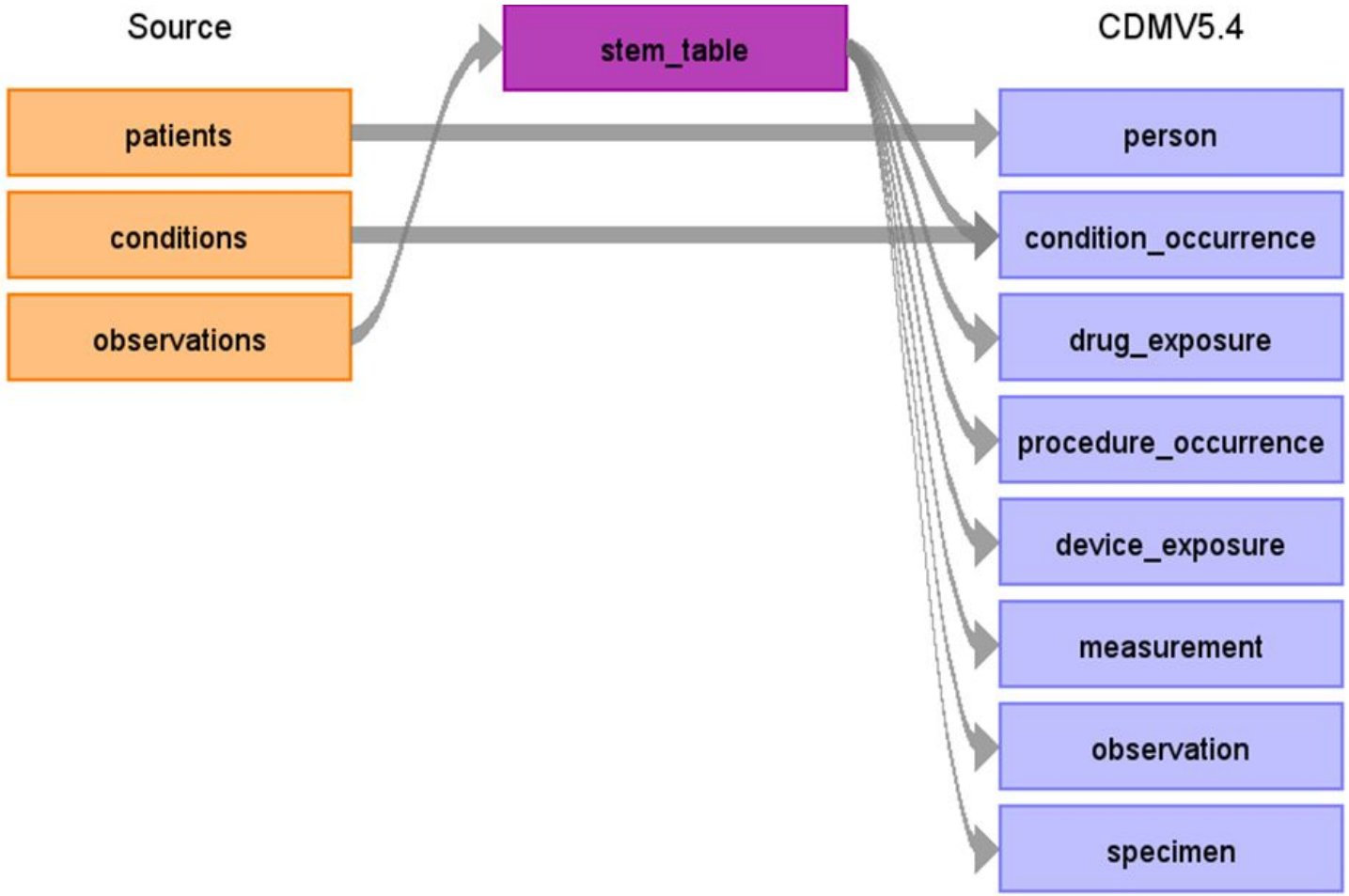
DETAILS

Domain ID	Observation
Concept Class ID	Clinical Observation
Vocabulary ID	LOINC
Concept ID	43054909
Concept code	72166-2
Validity	Valid
Concept	Standard

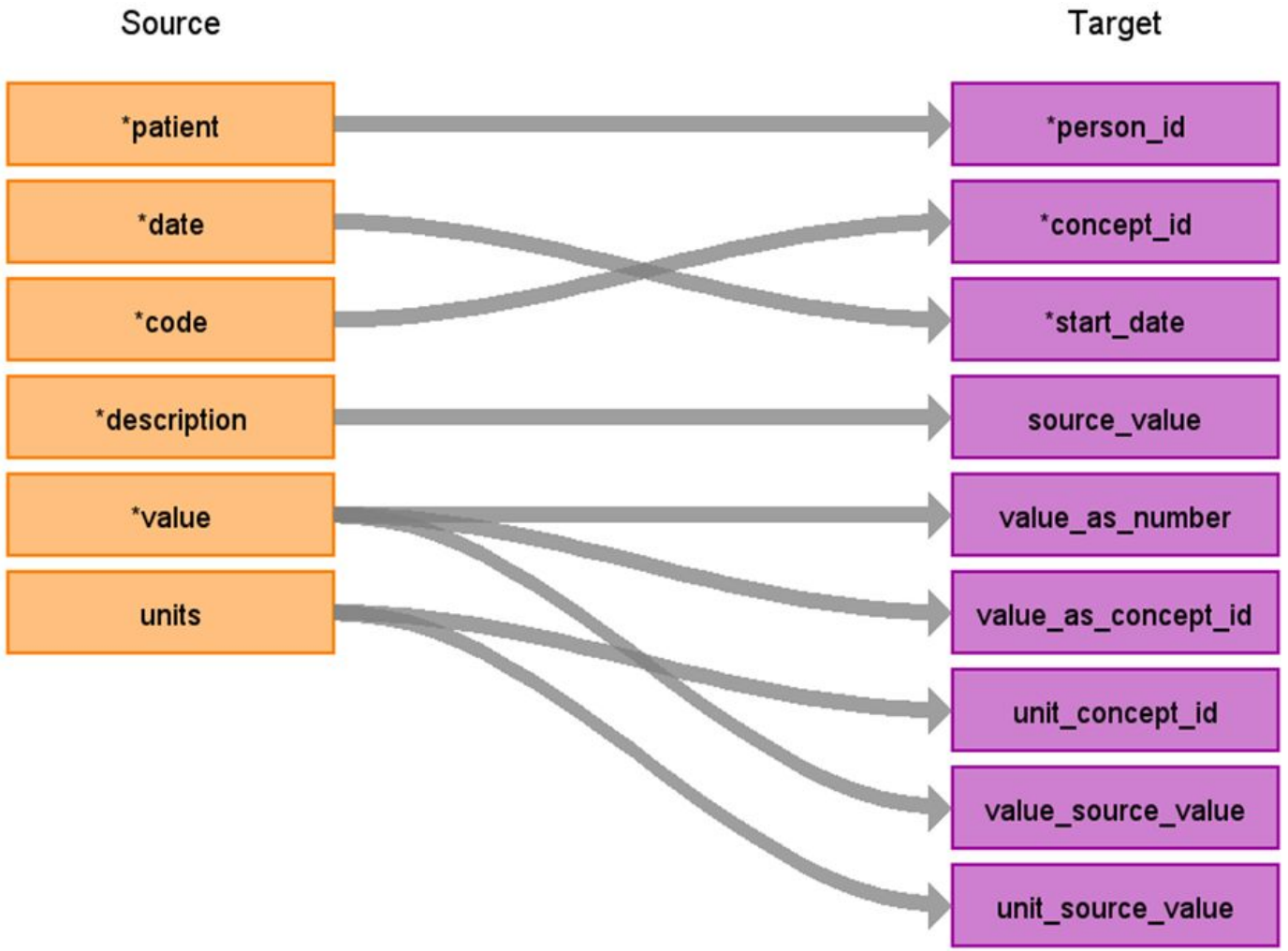
LANGUAGE	SYNONYM CONCEPT
English	Finding; Findings; H+P; H+P.HX; Nominal; P prime; Point in time; Random; Tobac smoke stat
English	Tobac smoke stat
Chinese	烟草吸食状况:发现:时间点:^患者:名义型

Valid start	28-Dec-2012
Valid end	31-Dec-2099

A solution: map to an intermediate table



And map intermediate to domains based on target concept



```sql
INSERT INTO cdm.stem_table
(
    domain_id,
    person_id,
    concept_id,
    start_date,
    type_concept_id,
    source_value,
    value_as_concept_id,
    unit_concept_id,
    unit_source_value,
    value_as_number,
    value_source_value
)

SELECT
    concept.domain_id AS domain_id,
    person.person_id,
    concept.concept_id AS concept_id,
    date AS start_date,
    32817 AS type_concept_id,
    description AS source_value,
    stcm_value.target_concept_id AS value_concept_id,
    stcm_unit.target_concept_id AS unit_concept_id,
    units AS unit_source_value,
    NULLIF(regexp_replace(value, '[^\.\d]','','g'), '')::numeric AS value_as_number,
    value AS value_source_value
FROM native.observations
LEFT JOIN cdm.concept AS concept
    ON code = concept.concept_code
LEFT JOIN cdm.person AS person
    ON observations.patient = person.person_source_value
LEFT JOIN cdm.source_to_concept_map AS stcm_value
    ON value = stcm_value.source_code
   AND stcm_value.source_vocabulary_id = '2'
LEFT JOIN cdm.source_to_concept_map AS stcm_unit
    ON units = stcm_unit.source_code
   AND stcm_unit.source_vocabulary_id = '1';
```

Measurement/Observation/…

STEM Table (intermediate)

domain_id character varying (20)	person_id integer	concept_id integer	start_date date	type_concept_id integer	value_as_number numeric	value_as_concept_id integer	unit_concept_id integer	unit_source_value character varying (50)	source_value character varying (250)	unit_source_value character varying (50)	value_source_value character varying (250)
Observation	1	3046853	2017-03-30	32817	[null]	[null]	[null]	NULL	Race	NULL	White
Measurement	1	3015182	2017-03-30	32817	40.5	[null]	8583	fL	Erythrocyte distribution width [Entitic vol...	fL	40.5
Measurement	1	3038553	2017-03-30	32817	28.6	[null]	9531	kg/m2	Body mass index (BMI) [Ratio]	kg/m2	28.6
Measurement	1	3043111	2017-03-30	32817	9.6	[null]	8583	fL	Platelet mean volume [Entitic volume] in ...	fL	9.6
Observation	1	37020846	2017-03-30	32817	[null]	45878245	[null]	NULL	At any point in the past 2 years has seas...	NULL	No
Observation	1	40766311	2017-03-30	32817	33375	[null]	[null]	/a	What was your best estimate of the total...	/a	33375
Measurement	1	3009744	2017-03-30	32817	34.3	[null]	8713	g/dL	MCHC [Mass/volume] by Automated cou...	g/dL	34.3
Observation	1	43054909	2017-03-30	32817	[null]	45883458	[null]	NULL	Tobacco smoking status	NULL	Ex-smoker (finding)
Observation	1	40758879	2017-03-30	32817	0.0	[null]	44777566	{score}	Patient Health Questionnaire 2 item (PH...	{score}	0.0
Measurement	1	3000905	2017-03-30	32817	4.4	[null]	44777519	10*3/uL	Leukocytes [#/volume] in Blood by Auto...	10*3/uL	4.4
Observation	1	37020172	2017-03-30	32817	[null]	45878245	[null]	NULL	Are you worried about losing your housin...	NULL	No
Observation	1	40770471	2017-03-30	32817	[null]	37079092	[null]	NULL	Employment status - current	NULL	Full-time work
Observation	1	42528763	2017-03-30	32817	[null]	37079292	[null]	NULL	Highest level of education	NULL	High school diploma or GED
Measurement	1	3036277	2017-03-30	32817	176.5	[null]	8582	cm	Body Height	cm	176.5
Observation	1	37020774	2017-03-30	32817	[null]	37079033	[null]	NULL	In the past year have you or any family ...	NULL	Clothing
Measurement	1	3027018	2017-03-30	32817	68.0	[null]	8541	/min	Heart rate	/min	68.0
Observation	1	42868746	2017-03-30	32817	0.0	[null]	44777566	{score}	Generalized anxiety disorder 7 item (GAD...	{score}	0.0
Observation	1	46234808	2017-03-30	32817	0.0	[null]	44777566	{score}	Total score [HARK]	{score}	0.0
Observation	1	37020730	2017-03-30	32817	[null]	37079425	[null]	NULL	Has lack of transportation kept you from...	NULL	Yes it has kept me from n...
Measurement	1	3004249	2017-03-30	32817	134.0	[null]	8876	mm[Hg]	Systolic Blood Pressure	mm[Hg]	134.0
Measurement	1	3012030	2017-03-30	32817	32.4	[null]	8564	pg	MCH [Entitic mass] by Automated count	pg	32.4
Measurement	1	3023314	2017-03-30	32817	39.6	[null]	8554	%	Hematocrit [Volume Fraction] of Blood b...	%	39.6
Measurement	1	3020416	2017-03-30	32817	5.0	[null]	8815	10*6/uL	Erythrocytes [#/volume] in Blood by Auto...	10*6/uL	5.0

```sql
INSERT INTO cdm.measurement
(
    measurement_id,
    person_id,
    measurement_concept_id,
    measurement_date,
    measurement_datetime,
    measurement_time,
    measurement_type_concept_id,
    operator_concept_id,
    value_as_number,
    value_as_concept_id,
    unit_concept_id,
    range_low,
    range_high,
    provider_id,
    visit_occurrence_id,
    visit_detail_id,
    measurement_source_value,
    measurement_source_concept_id,
    unit_source_value,
    unit_source_concept_id,
    value_source_value
)
SELECT
    stem_table.id AS measurement_id,
    stem_table.person_id,
    coalesce(stem_table.concept_id, 0) AS measurement_concept_id,
    stem_table.start_date AS measurement_date,
    stem_table.start_datetime AS measurement_datetime,
    NULL AS measurement_time,
    stem_table.type_concept_id AS measurement_type_concept_id,
    stem_table.operator_concept_id,
    stem_table.value_as_number,
    stem_table.value_as_concept_id,
    stem_table.unit_concept_id,
    stem_table.range_low,
    stem_table.range_high,
    stem_table.provider_id,
    stem_table.visit_occurrence_id,
    stem_table.visit_detail_id,
    stem_table.source_value AS measurement_source_value,
    stem_table.source_concept_id AS measurement_source_concept_id,
    stem_table.unit_source_value,
    stem_table.unit_source_concept_id,
    stem_table.value_source_value
FROM cdm.stem_table
WHERE stem_table.domain_id = 'Measurement';
```

```sql
INSERT INTO cdm.observation
(
    observation_id,
    person_id,
    observation_concept_id,
    observation_date,
    observation_datetime,
    observation_type_concept_id,
    value_as_number,
    value_as_string,
    value_source_value,
    value_as_concept_id,
    qualifier_concept_id,
    unit_concept_id,
    provider_id,
    visit_occurrence_id,
    visit_detail_id,
    observation_source_value,
    observation_source_concept_id,
    unit_source_value,
    qualifier_source_value,
    observation_event_id,
    obs_event_field_concept_id
)
SELECT
    stem_table.id   AS  observation_id,
    stem_table.person_id,
    coalesce(stem_table.concept_id, 0) AS observation_concept_id,
    stem_table.start_date AS observation_date,
    stem_table.start_datetime AS observation_datetime,
    stem_table.type_concept_id AS observation_type_concept_id,
    stem_table.value_as_number,
    stem_table.value_as_string,
    stem_table.value_source_value,
    stem_table.value_as_concept_id,
    stem_table.qualifier_concept_id,
    stem_table.unit_concept_id,
    stem_table.provider_id,
    stem_table.visit_occurrence_id,
    stem_table.visit_detail_id,
    stem_table.source_value AS observation_source_value,
    stem_table.source_concept_id AS observation_source_concept_id,
    stem_table.unit_source_value,
    stem_table.qualifier_source_value,
    stem_table.event_id AS observation_event_id,
    stem_table.event_field_concept_id AS obs_event_field_concept_id
FROM cdm.stem_table
WHERE stem_table.domain_id = 'Observation';
```

OMOP - Measurement

measurement_id [PK] integer	person_id integer	measurement_concept_id integer	measurement_type_concept_id integer	value_as_number numeric	unit_concept_id integer	unit_source_value character varying (50)	measurement_source_value character varying (500)
561	1	3036277	32817	176.5	8582	cm	Body Height
183	1	3024171	32817	15.0	8541	/min	Respiratory rate
498	1	3012030	32817	32.4	8564	pg	MCH [Entitic mass] by Automated count
288	1	3043111	32817	9.6	8583	fL	Platelet mean volume [Entitic volume] in Blood by Automated count
309	1	3038553	32817	28.6	9531	kg/m2	Body mass index (BMI) [Ratio]
253	1	3025315	32817	89.1	9529	kg	Body Weight
50	1	3027018	32817	68.0	8541	/min	Heart rate
428	1	3000963	32817	17.1	8713	g/dL	Hemoglobin [Mass/volume] in Blood
29	1	3024929	32817	212.5	44777519	10*3/uL	Platelets [#/volume] in Blood by Automated count
1	1	3002736	32817	281.8	8583	fL	Platelet distribution width [Entitic volume] in Blood by Automated co..
400	1	3000905	32817	4.4	44777519	10*3/uL	Leukocytes [#/volume] in Blood by Automated count
603	1	3012888	32817	81.0	8876	mm[Hg]	Diastolic Blood Pressure
92	1	43055141	32817	3.0	44777566	{score}	Pain severity - 0-10 verbal numeric rating [Score] - Reported
526	1	3020416	32817	5.0	8815	10*6/uL	Erythrocytes [#/volume] in Blood by Automated count
512	1	3023599	32817	84.4	8583	fL	MCV [Entitic volume] by Automated count
169	1	3004249	32817	134.0	8876	mm[Hg]	Systolic Blood Pressure
225	1	3015182	32817	40.5	8583	fL	Erythrocyte distribution width [Entitic volume] by Automated count
337	1	3023314	32817	39.6	8554	%	Hematocrit [Volume Fraction] of Blood by Automated count
134	1	3009744	32817	34.3	8713	g/dL	MCHC [Mass/volume] by Automated count

OMOP - Observation

observation_id [PK] integer	person_id integer	observation_concept_id integer	observation_source_value character varying (250)	observation_type_concept_id integer	value_as_number numeric	value_as_concept_id integer	value_source_value character varying (250)
302	1	3046853	Race	32817	[null]	[null]	White
22	1	42869557	Housing status	32817	[null]	37079501	I have housing
694	1	37020116	Stress level	32817	[null]	45883172	Not at all
78	1	40766239	How many people are living or staying at this address?	32817	8.0	[null]	8.0
687	1	37020846	At any point in the past 2 years has season or migrant farm...	32817	[null]	45878245	No
127	1	46235654	Primary insurance	32817	[null]	45877444	Private insurance
680	1	37021580	Have you been discharged from the armed forces of the Uni...	32817	[null]	45878245	No
477	1	46235507	Within the last year have you been afraid of your partner or ...	32817	[null]	45878245	No
351	1	40758879	Patient Health Questionnaire 2 item (PHQ-2) total score [Re...	32817	0.0	[null]	0.0
673	1	37020172	Are you worried about losing your housing?	32817	[null]	45878245	No
358	1	40759172	Do you consider yourself Hispanic/Latino?	32817	[null]	45878245	No
666	1	37020774	In the past year have you or any family members you live wi...	32817	[null]	37079033	Clothing
365	1	40759918	Address	32817	17041	[null]	170 Blick Lock Suite 41
659	1	37020730	Has lack of transportation kept you from medical appointm...	32817	[null]	37079425	Yes it has kept me from non-...
393	1	40766311	What was your best estimate of the total income of all famil...	32817	33375	[null]	33375
414	1	40770471	Employment status - current	32817	[null]	37079092	Full-time work
652	1	37020032	How often do you see or talk to people that you care about ...	32817	5	37079490	5 or more times a week
421	1	42868746	Generalized anxiety disorder 7 item (GAD-7) total score [Re...	32817	0.0	[null]	0.0
645	1	37021367	In the past year have you spent more than 2 nights in a row...	32817	[null]	45878245	No
442	1	43054909	Tobacco smoking status	32817	[null]	45883458	Ex-smoker (finding)
638	1	37020108	Do you feel physically and emotionally safe where you curre...	32817	[null]	45877994	Yes

CDM source

```
INSERT INTO cdm.cdm_source
(
    cdm_source_name,
    cdm_source_abbreviation,
    cdm_holder,
    source_description,
    source_documentation_reference,
    cdm_etl_reference,
    source_release_date,
    cdm_release_date,
    cdm_version,
    cdm_version_concept_id,
    vocabulary_version
)
VALUES
(
    'Synthea ETL',
    'Synthea',
    'ETL Workshop peepz',
    'Synthea data using 50k generated patients',
    'link to source doc',
    'link to repo',
    '2025-12-31',
    CURRENT_DATE,
    'v5.4',
    756265,
    'v5.0 27-FEB-26'
);
```

https://ohdsi.github.io/CommonDataModel/sqlScripts.html#v54

Introduction	Maxim Moinat
ETL Implementation demo (part 1)	Liam Glueck
ETL Implementation demo (part 2)	Anne van Winzum
Q&A	
ETL Execution Considerations	Anne van Winzum
Loading new OMOP vocabulary	Liam Glueck
Release comparison	Maxim Moinat


OHDSI
Observational Health Data Sciences and Informatics

Questions?

ETL Execution

Anne van Winzum and Stefan Payrable

Orchestration

- Now we have a quick one-off ETL, but now we need to get a maintainable, reproducible and performant ETL process.
- Ok, great, but how?

What to consider for the ETL pipeline

- Language - Python/SQL/R
- What does the ETL pipeline need to have besides running of the transformations?
  - Load the source data
  - Create the OMOP CDM tables and load the vocabularies
  - Logging of issues/progress
  - Performance!

Performance (1/4) Constraint handling

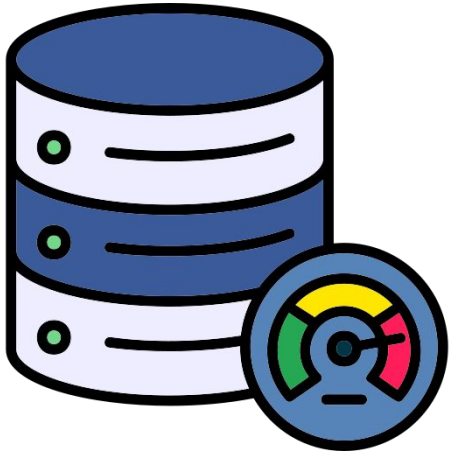
Inserting data into a heavily indexed database (like OMOP) negatively affects performance

➔ Insert data before applying indexes and constraints

➔ Apply some OMOP PKs/indexes early for faster lookups; e.g. vocabulary tables, person_source_value

If your source data is also in a database, some well-placed indexes can make a big difference.

➔ Good candidate columns are those used in WHERE clauses and JOIN operations



Performance (2/4) DB configuration

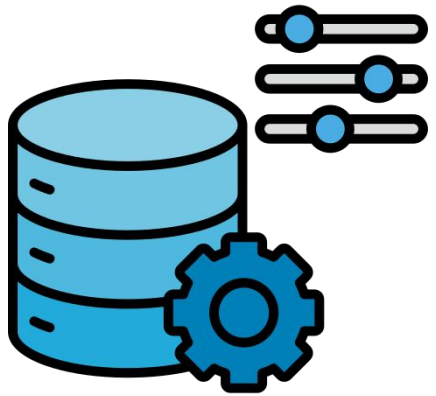
Default database settings are not always the best for the typical OMOP use cases.

A default Postgres DB is optimized for handling many (small) transactions concurrently, not bulk insert or complex queries.

Changing these settings can significantly improve performance

PGTune

Check PGTune with type 'Data warehouse' for a tuning suggestion based on your hardware



https://pgtune.leopard.in.ua/

Performance (3/4) Batching

Processing data in smaller batches can help, especially when aggregating data in row-oriented databases like Postgres

Prevent having to do aggregation/sorting on larger-than-memory data (avoids disk spilling)

E.g. the DRUG_ERA query can benefit from smaller groups of persons being processed in parallel.

Performance (4/4) Engine/Platform choice

If your traditional database has hit a performance limit, you can consider a specialized engine for large data processing

Highly-performant open-source solutions are available

Incremental data loading

An incremental data loading approach can be useful if both of the following are true:

- Your source dataset is frequently updated
- Converting the full dataset to OMOP is costly and/or takes a long time

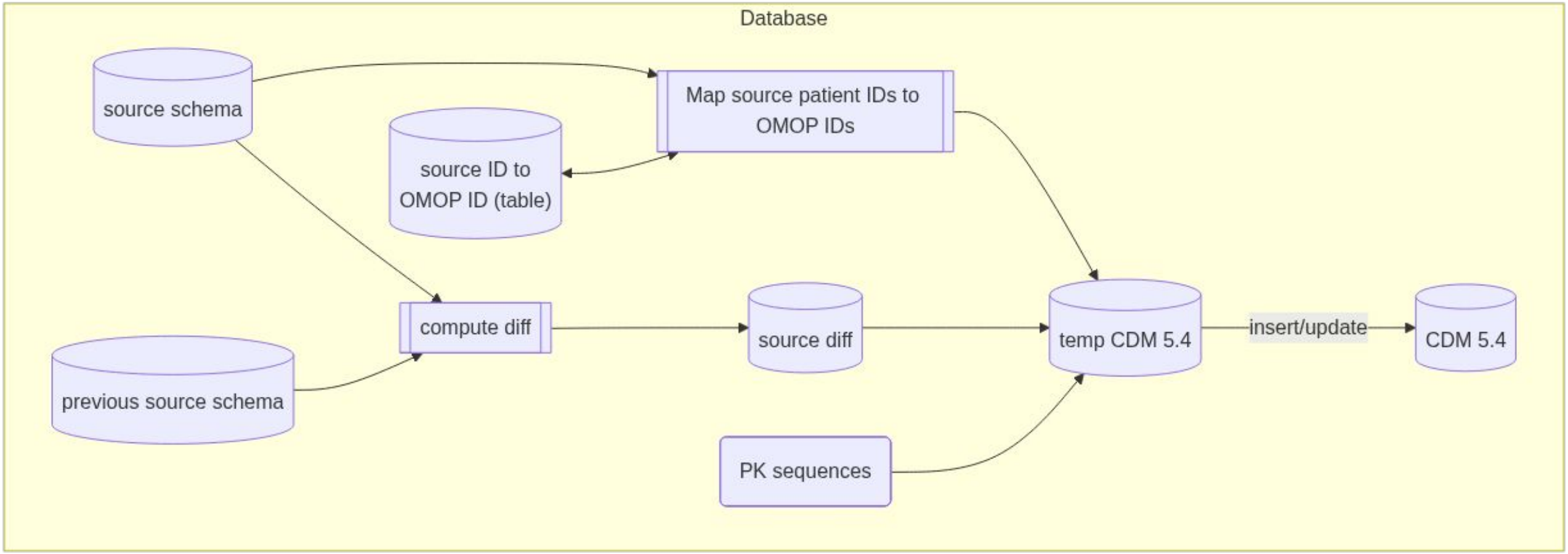
Incremental loading can improve the efficiency of your pipeline, but it comes at the price of increased complexity

Incremental data loading

Multiple approaches are possible, depending on what type of changes can be expected in the source data

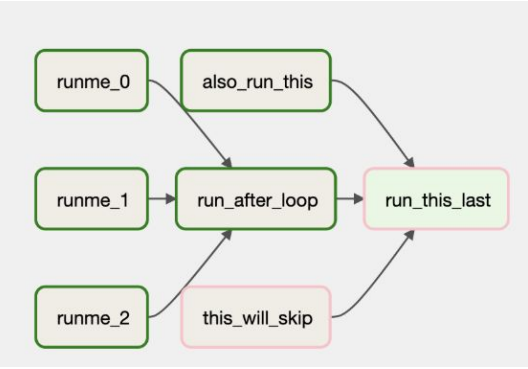
- Only new records?
- Updates/deletes in old records?
- Patient consent withdrawal?

Incremental data loading



ETL Pipeline Tooling Examples

- Python - delphyne (The Hyve)
- Python - Rabbit-in-a-Blender (AZ Delta)
- R – ETL-Synthea (OHDSI Github)
- Java - JCdmBuilder (OHDSI Github)
- C# - ETL-CDMBuilder (OHDSI Github)
- Frameworks: dbt, Pentaho, bonobo
- Airflow



Vocabulary Update

Liam Glueck

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Vocabulary Update

How often do you need to update your OMOP vocabularies?

- Every 6 months, a new vocabulary is released.
  - Every February & August

What changes take place between each vocabulary version?

- This varies depending on the release:
  - Feb 2026: RxNorm, ATC & CPT4 updates and more.
  - Aug 2025: Race/Ethnicicty, LOINC, SNOMED & ICD updates and more.

Vocabulary Update

What can you do to update (already made) custom mapping files?

What tools can help you with this?

- Update your vocabulary in Usagi, manual approve new concepts
- Use Kotobuki to update your mappings automatically
  - Github: https://github.com/thehyve/kotobuki

Kotobuki

- Search algorith uses concept relationships to find non-standard to standard mappings.
- Can handle one-to-many mappings and homonym search.
- Provides new mappings as well as mapping paths
- Can be used with CLI or via python.

CDM Release Comparison

Maxim Moinat

Harmonisation is not one-off

- Update with **more recent data**
- Add **more sources**
- Improve **data quality**
  - Expand vocabulary mappings
  - Adapt to (new) conventions, e.g. observation period
- New OMOP CDM version
- Review each new release, and compare

Extract Load Transform (ETL) journey

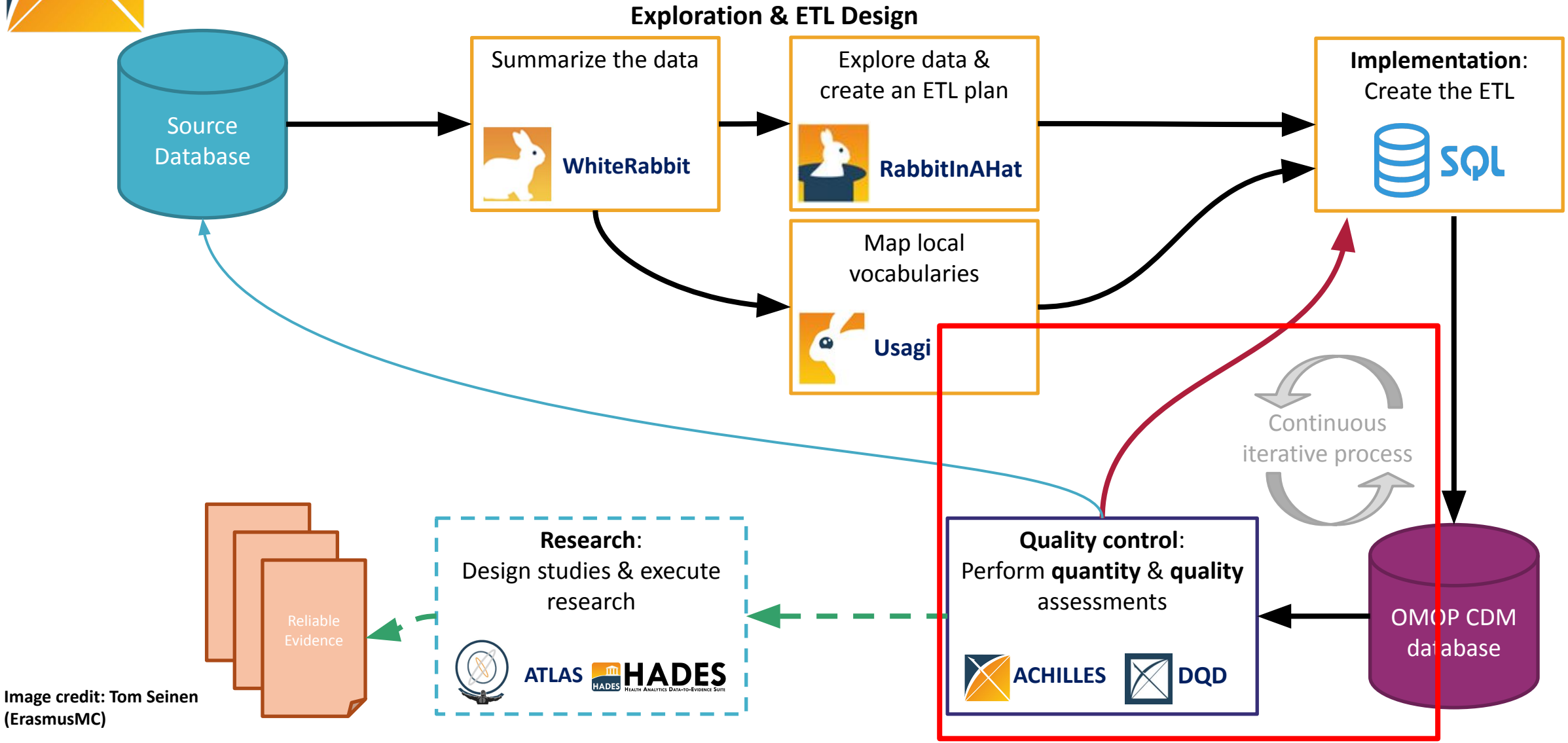


Image credit: Tom Seinen (ErasmusMC)

15:30 – 17:30	**Reviewing results in OHDSI shiny apps** Melissa Leung, Berta Raventós (EMC) <u>Description</u> • Hands-on session to exploring OHDSI Shiny Applications, including apps for phenotype assessment and those with study results • Learn how to navigate and review Shiny Applications • Learn how to identify potential inconsistencies (especially valuable for researchers leading their own studies or	**Data Quality Assessment Framework & Tools** Clair Blacketer, Anthony Sena (J&J) <u>Description</u> • Data Quality Dashboard (DQD) and other recent developments for tools to assess data quality • Hands-on exercise for running DQD to identify and address ETL conversion issues • Data quality considerations for network studies <u>Target audience</u> Anyone responsible for assessing and improving the quality of OMOP CDM ETL or data quality for study	**Hands-on characterization session using OHDSI/DARWIN packages** Adam Black (EMC), Marta Alcalde-Herraiz, Moronfoluwa Akintola, Martí Català (UO) <u>Description</u> • Hands-on session focused on cohort characterization using OHDSI/DARWIN tools • Connecting to OMOP CDM data in a tidy, user-friendly way • Programmatically creating cohorts with inclusion and exclusion criteria • Running phenotype

Comparing results

- New DQD failures?
- Decrease in records, where increase expected?
- Drastic change in target concept count

Ares - Combine and compare results from Achilles and DQD

SYNTHEA SYNTHETIC HEALTH DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

DataQualityDashboard Version: 2.6.1
Results generated at 2024-10-06 20:49:37 in 2 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	409	15	424	96%	4	0	4	100%	413	15	428	96%
Conformance	902	1	903	100%	137	0	137	100%	1039	1	1040	100%
Completeness	442	10	452	98%	17	0	17	100%	459	10	469	98%
Total	1753	26	1779	99%	158	0	158	100%	1911	26	1937	**99%**

495 out of 1911 passed checks are Not Applicable, due to empty tables or fields.
4 out of 26 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 98% (1416/1438).

ARES

A Research **E**xploration **S**ystem that facilitates exploration of patient level, observational data research accompanied by source data characterization and quality assessment ensuring that results are presented with proper context.

EXPLORE DATA SOURCES

Report Category: Data Source

Data Source: DE

Report: Data Source Overview

Source Overview

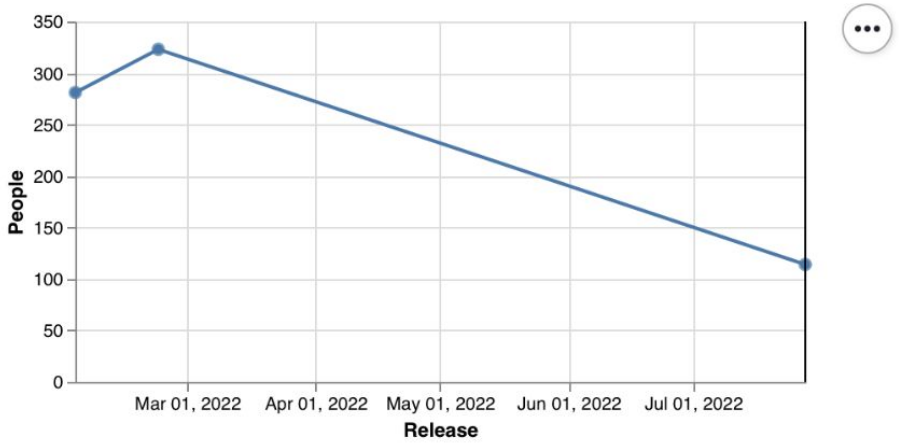
Releases

3

Average Days Between Releases

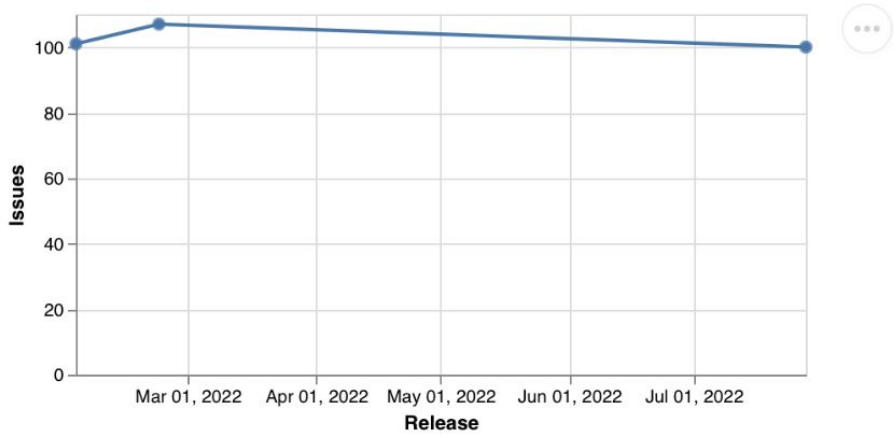
88

Population History



View export query

Data Quality Issues History



View export query

Thanks for your Attention!