OMOP
Common Data Model (CDM)
& Extract-Transform-Load (ETL)
Tutorial

Rimma Belenkaya (Memorial Sloan Kettering)
Karthik Natarajan (Columbia University)
Mark Velez (Columbia University)
Erica Voss (Janssen R&D Epidemiology Analytics)

24 September 2016

Please copy the contents of the
USB drive to your hard disk now.
You will need ~45GB free disk space available.

Rimma Belenkaya Memorial Sloan Kettering	Karthik Natarajan Columbia University
Mark Velez Columbia University	Erica Voss Janssen R&D Epidemiology Analytics

Teaching Assistants

Anthony Sena Janssen R&D Epidemiology Analytics	Jungmi Han Columbia University

Ground Rules

- We are recording today's session, so presenters should repeat questions.
- We may table source specific questions.
- The Virtual Machine (VM) distributed today on USB, please return.
- If we cannot get the VM working on your machine let's try to buddy you up. Do not worry the presentation will still walk you through the content.
- This course will not focus on the Vocabulary, however the Vocabulary is critical to the Common Data Model and the ETL process.

Agenda

Time	Type	Section
8:00AM-8:15AM		Introductions
8:15AM-9:15AM	*Foundational*	What is OMOP/OHDSI? OMOP Common Data Model (CDM) – Why and How
9:15AM-10:00AM		How to retrieve data from OMOP CDM
10:00AM-10:15AM		Break
10:15AM-10:45AM	*Implementation*	Setup and Performing of an Extract Transform and Load process into the CDM
10:45AM-11:30AM		Using WhiteRabbit and Rabbit-In-A-Hat to Build an ETL
11:30AM-11:45AM	*Evaluation*	Testing and Quality Assurance
11:45AM-12:00PM		Wrap up

OHDSI

Observational Health Data Sciences and Informatics

Foundational

What is OMOP/OHDSI? OMOP Common Data Model (CDM) – Why and How

Introduction of OMOP/OHDSI

OHDSI: Observational Health Data Sciences and Informatics is a research collaborative coordinated through Columbia University

Who?

–Multiple stakeholders: academia, government, industry

–Multiple disciplines: statistics, epidemiology, informatics, clinical sciences

Why? To generate evidence about all aspects of healthcare

Where? Multiple geographies: US, Europe, Asia-Pacific, 20 countries. OHDSI collaborators access a network of 600 mln patients

How? By developing analytical methods and tools based on the data standardized to OMOP Common Data Model (CDM) and vocabulary

OMOP Common Data Model (CDM) What is it and why have one?

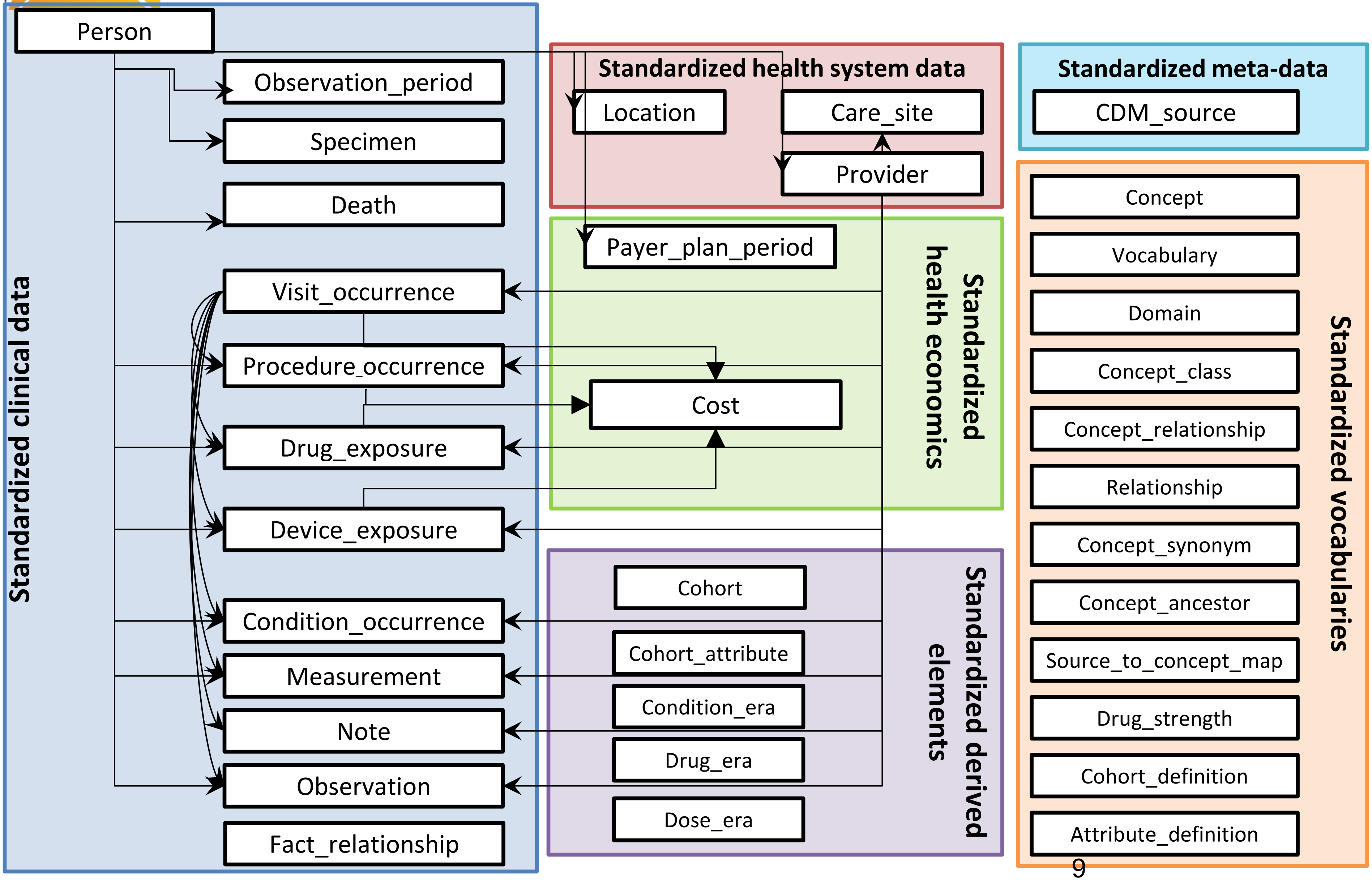
What?

- A standardized way to represent data structure (CDM) and content (vocabulary)
- One model to accommodate data coming from disparate data sources
  - –administrative claims, electronic health records
  - –EHRs from both inpatient and outpatient settings
  - –registries and longitudinal surveys
  - –data sources both within and outside of US

Why?

- Enable standardization of structure and content to support a systematic and reproducible process to efficiently generate evidence
- Support collaborative research both within and outside of US

OMOP CDM v5.0.1



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_map

Drug_strength

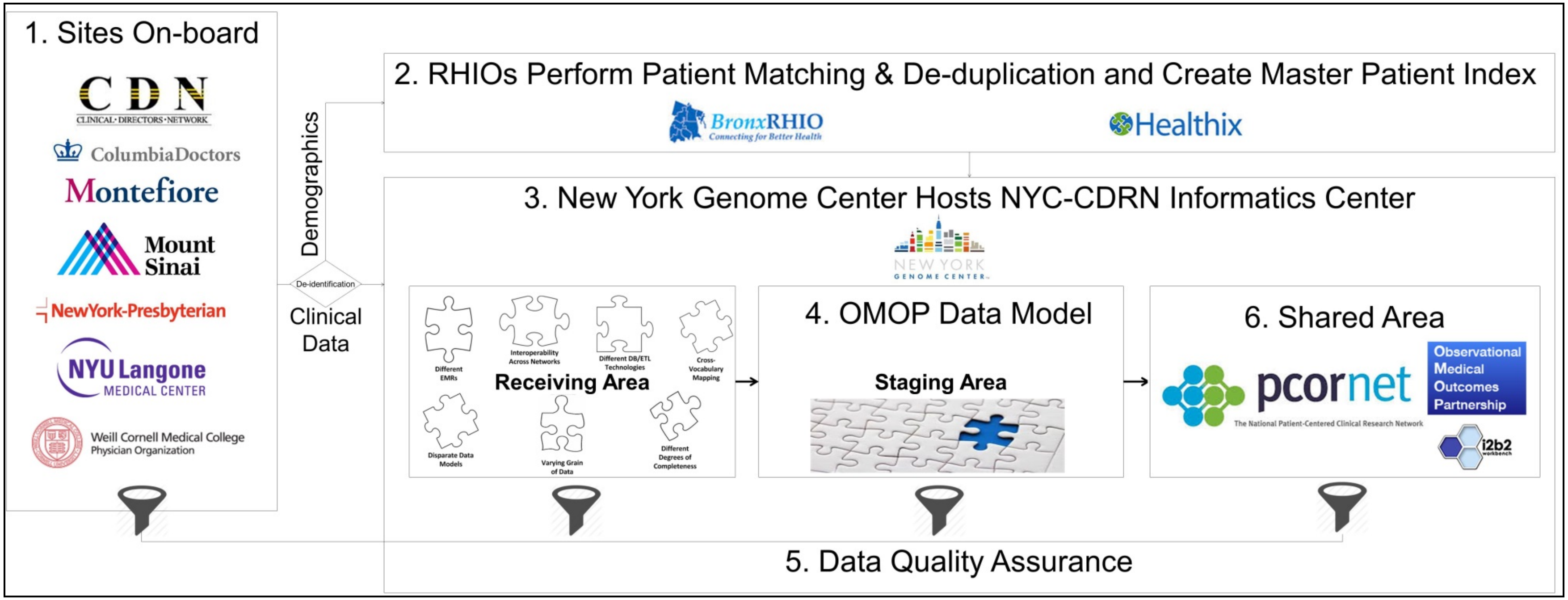
Cohort_definition

Attribute_definition

OMOP CDM Design Principles

- Relational design but platform independent
  - Integrated with Controlled Vocabulary
  - Domain (subject area) based
  - Patient centric
  - Uniformly integrates data from heterogeneous data sources: EMR, claims, registries
- Built for analytical purposes, extended/developed based on analytic use cases
- Extendable, both vocabulary (new vocabs, local concepts) and CDM (Observation)

NYC-CDRN Experience



1. Sites On-board

2. RHIOs Perform Patient Matching & De-duplication and Create Master Patient Index

3. New York Genome Center Hosts NYC-CDRN Informatics Center

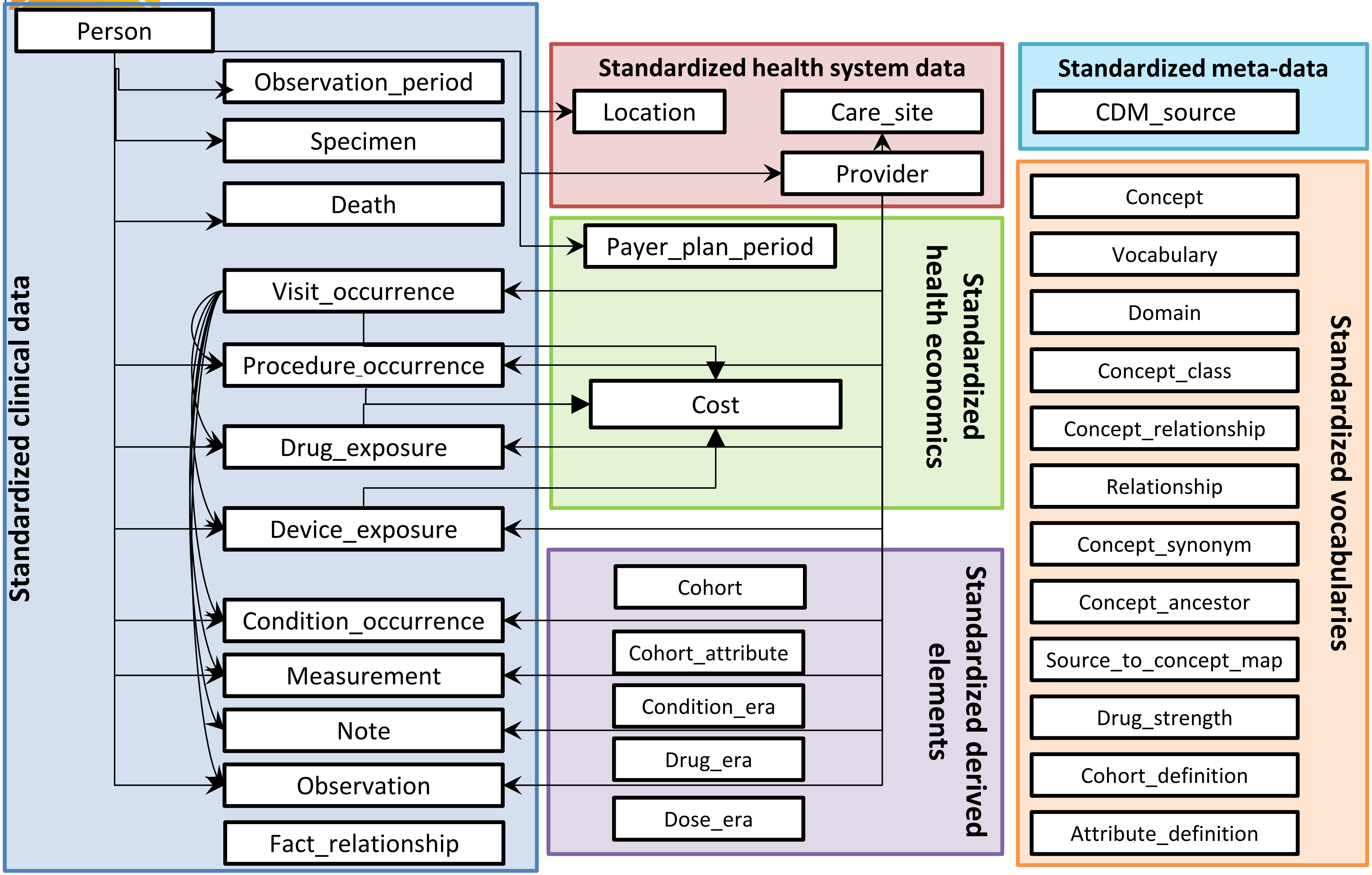
4. OMOP Data Model

5. Data Quality Assurance

6. Shared Area

CDN
CLINICAL · DIRECTORS · NETWORK
ColumbiaDoctors
Montefiore
Mount Sinai
NewYork-Presbyterian
NYU Langone MEDICAL CENTER
Weill Cornell Medical College Physician Organization
Demographics
De-identification
Clinical Data
BronxRHIO Connecting for Better Health
Healthix
NEW YORK GENOME CENTER
Different EMRs
Interoperability Across Networks
Different DB/ETL Technologies
Cross-Vocabulary Mapping
Receiving Area
Disparate Data Models
Varying Grain of Data
Different Degrees of Completeness
Staging Area
pcornet
The National Patient-Centered Clinical Research Network
Observational Medical Outcomes Partnership
i2b2 workbench

OMOP CDM v5.0.1



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_map

Drug_strength

Cohort_definition

Attribute_definition

OMOP Common Vocabulary Model

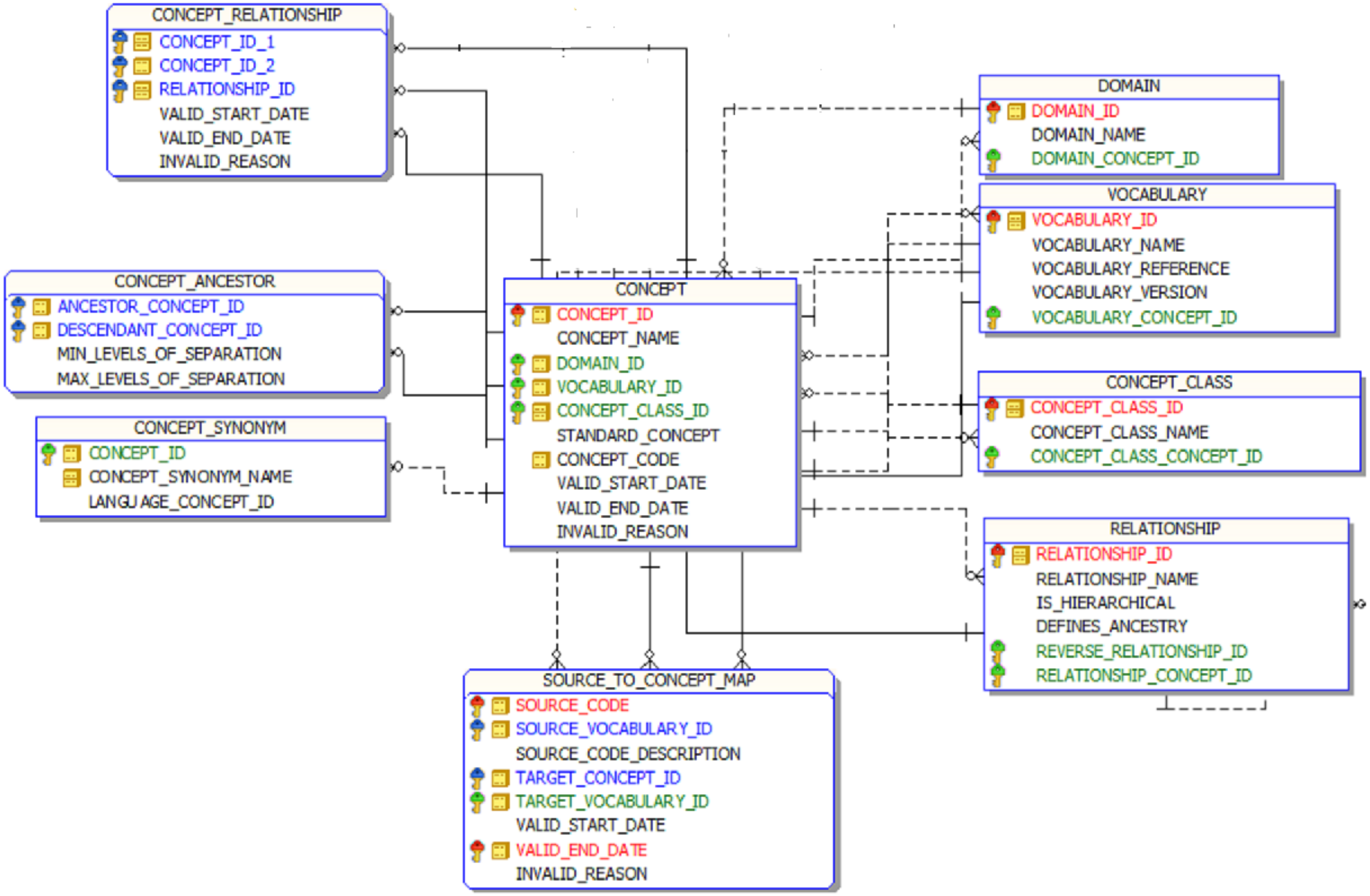
What it is

- Standardized structure to house existing vocabularies used in the public domain
- Compiled standards from disparate public and private sources and some OMOP-grown concepts
- Built on the shoulders of National Library of Medicine's Unified Medical Language System (UMLS)

What it's not

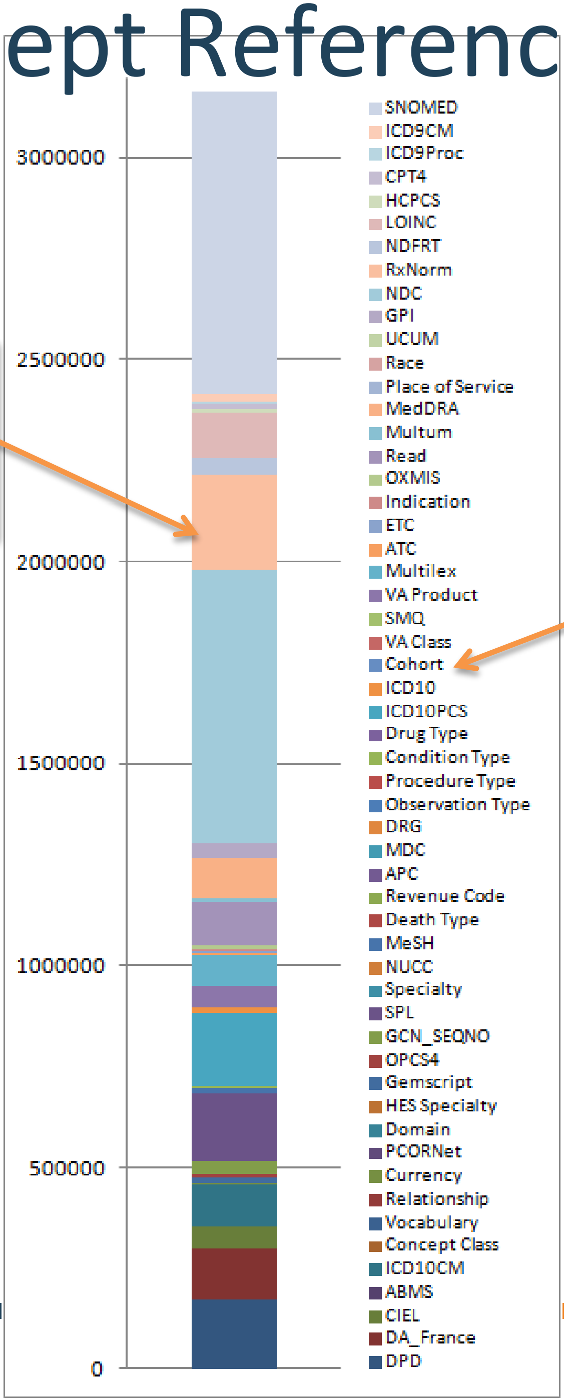
- Static dataset – the vocabulary updates regularly to keep up with the continual evolution of the sources
- Finished product – vocabulary maintenance and improvement is ongoing activity that requires community participation and support

OMOP Common Vocabulary Model



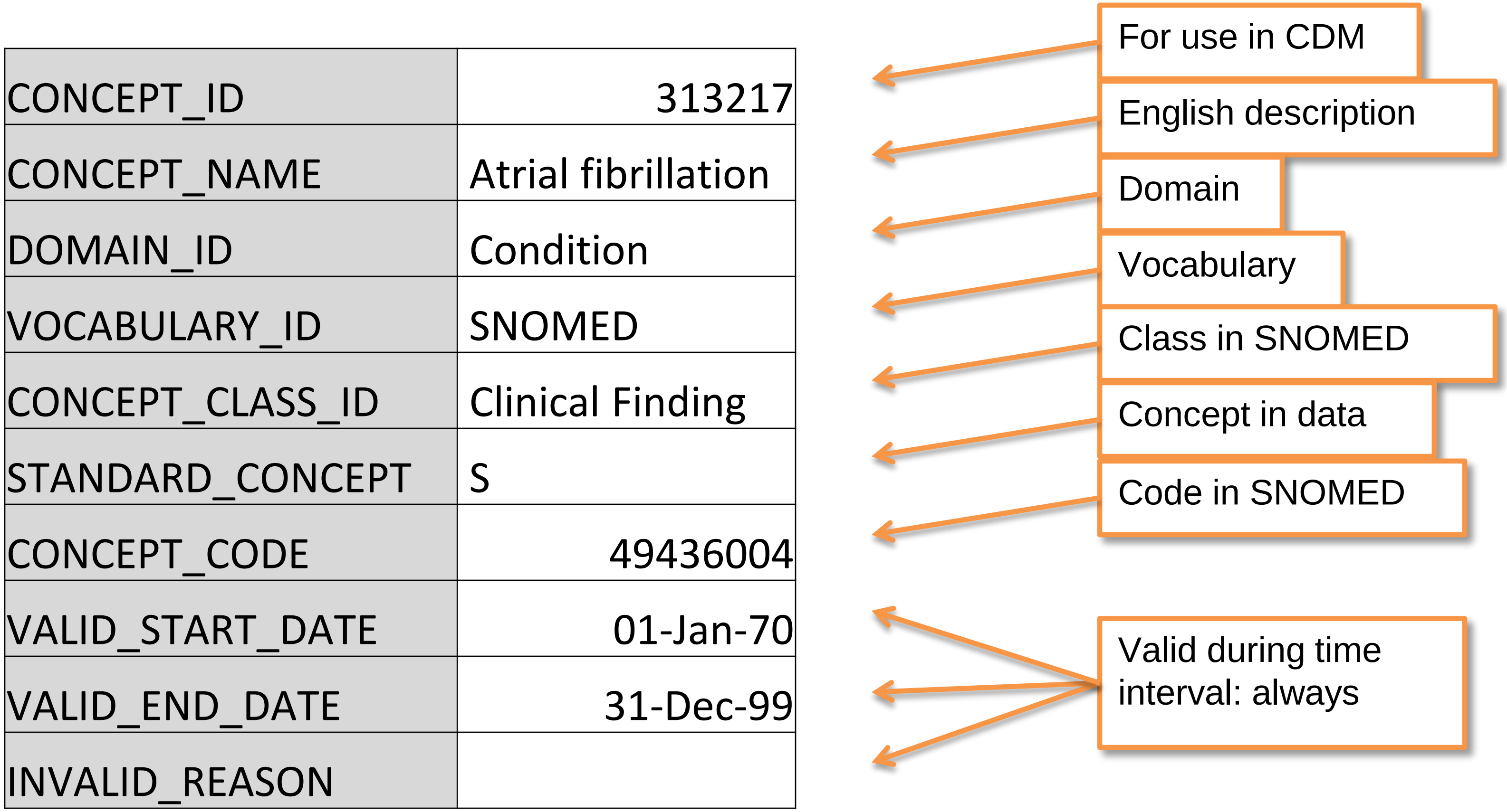
Single Concept Reference Table

All vocabularies stacked up in one table



Vocabulary ID

What's in a Concept



CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-70
VALID_END_DATE	31-Dec-99
INVALID_REASON	

For use in CDM

English description

Domain

Vocabulary

Class in SNOMED

Concept in data

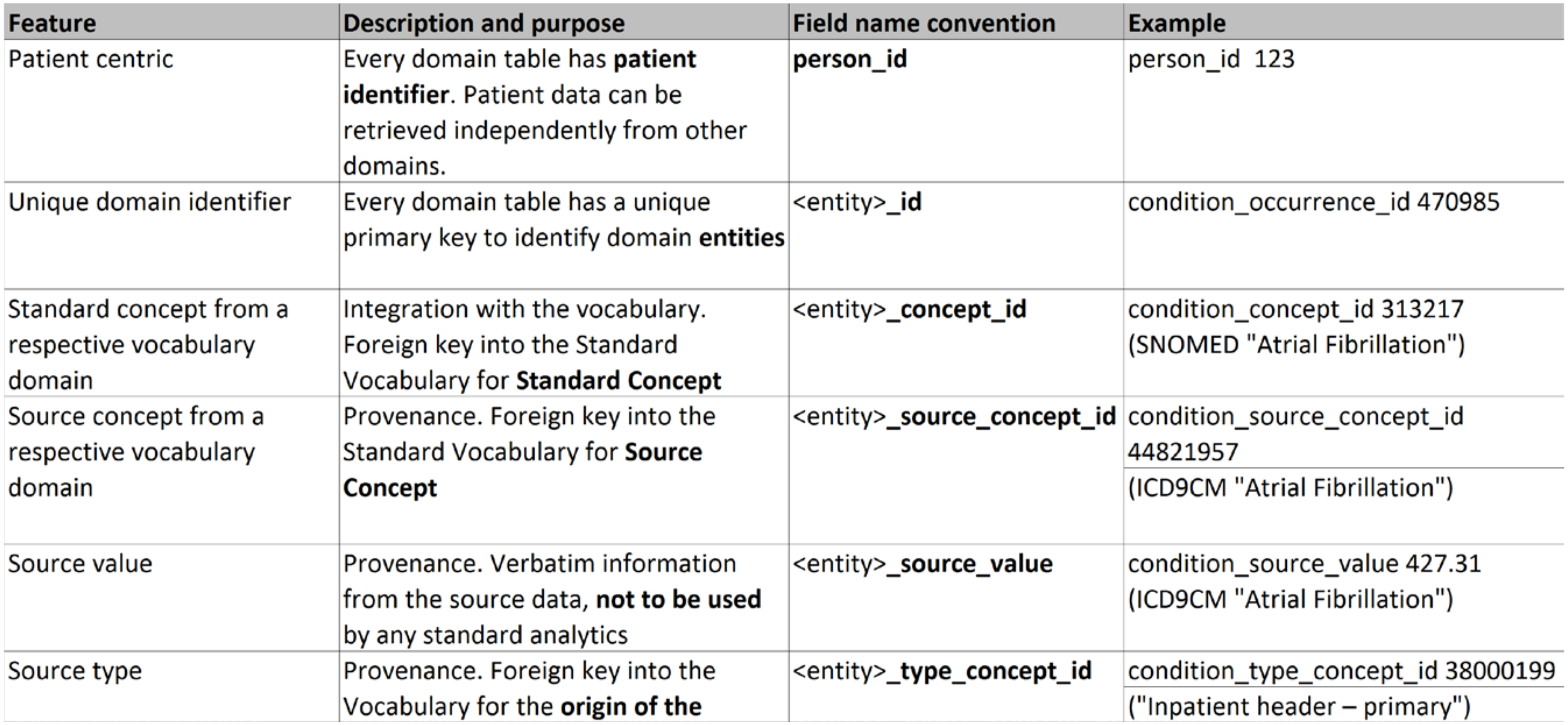
Code in SNOMED

Valid during time interval: always

OMOP Vocabulary Model Design Principles

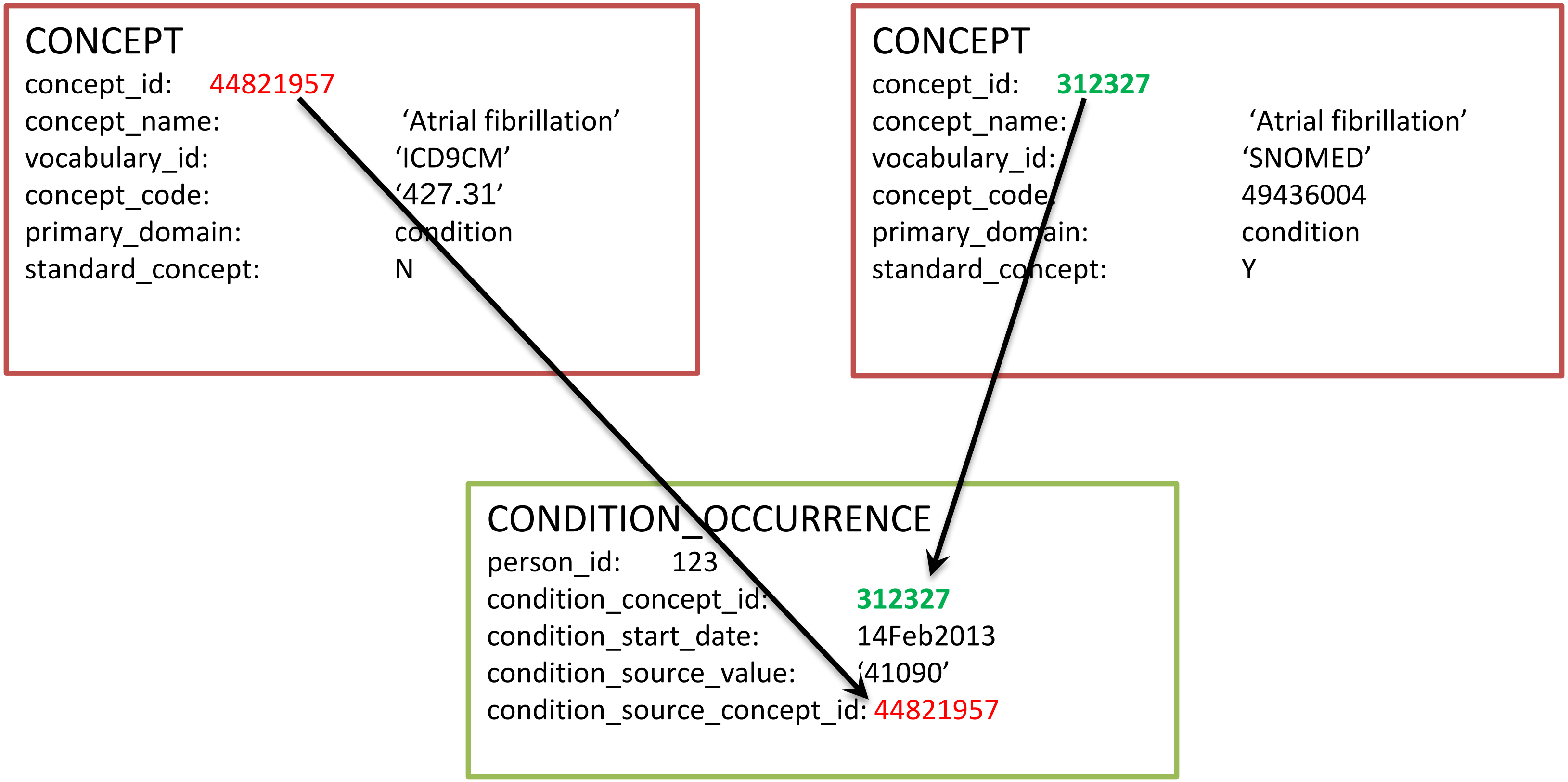
- Uniform structure
  - All concepts are in one table
  - All concept relationships are in one table, including mappings from source to standard vocabularies
- Formalized integration with Common Data Model via concept domain
  - Direction of ETL is informed by concept domain
- Relationships are bi-directional
- Hierarchical relationships have additional representation in the model to support efficient data retrieval

OMOP CDM Standard Domain Features



Feature	Description and purpose	Field name convention	Example
Patient centric	Every domain table has **patient identifier**. Patient data can be retrieved independently from other domains.	**person_id**	person_id 123
Unique domain identifier	Every domain table has a unique primary key to identify domain **entities**	<entity>**_id**	condition_occurrence_id 470985
Standard concept from a respective vocabulary domain	Integration with the vocabulary. Foreign key into the Standard Vocabulary for **Standard Concept**	<entity>**_concept_id**	condition_concept_id 313217 (SNOMED "Atrial Fibrillation")
Source concept from a respective vocabulary domain	Provenance. Foreign key into the Standard Vocabulary for **Source Concept**	<entity>**_source_concept_id**	condition_source_concept_id 44821957 (ICD9CM "Atrial Fibrillation")
Source value	Provenance. Verbatim information from the source data, **not to be used** by any standard analytics	<entity>**_source_value**	condition_source_value 427.31 (ICD9CM "Atrial Fibrillation")
Source type	Provenance. Foreign key into the Vocabulary for the **origin of the**	<entity>**_type_concept_id**	condition_type_concept_id 38000199 ("Inpatient header – primary")

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

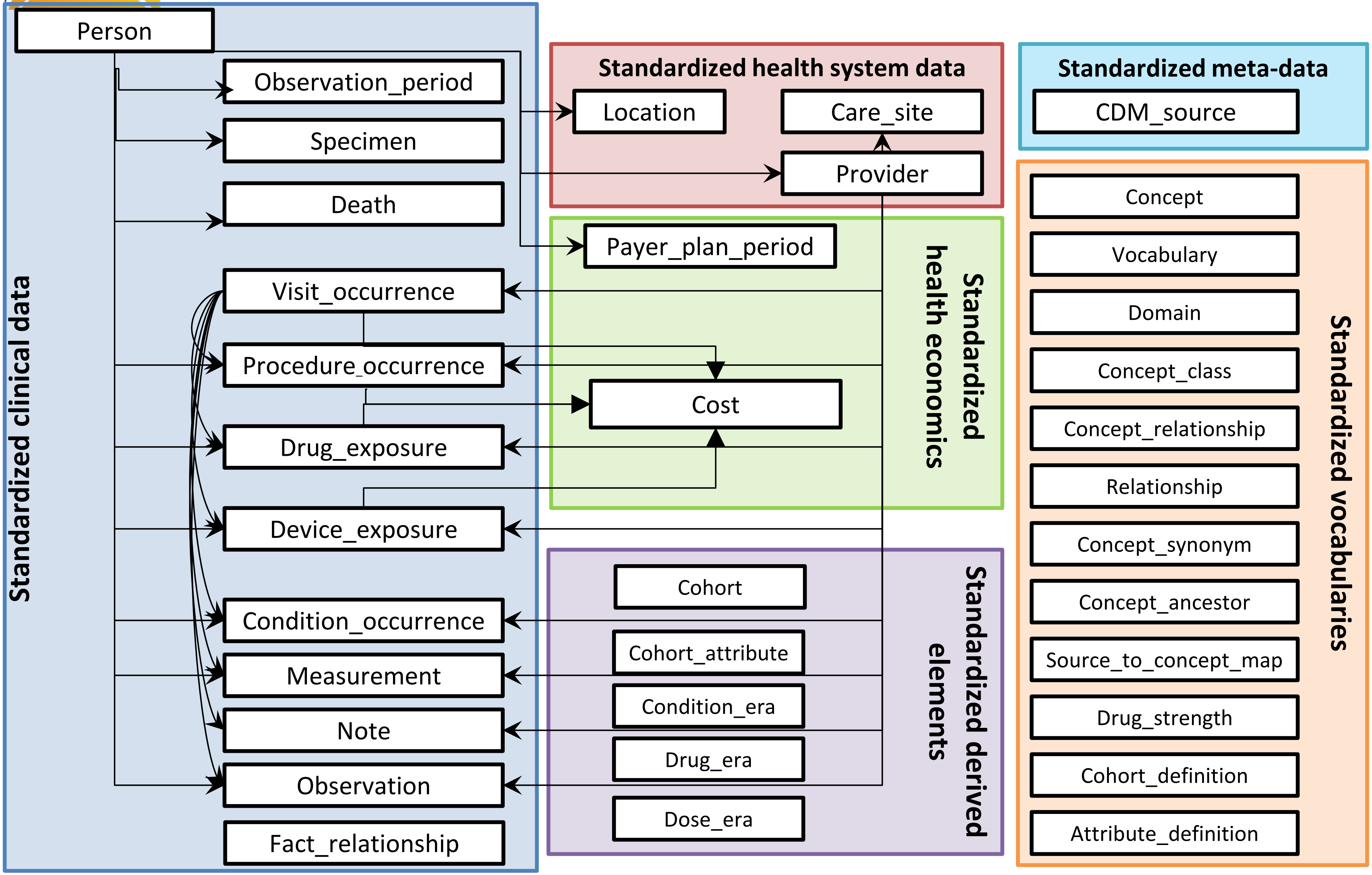
CONCEPT

concept_id:	312327
concept_name:	‘Atrial fibrillation’
vocabulary_id:	‘SNOMED’
concept_code:	49436004
primary_domain:	condition
standard_concept:	Y

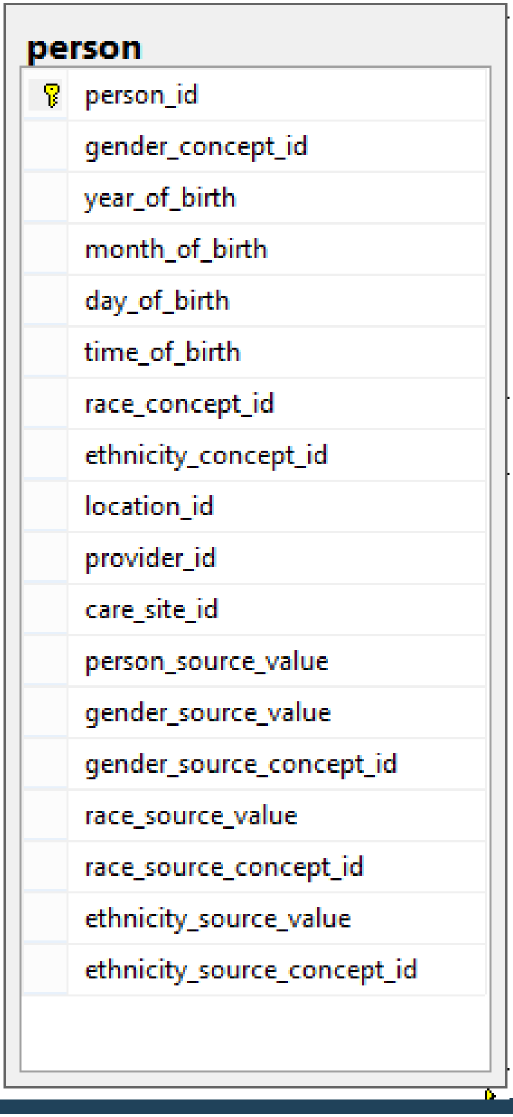
CONDITION_OCCURRENCE

person_id:	123
condition_concept_id:	312327
condition_start_date:	14Feb2013
condition_source_value:	‘41090’
condition_source_concept_id:	44821957

OMOP CDM v5.0.1



PERSON

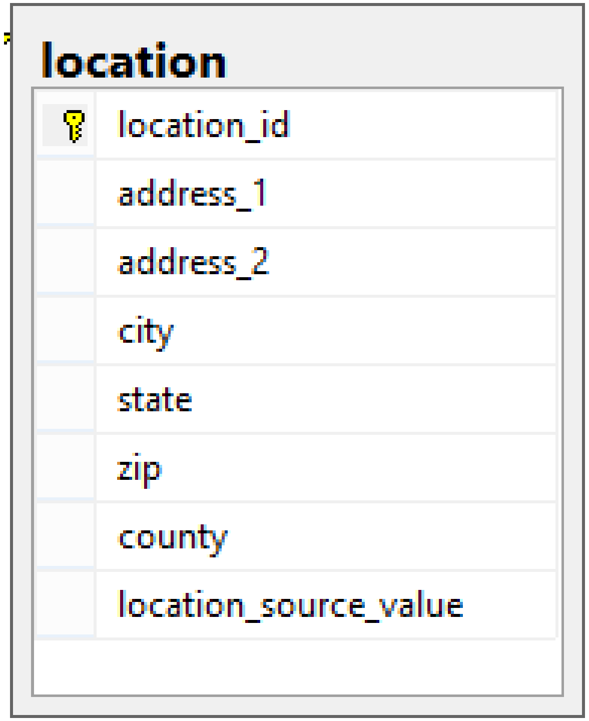


person

- person_id
- gender_concept_id
- year_of_birth
- month_of_birth
- day_of_birth
- time_of_birth
- race_concept_id
- ethnicity_concept_id
- location_id
- provider_id
- care_site_id
- person_source_value
- gender_source_value
- gender_source_concept_id
- race_source_value
- race_source_concept_id
- ethnicity_source_value
- ethnicity_source_concept_id

- Need to create one unique record per person (not multiple rows per move)
- Vocabulary for gender, race, ethnicity: HL7 administrative
- No history of location/demographics: need to select latest available
- Location peculiarity: foreign key to the LOCATION table that contains one record per each unique location
- Year of birth required…day/month optional

LOCATION



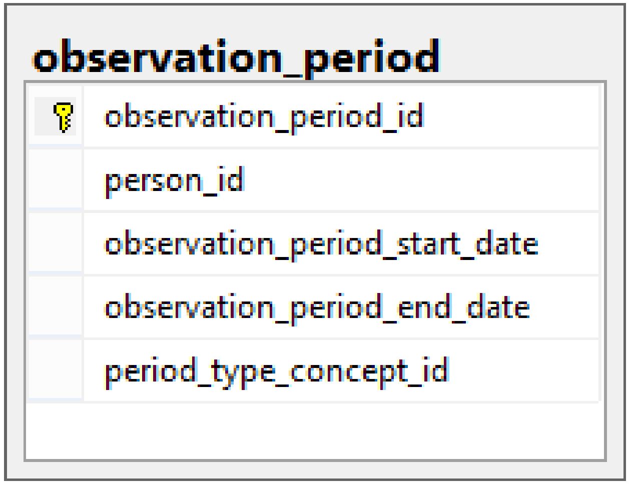
location

location_id
address_1
address_2
city
state
zip
county
location_source_value

- Contains one record per each unique location

- Location is highly variable across sources, of limited use thus far

OBSERVATION_PERIOD

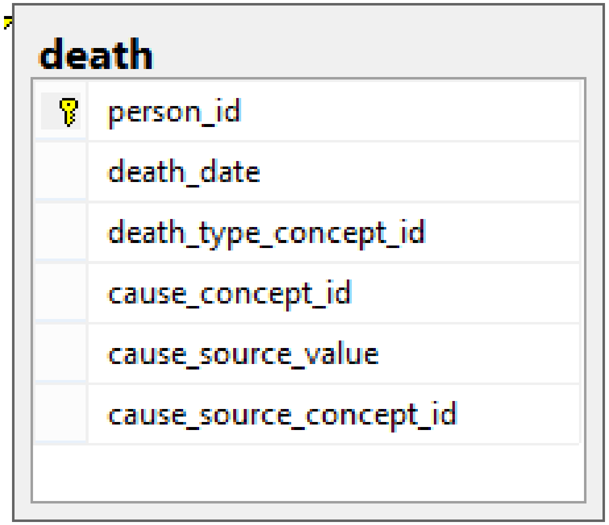


observation_period

- observation_period_id
- person_id
- observation_period_start_date
- observation_period_end_date
- period_type_concept_id

- Spans of time where data source has capture of data
- Required to run analytical methods
- One person may have multiple periods if there is interruption in data capture
- Challenge: determine observation periods based on the source data

DEATH

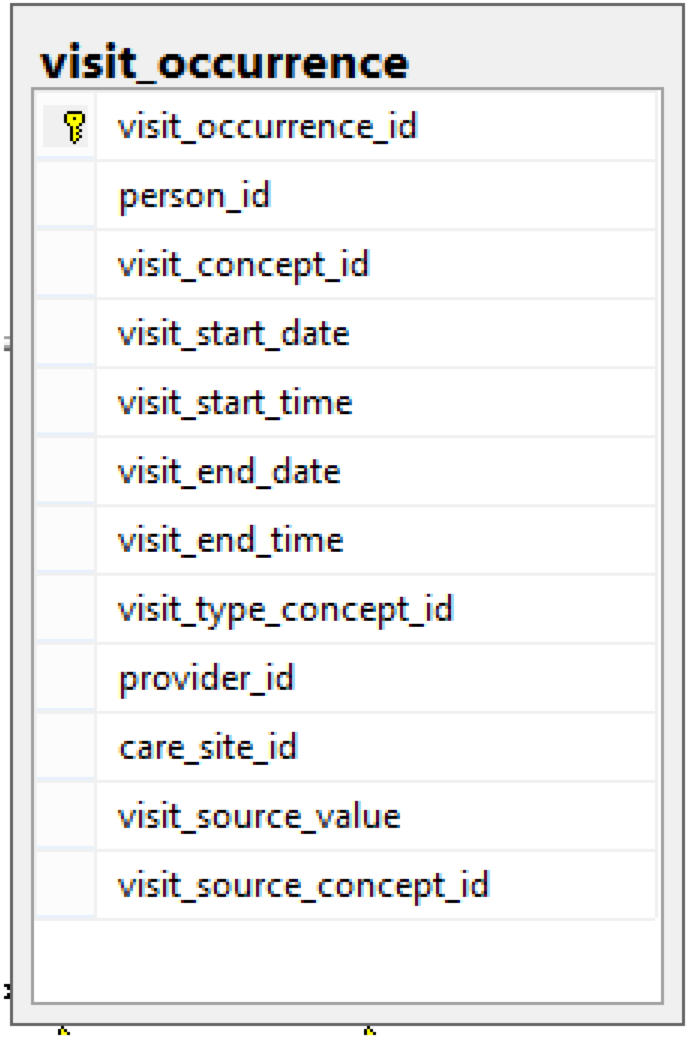


death

- 🔑 person_id
- death_date
- death_type_concept_id
- cause_concept_id
- cause_source_value
- cause_source_concept_id

- Can have death without cause
- Can only have 1 death per person

VISIT_OCCURRENCE

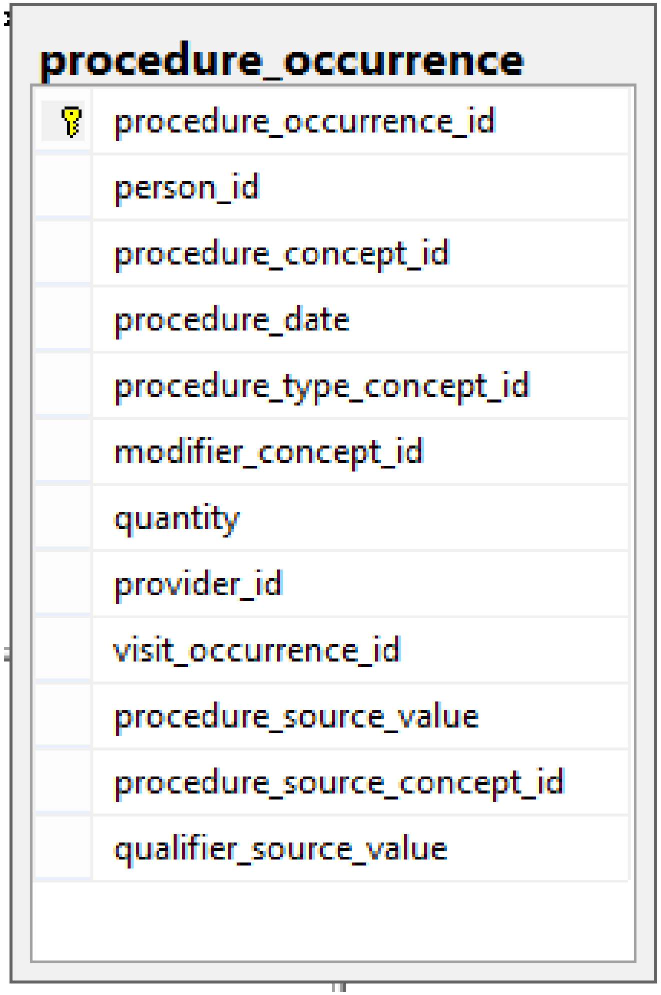


visit_occurrence

- visit_occurrence_id
- person_id
- visit_concept_id
- visit_start_date
- visit_start_time
- visit_end_date
- visit_end_time
- visit_type_concept_id
- provider_id
- care_site_id
- visit_source_value
- visit_source_concept_id

- Visits <> ‘Encounters’:
  - –claims often need to be consolidated to minimize double-counting
  - –inpatient transitions are not covered
- Visit Types
  - –Inpatient
  - –Emergency room
  - –Inpatient/Emergency - new
  - –Outpatient
  - –Long-term care
- Vocabulary: OMOP
- Other attributes: time of visit start/end, provider, admitting source, discharge disposition

PROCEDURE_OCCURRENCE



procedure_occurrence

procedure_occurrence_id

person_id

procedure_concept_id

procedure_date

procedure_type_concept_id

modifier_concept_id

quantity

provider_id

visit_occurrence_id

procedure_source_value

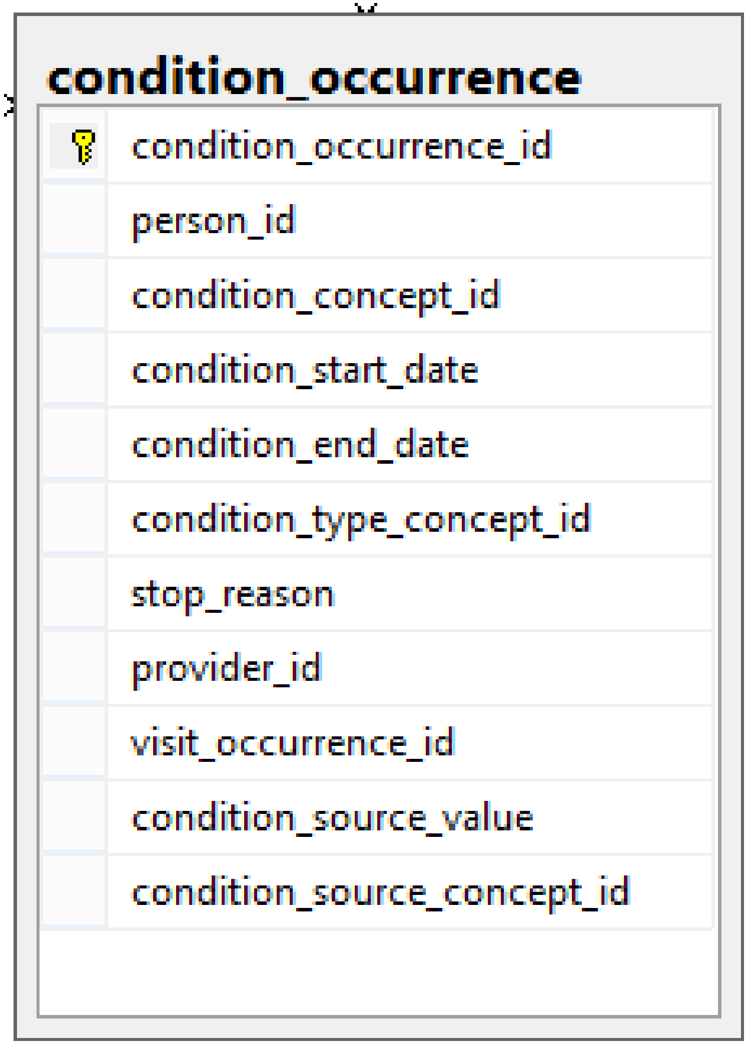
procedure_source_concept_id

qualifier_source_value

- Vocabularies: CPT-4, HCPCS, ICD-9 Procedures, ICD-10 Procedures, LOINC, SNOMED

- Procedures have the least standardized vocabularies that causes some redundancy

CONDITION_OCCURRENCE

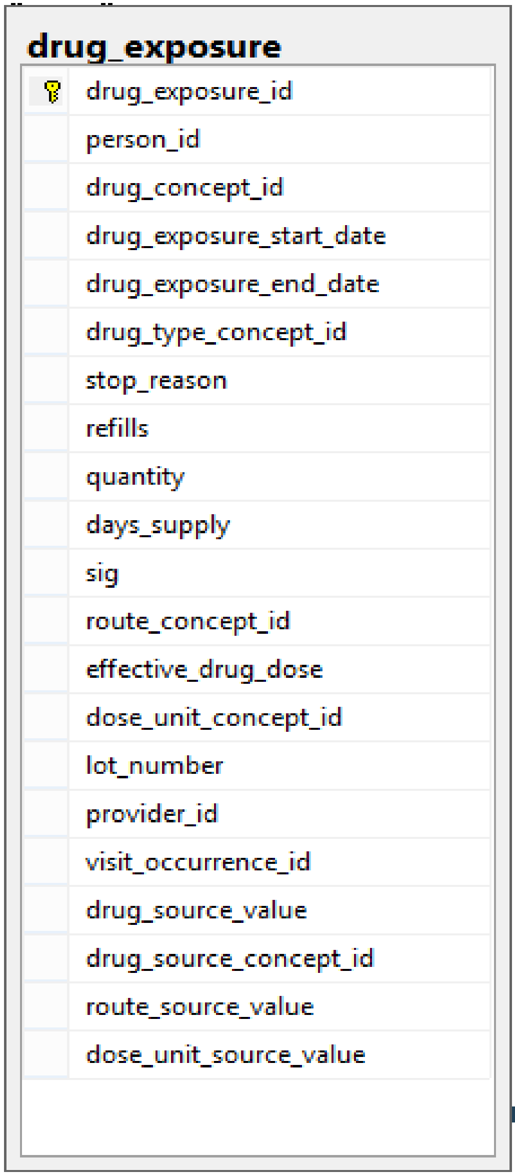


condition_occurrence

- 🔑 condition_occurrence_id
- person_id
- condition_concept_id
- condition_start_date
- condition_end_date
- condition_type_concept_id
- stop_reason
- provider_id
- visit_occurrence_id
- condition_source_value
- condition_source_concept_id

- Vocabulary: SNOMED -> classification
- Data sources:
  - Billing diagnosis (inpatient, outpatient)
  - Problem list
- Individual records <> distinct episodes

DRUG_EXPOSURE

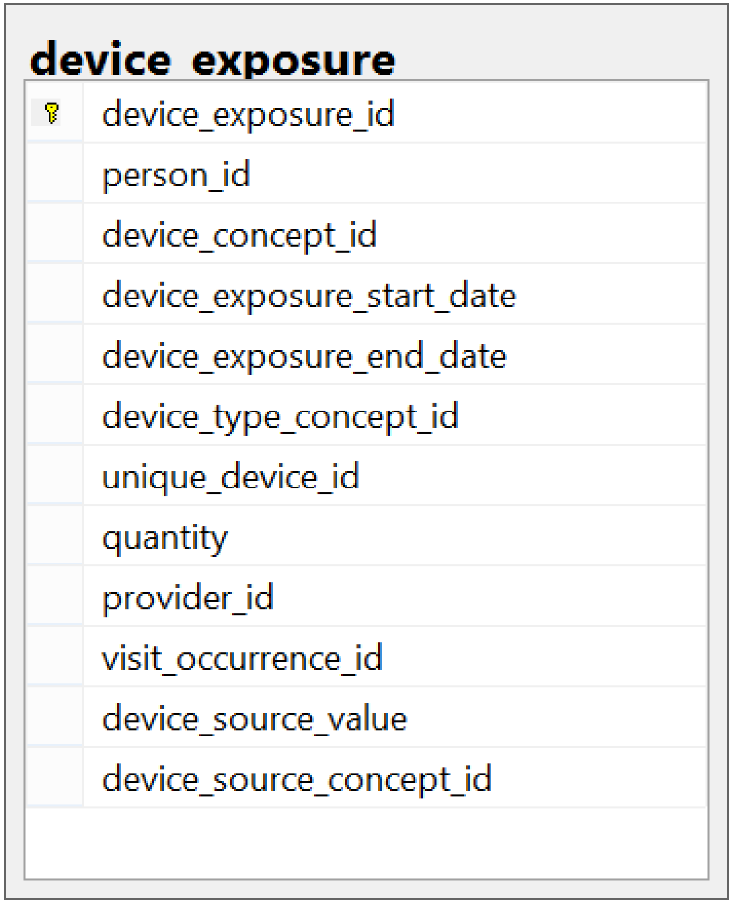


drug_exposure

- drug_exposure_id
- person_id
- drug_concept_id
- drug_exposure_start_date
- drug_exposure_end_date
- drug_type_concept_id
- stop_reason
- refills
- quantity
- days_supply
- sig
- route_concept_id
- effective_drug_dose
- dose_unit_concept_id
- lot_number
- provider_id
- visit_occurrence_id
- drug_source_value
- drug_source_concept_id
- route_source_value
- dose_unit_source_value

- Vocabulary: RxNorm-> classifications by drug class and indication
- Data sources:
  - Pharmacy dispensing
  - Prescriptions written
  - Medication history
- Source fields may vary, but so inference of drug exposure end may vary

DEVICE_EXPOSURE

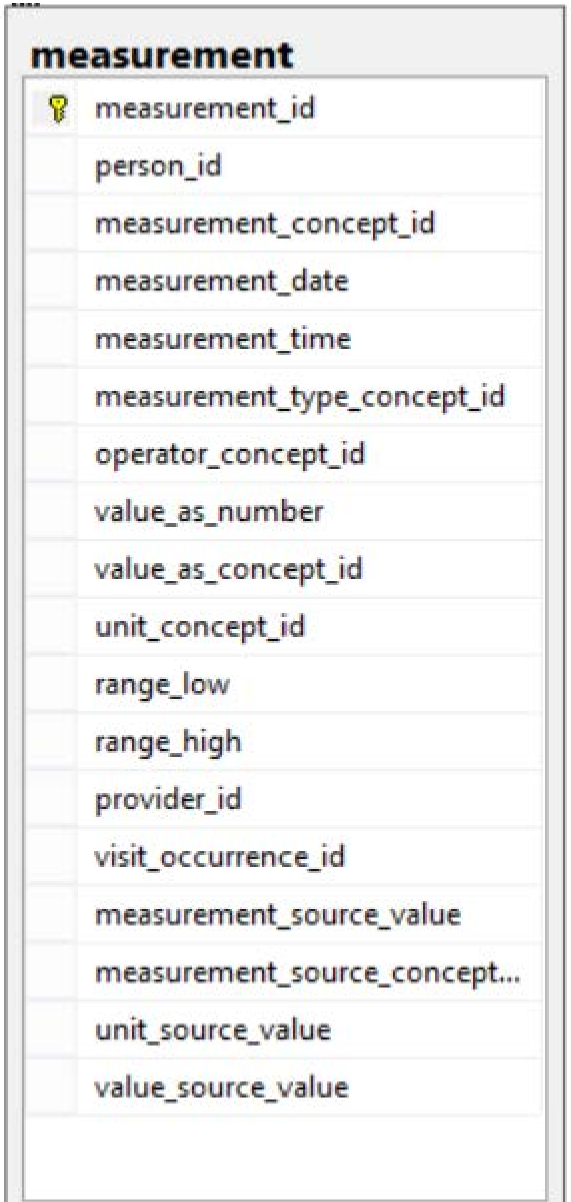


device_exposure

- 🔑 device_exposure_id
- person_id
- device_concept_id
- device_exposure_start_date
- device_exposure_end_date
- device_type_concept_id
- unique_device_id
- quantity
- provider_id
- visit_occurrence_id
- device_source_value
- device_source_concept_id

- OMOP CDM is the only data model supporting devices
- Accommodates FDA unique device identifiers (UDI) even though most data sources don’t have them yet

MEASUREMENT

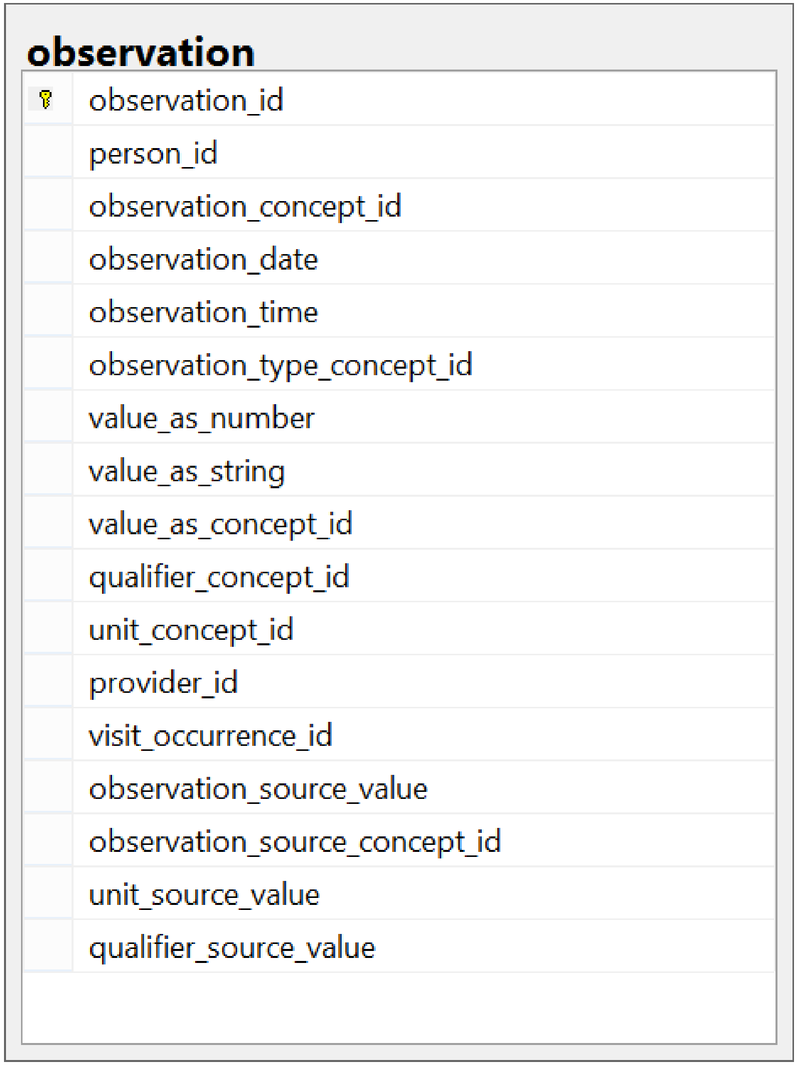


measurement

- 🔑 measurement_id
- person_id
- measurement_concept_id
- measurement_date
- 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
- measurement_source_value
- measurement_source_concept...
- unit_source_value
- value_source_value

- EAV design
- Vocabulary: LOINC, SNOMED
- Data sources: structured, quantitative measures, such as laboratory tests
- Measures have associated units
  - Measurement units vocabulary: UCUM
- No free format for measurement results

OBSERVATION



observation

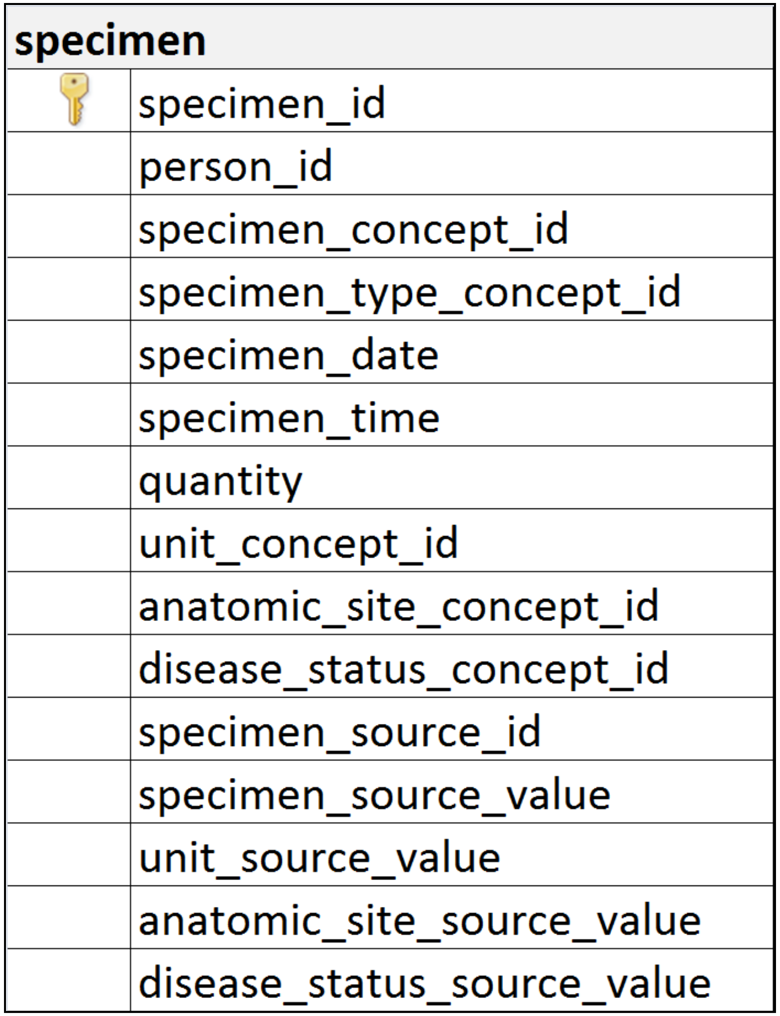
🔑 observation_id
person_id
observation_concept_id
observation_date
observation_time
observation_type_concept_id
value_as_number
value_as_string
value_as_concept_id
qualifier_concept_id
unit_concept_id
provider_id
visit_occurrence_id
observation_source_value
observation_source_concept_id
unit_source_value
qualifier_source_value

- Catch-all EAV design to capture all other data:
  - observation: ‘question’
  - value: ‘answer’
    - Can be numeric, concept, or string (e.g. free text)

- Instrument for CDM extension, playpen

- Not all ‘questions’ are standardized, source value can accommodate ‘custom’ observations (particularly pertinent in registries)

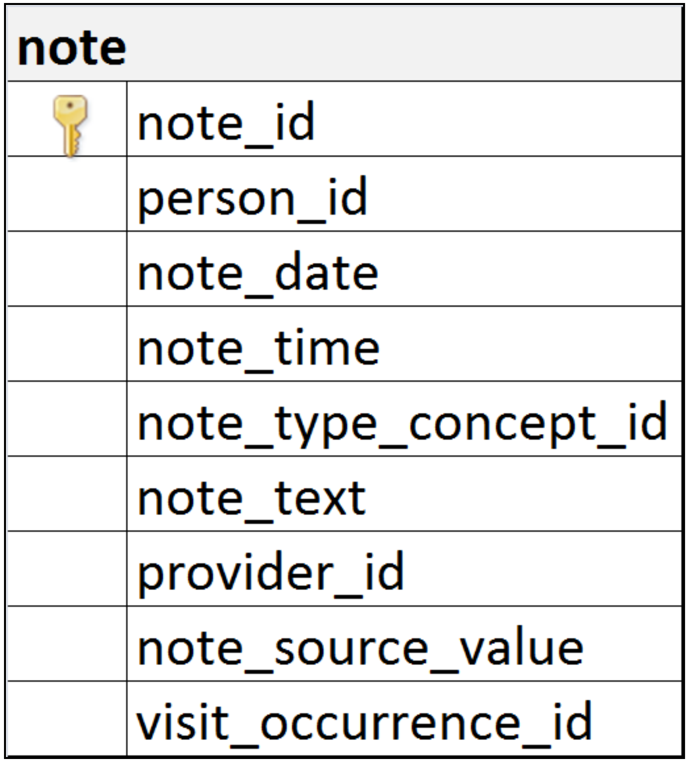
SPECIMEN



	specimen
🔑	specimen_id
	person_id
	specimen_concept_id
	specimen_type_concept_id
	specimen_date
	specimen_time
	quantity
	unit_concept_id
	anatomic_site_concept_id
	disease_status_concept_id
	specimen_source_id
	specimen_source_value
	unit_source_value
	anatomic_site_source_value
	disease_status_source_value

- To capture of biomarker / tissue bank

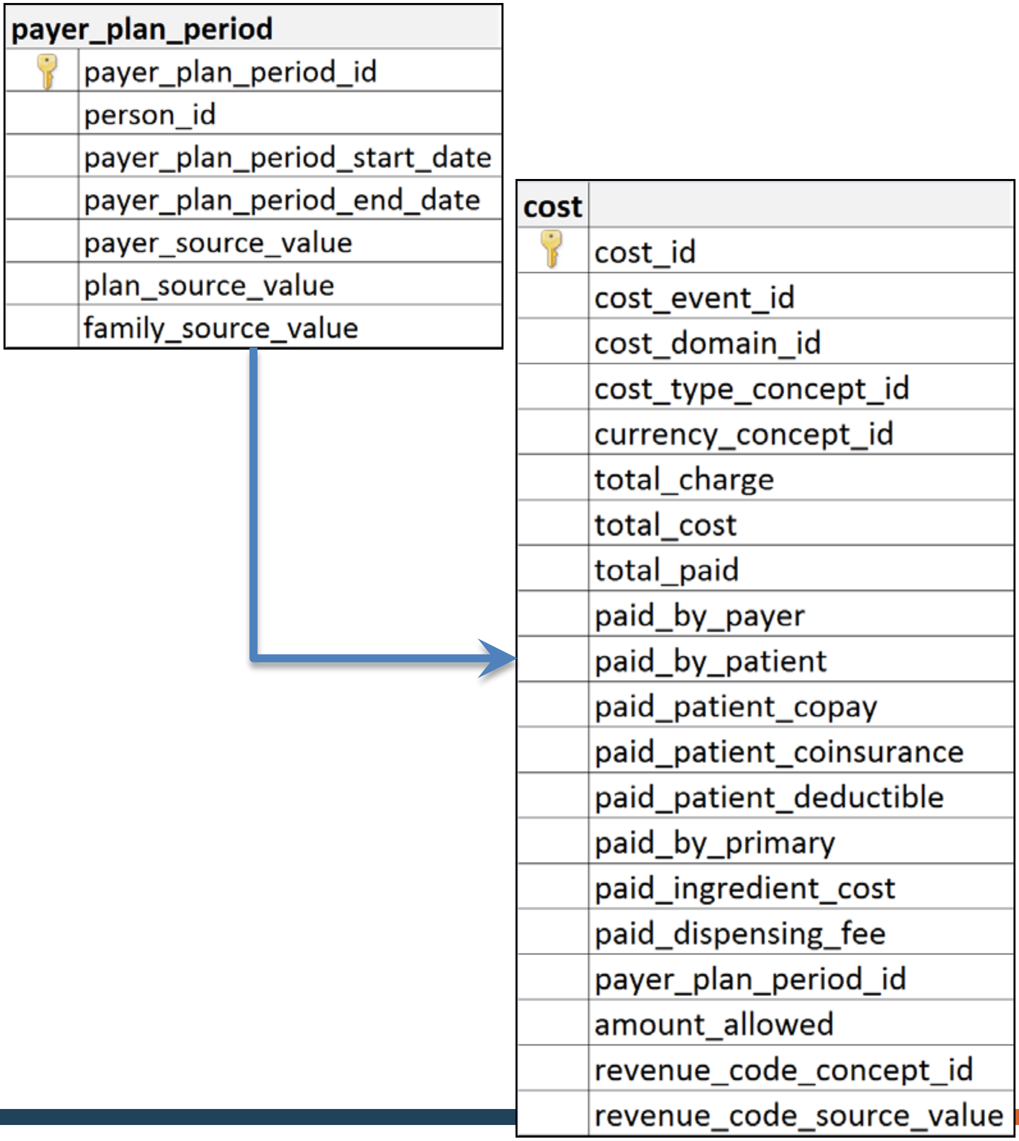
NOTE



note	
🔑	note_id
	person_id
	note_date
	note_time
	note_type_concept_id
	note_text
	provider_id
	note_source_value
	visit_occurrence_id

- To capture unstructured free text
- Coming soon in CDM 5.x: NLP and LOINC Clinical Document Ontology (CDO) annotations

Health Economics



payer_plan_period	
🔑	payer_plan_period_id
	person_id
	payer_plan_period_start_date
	payer_plan_period_end_date
	payer_source_value
	plan_source_value
	family_source_value

cost	
🔑	cost_id
	cost_event_id
	cost_domain_id
	cost_type_concept_id
	currency_concept_id
	total_charge
	total_cost
	total_paid
	paid_by_payer
	paid_by_patient
	paid_patient_copay
	paid_patient_coinsurance
	paid_patient_deductible
	paid_by_primary
	paid_ingredient_cost
	paid_dispensing_fee
	payer_plan_period_id
	amount_allowed
	revenue_code_concept_id
	revenue_code_source_value

- All costs consolidated into one table COST table
- Costs tied to respective observation records
- Domain is determined by cost_domain_id (e.g. visit, condition, etc.)

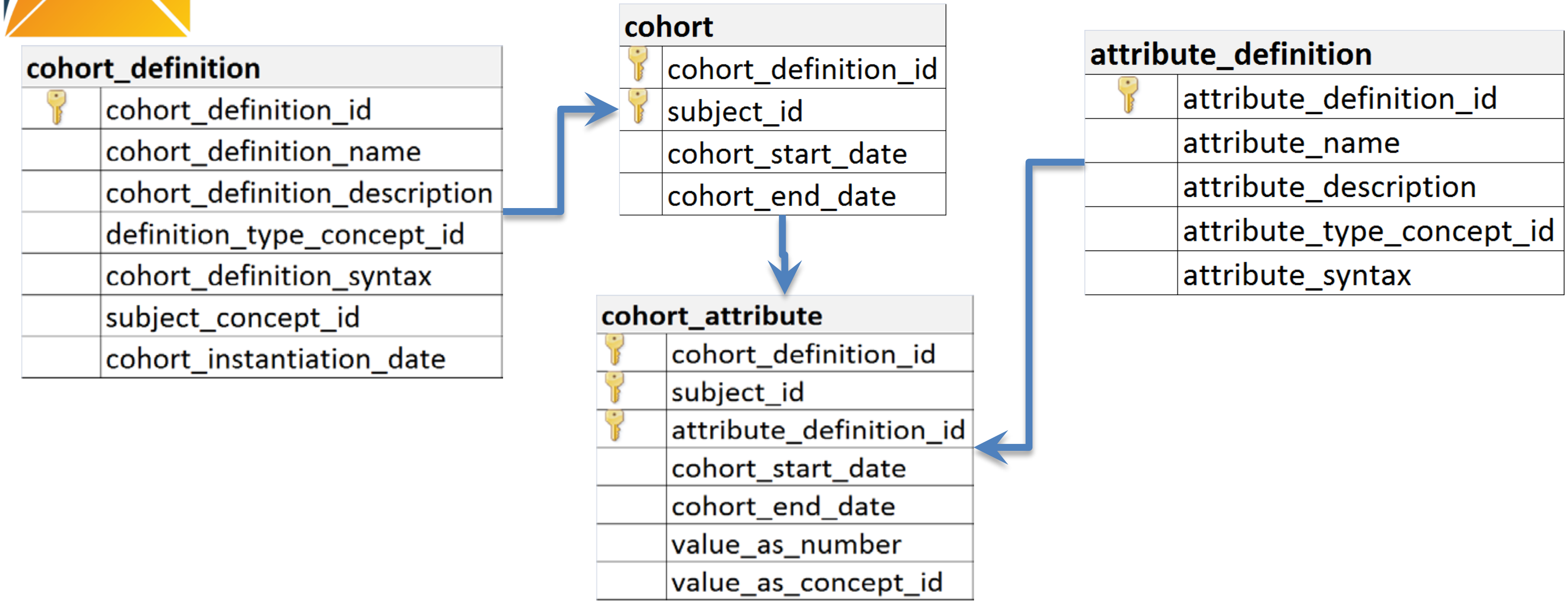
OMOP CDM Service Tables

- CDM_SOURCE
  - Provenance, integration, metadata
  - Future extension to individual domains
- FACT_RELATIONSHIP
  - Linkage between related observations
  - Example: systolic and diastolic blood pressure

Motivation for Standardized Derived Elements

- Derived elements intended to supplement- not replace- raw data
  - If derived assumptions don’t meet a specific use case, don’t use them
- Promotes transparency and consistency in research by having standard processes applies across analyses
- Increased efficiency by processing key data elements once at ETL-time, rather than requiring each analysis to figure it out at each analysis run-time
- Key standardized elements available in OMOP CDMv5:
  - Cohort – standardize definition and syntax for defining populations that meet inclusion criteria
  - Drug era – standardize inference of length of exposure to product for all active ingredients
  - Dose era – standardize estimation of daily dose for periods of exposure to all drug products
  - Condition era – standardize aggregation of episodes of care, delineating between acute vs. chronic conditions

Cohort Management



	cohort_definition
🔑	cohort_definition_id
	cohort_definition_name
	cohort_definition_description
	definition_type_concept_id
	cohort_definition_syntax
	subject_concept_id
	cohort_instantiation_date

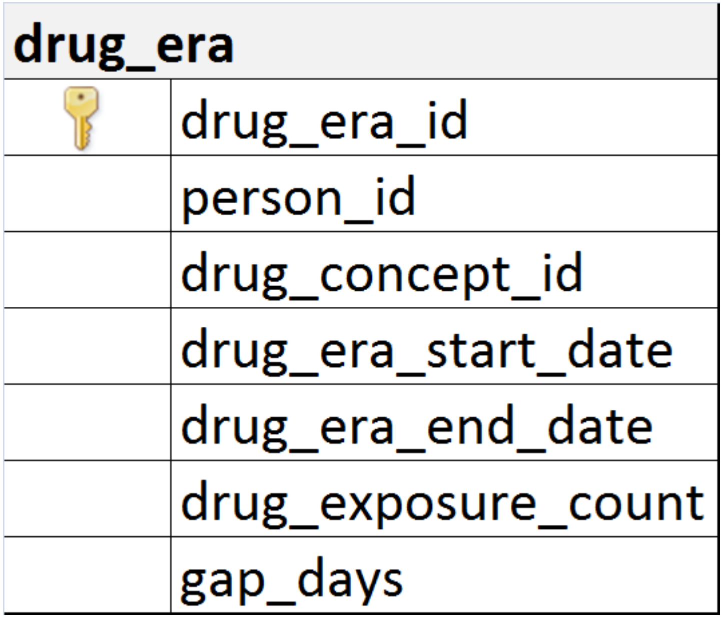
	cohort
🔑	cohort_definition_id
🔑	subject_id
	cohort_start_date
	cohort_end_date

	attribute_definition
🔑	attribute_definition_id
	attribute_name
	attribute_description
	attribute_type_concept_id
	attribute_syntax

	cohort_attribute
🔑	cohort_definition_id
🔑	subject_id
🔑	attribute_definition_id
	cohort_start_date
	cohort_end_date
	value_as_number
	value_as_concept_id

1. **COHORT** table contains records of subjects that satisfy a given set of criteria for a duration of time.
2. The definition of the cohort is contained within the **COHORT_DEFINITIO**N table. It provides a standardized structure for maintaining the rules governing the inclusion of a subject into a cohort, and can store programming code to instantiate the cohort within the OMOP CDM.
3. **COHORT_ATTRIBUTE** table contains attributes associated with each subject within a cohort, as defined by a given set of criteria for a duration of time.
4. The definition of the Cohort Attribute is contained in the **ATTRIBUTE_DEFINITION** table.

DRUG_ERA

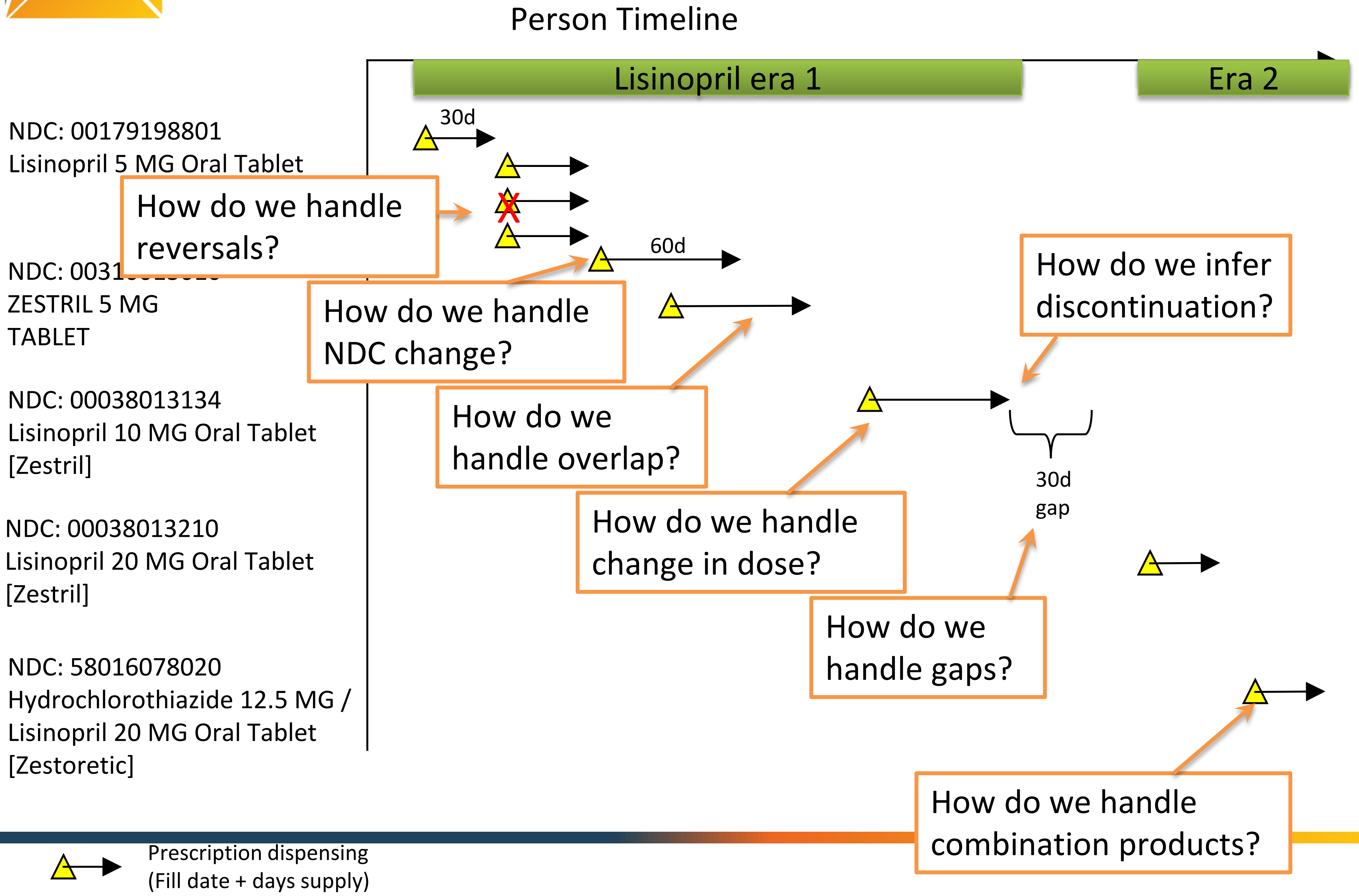


drug_era	
🔑	drug_era_id
	person_id
	drug_concept_id
	drug_era_start_date
	drug_era_end_date
	drug_exposure_count
	gap_days

- Standardized inference of length of exposure to product for all active ingredients

- Derived from records in DRUG_EXPOSURE under certain rules to produce continuous Drug Eras

Illustrating inferences needed within longitudinal pharmacy claims data for one patient



Person Timeline

Lisinopril era 1

Era 2

NDC: 00179198801
Lisinopril 5 MG Oral Tablet

NDC: 0031
ZESTRIL 5 MG
TABLET

NDC: 00038013134
Lisinopril 10 MG Oral Tablet
[Zestril]

NDC: 00038013210
Lisinopril 20 MG Oral Tablet
[Zestril]

NDC: 58016078020
Hydrochlorothiazide 12.5 MG /
Lisinopril 20 MG Oral Tablet
[Zestoretic]

How do we handle reversals?

How do we handle NDC change?

How do we handle overlap?

How do we handle change in dose?

How do we handle gaps?

How do we infer discontinuation?

How do we handle combination products?

Prescription dispensing
(Fill date + days supply)

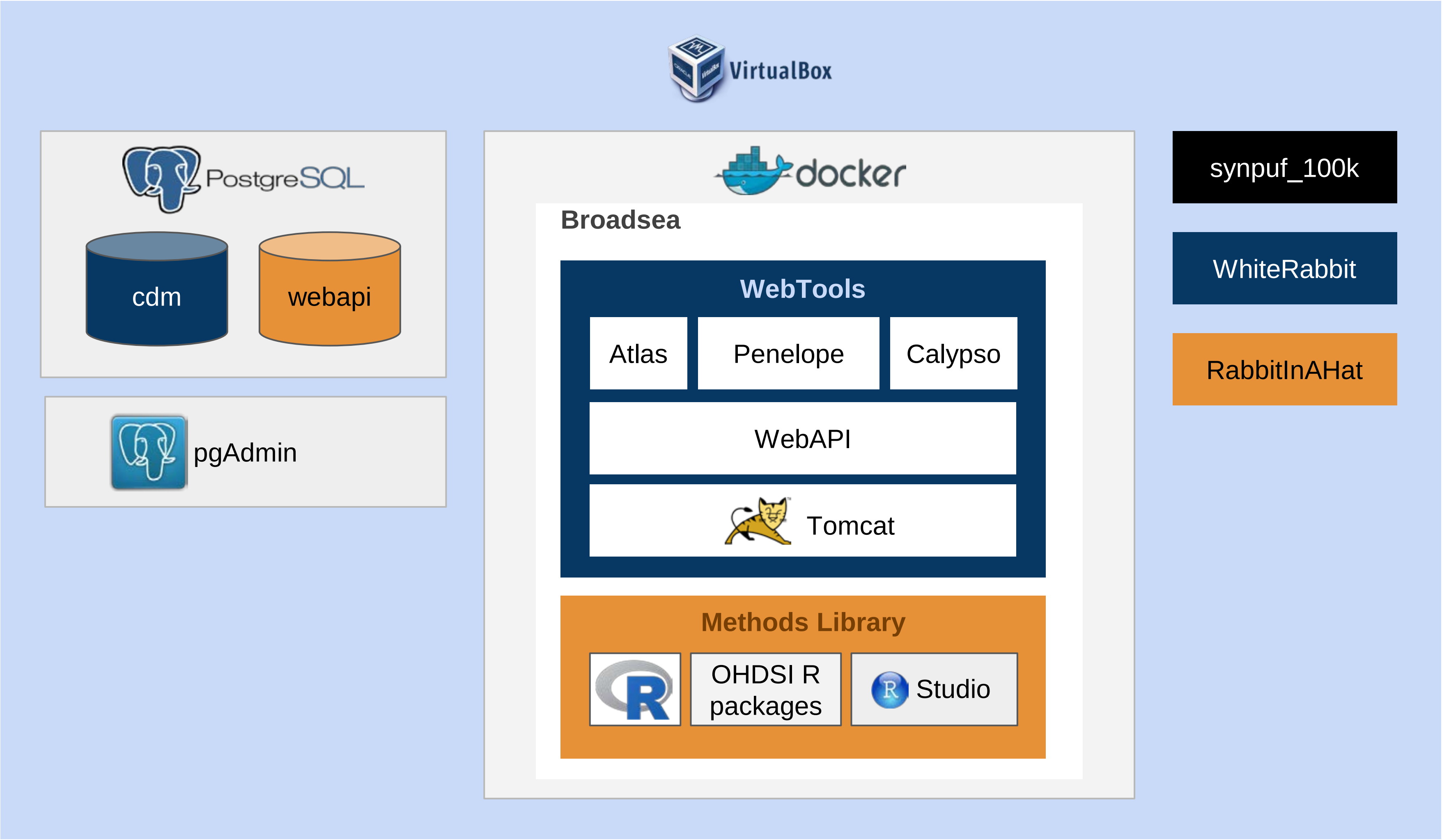
What makes OMOP CDM unique

- Specialized CDM - reflective of clinical domain, granular, well structured
- Vocabulary - uniformly structured and well curated
- Information Model - formalized connection between data model and conceptual model (Vocabulary)
- Specialized yet Extendable – new attributes and concepts can be added
- Supportive Community of developers and researchers
- Development driven by analytic use cases

Foundational

How to retrieve data from OMOP CDM

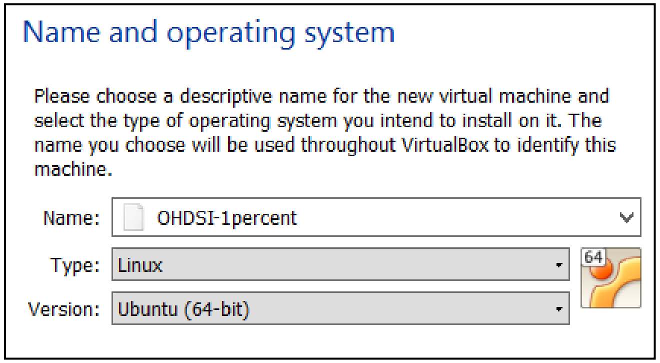
OHDSI in a Box



OHDSI in a Box – Setup

1. Open VM VirtualBox Manager

2. Click on New



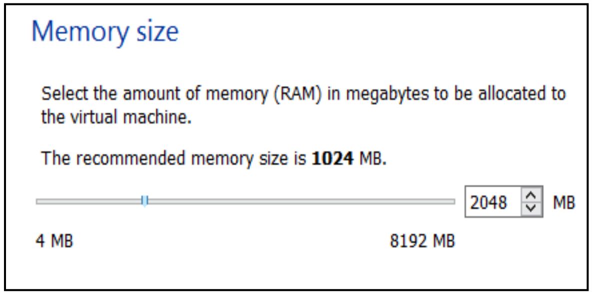
Name and operating system

Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name: OHDSI-1percent

Type: Linux

Version: Ubuntu (64-bit)



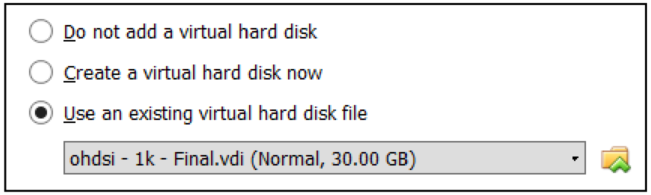
Memory size

Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.

The recommended memory size is **1024** MB.

2048 MB

4 MB 8192 MB

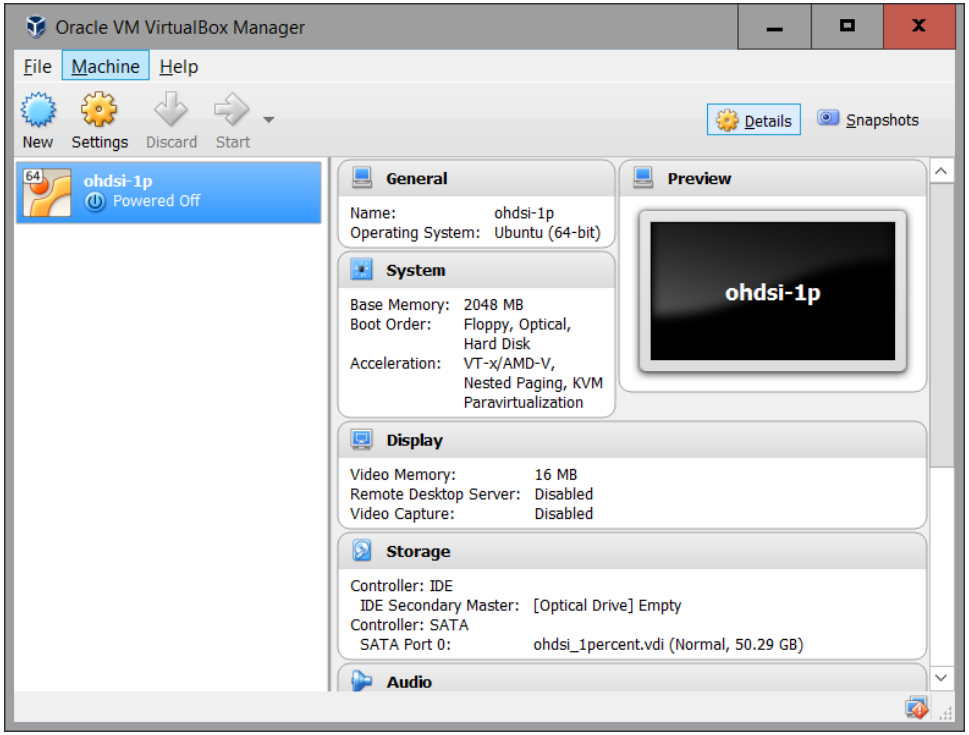


○ Do not add a virtual hard disk

○ Create a virtual hard disk now

● Use an existing virtual hard disk file

ohdsi - 1k - Final.vdi (Normal, 30.00 GB)



Oracle VM VirtualBox Manager

ohdsi-1p
Powered Off

General

Name: ohdsi-1p
Operating System: Ubuntu (64-bit)

Preview

ohdsi-1p

System

Base Memory: 2048 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

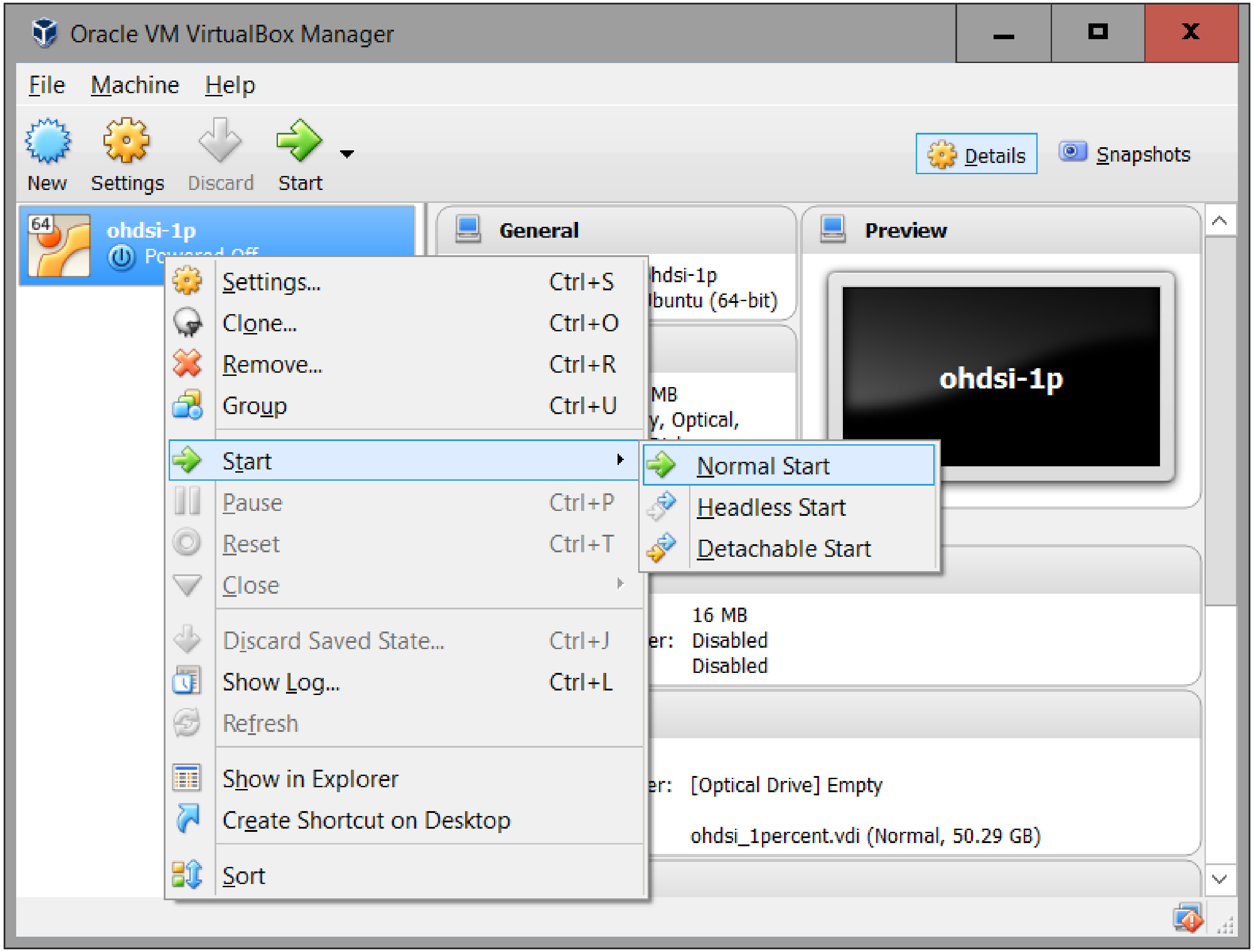
Video Memory: 16 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

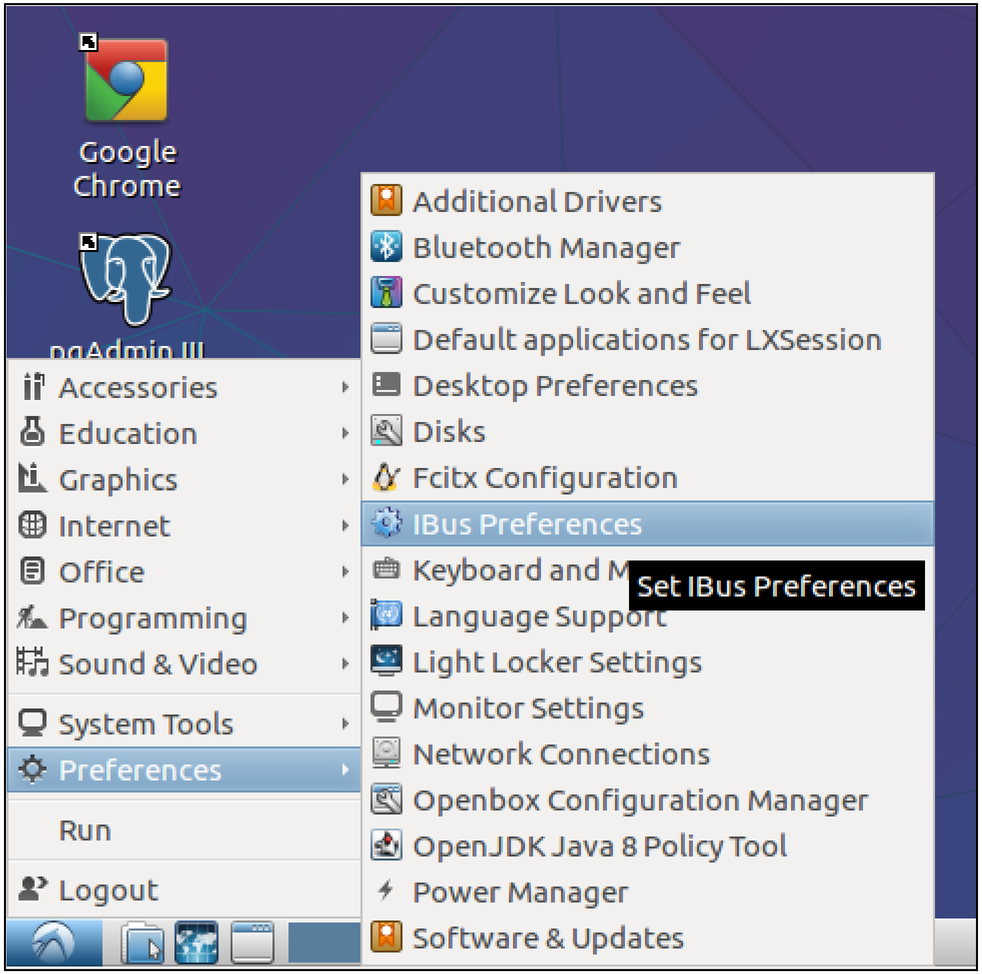
Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: ohdsi_1percent.vdi (Normal, 50.29 GB)

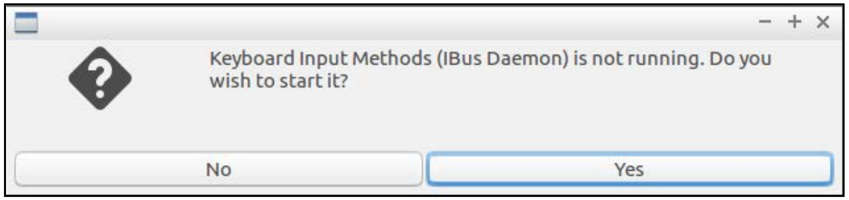
Audio

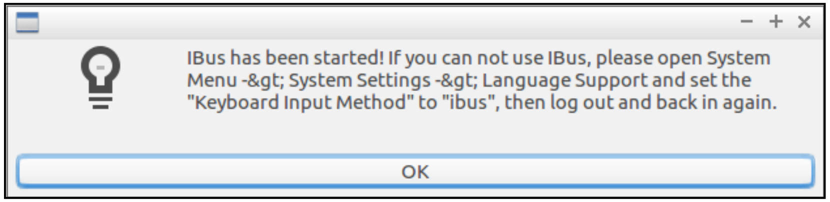
OHDSI in a Box – Start Up

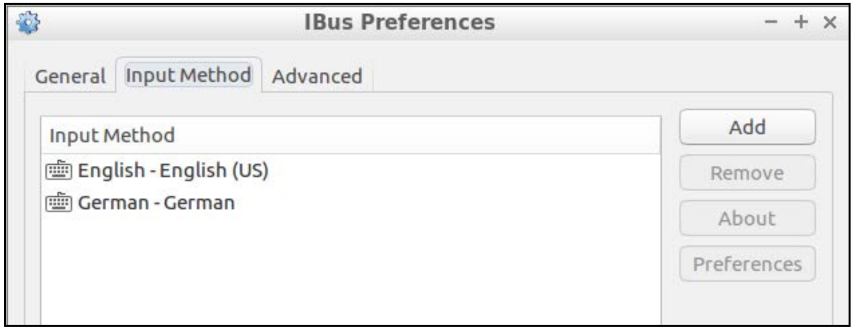


OHDSI in a Box – International Keyboards




Keyboard Input Methods (IBus Daemon) is not running. Do you wish to start it?

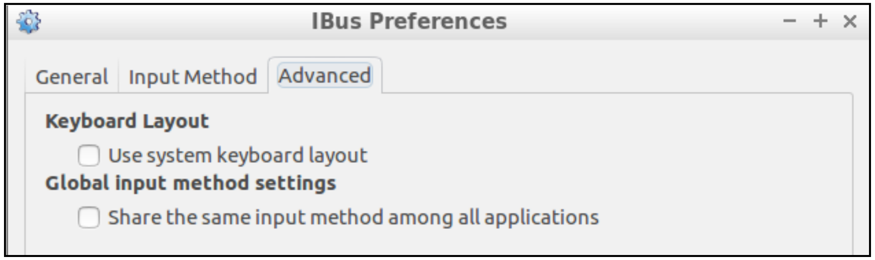

IBus has been started! If you can not use IBus, please open System Menu -&gt; System Settings -&gt; Language Support and set the "Keyboard Input Method" to "ibus", then log out and back in again.


IBus Preferences

Input Method

English - English (US)

German - German


IBus Preferences

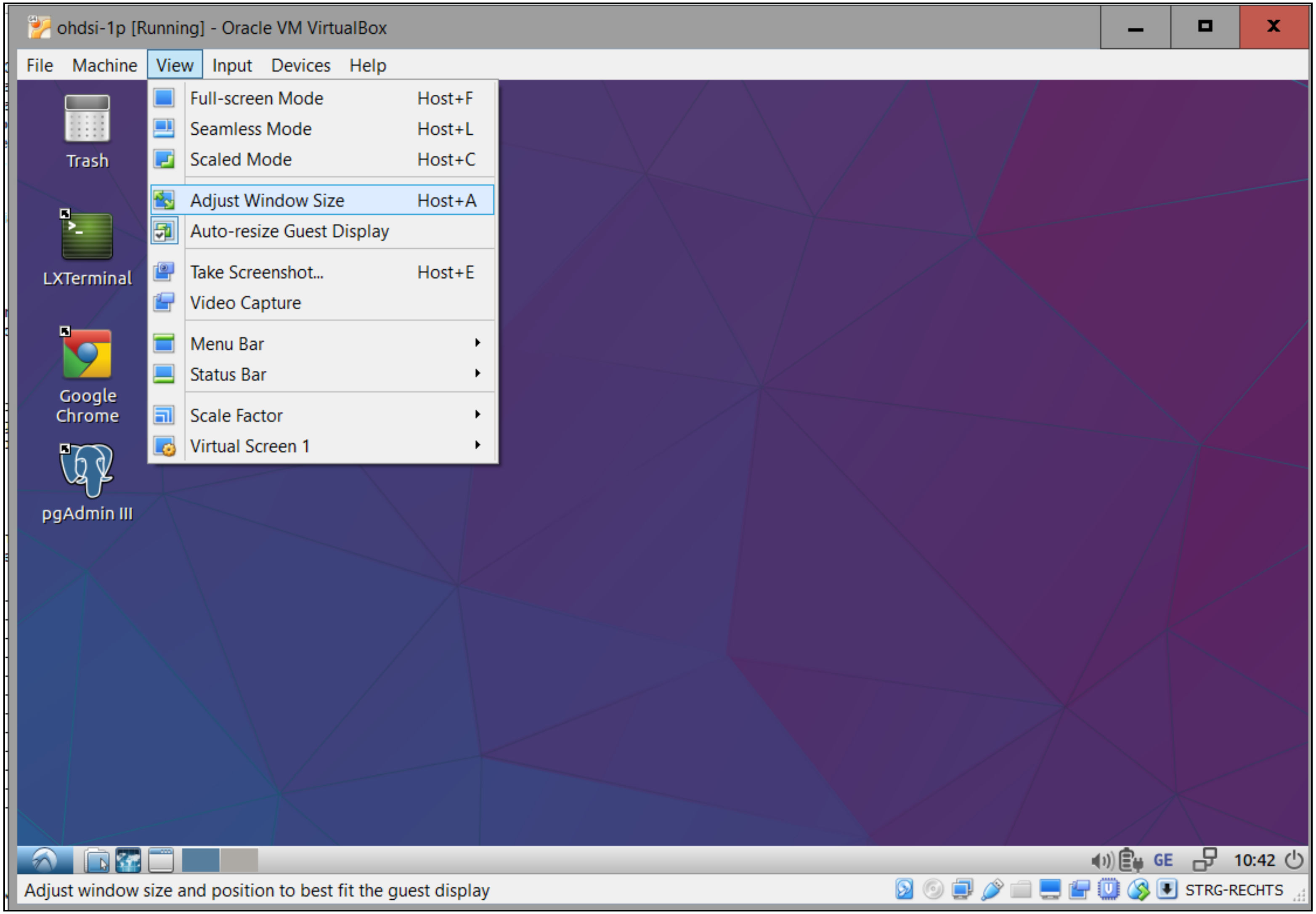
Keyboard Layout

☐ Use system keyboard layout

Global input method settings

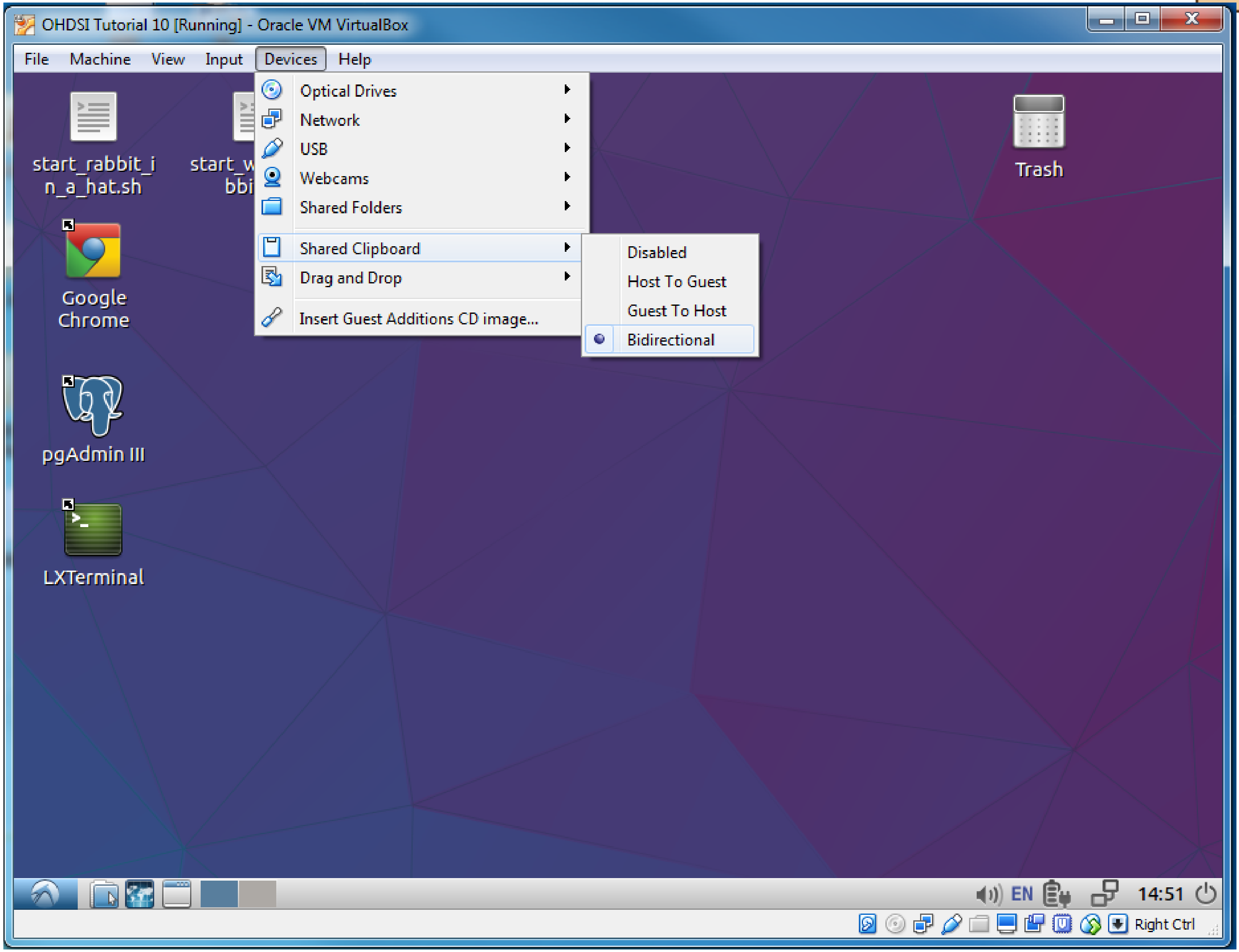
☐ Share the same input method among all applications

OHDSI in a Box – Adjust Resolution



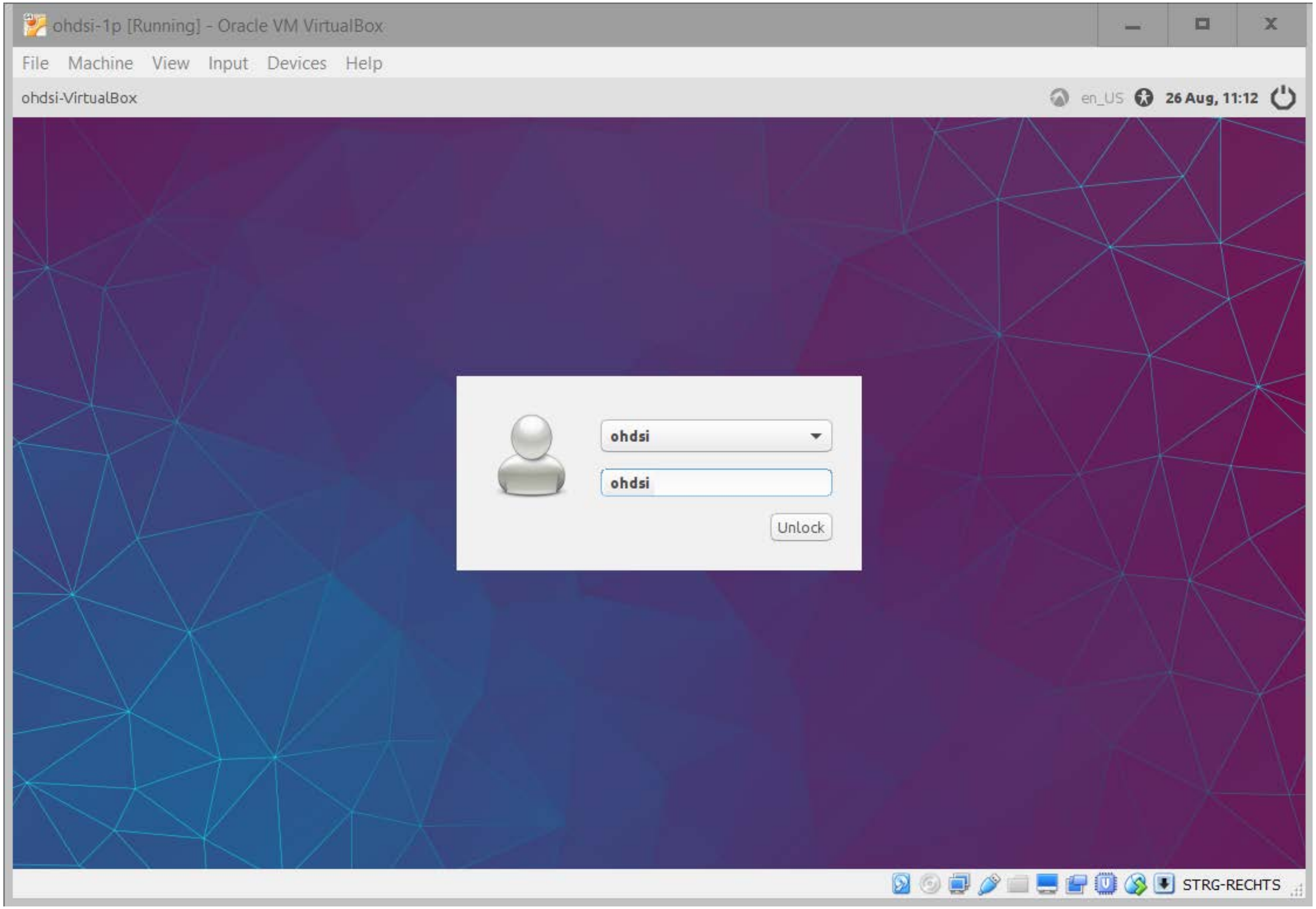
Adjust window size and position to best fit the guest display

OHDSI in a Box – Clipboard



Shared Clipboard
Disabled
Host To Guest
Guest To Host
● Bidirectional

OHDSI in a Box – Timeout

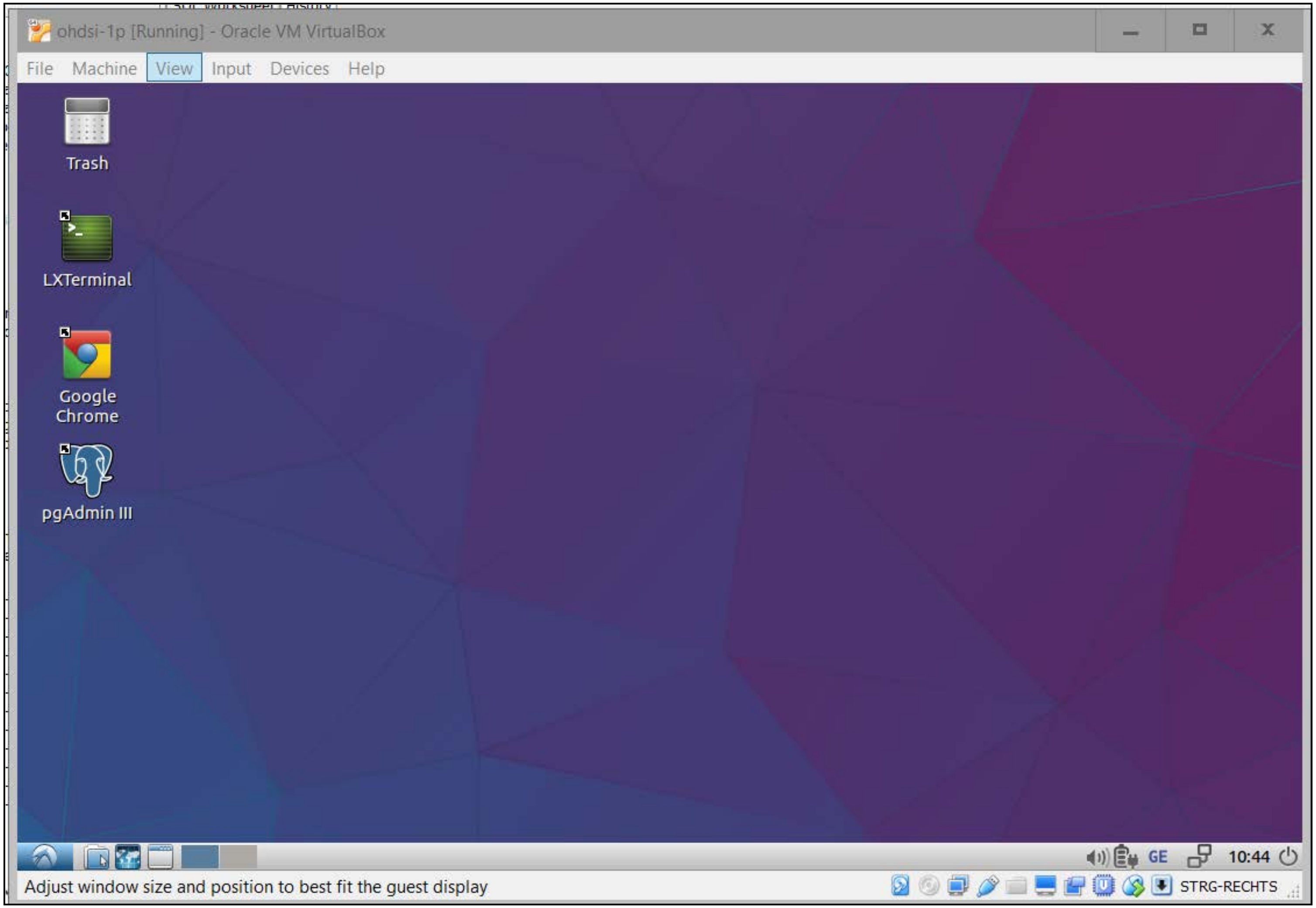


ohdsi-1p [Running] - Oracle VM VirtualBox

ohdsi

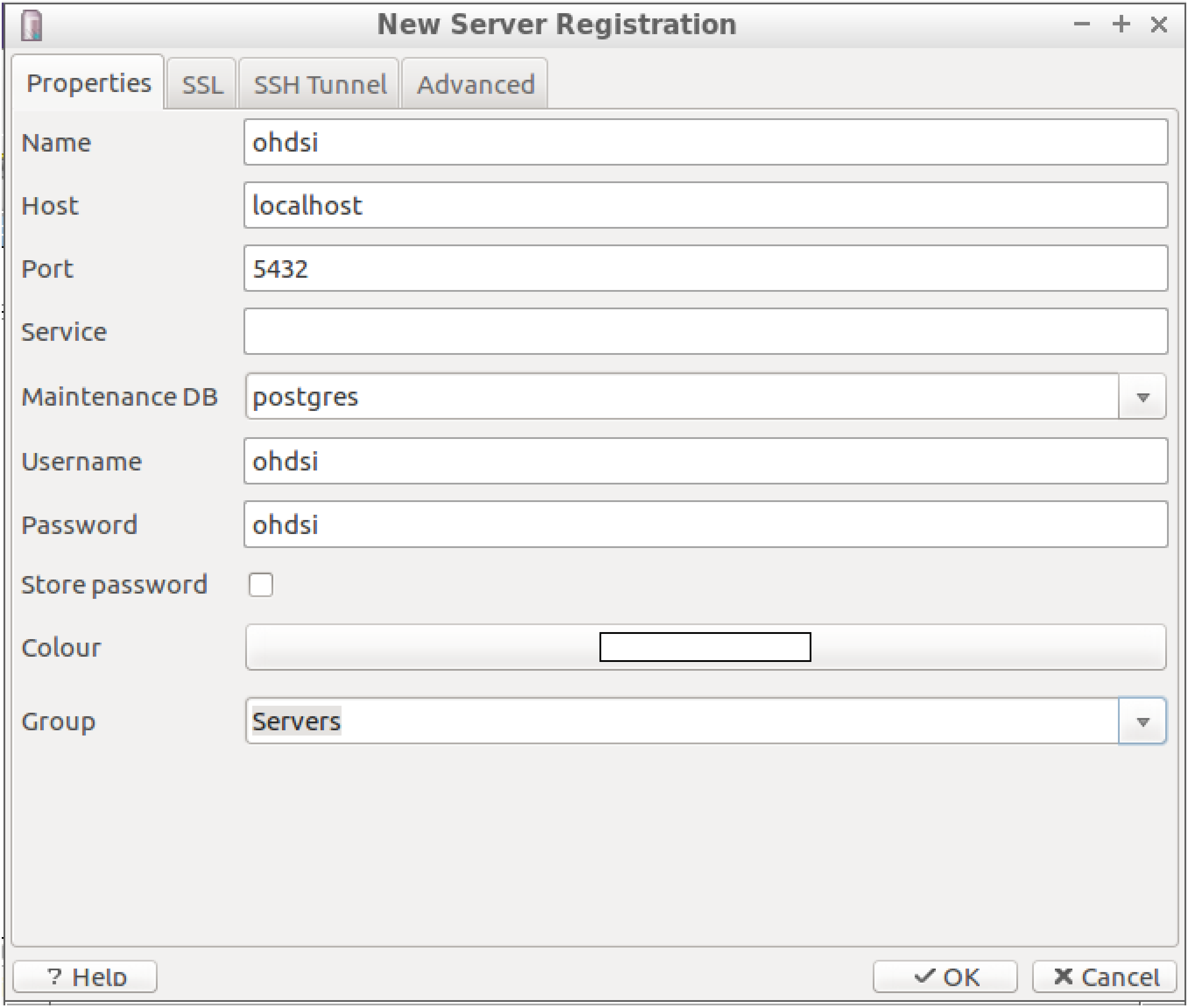
ohdsi

OHDSI in a Box – Ready



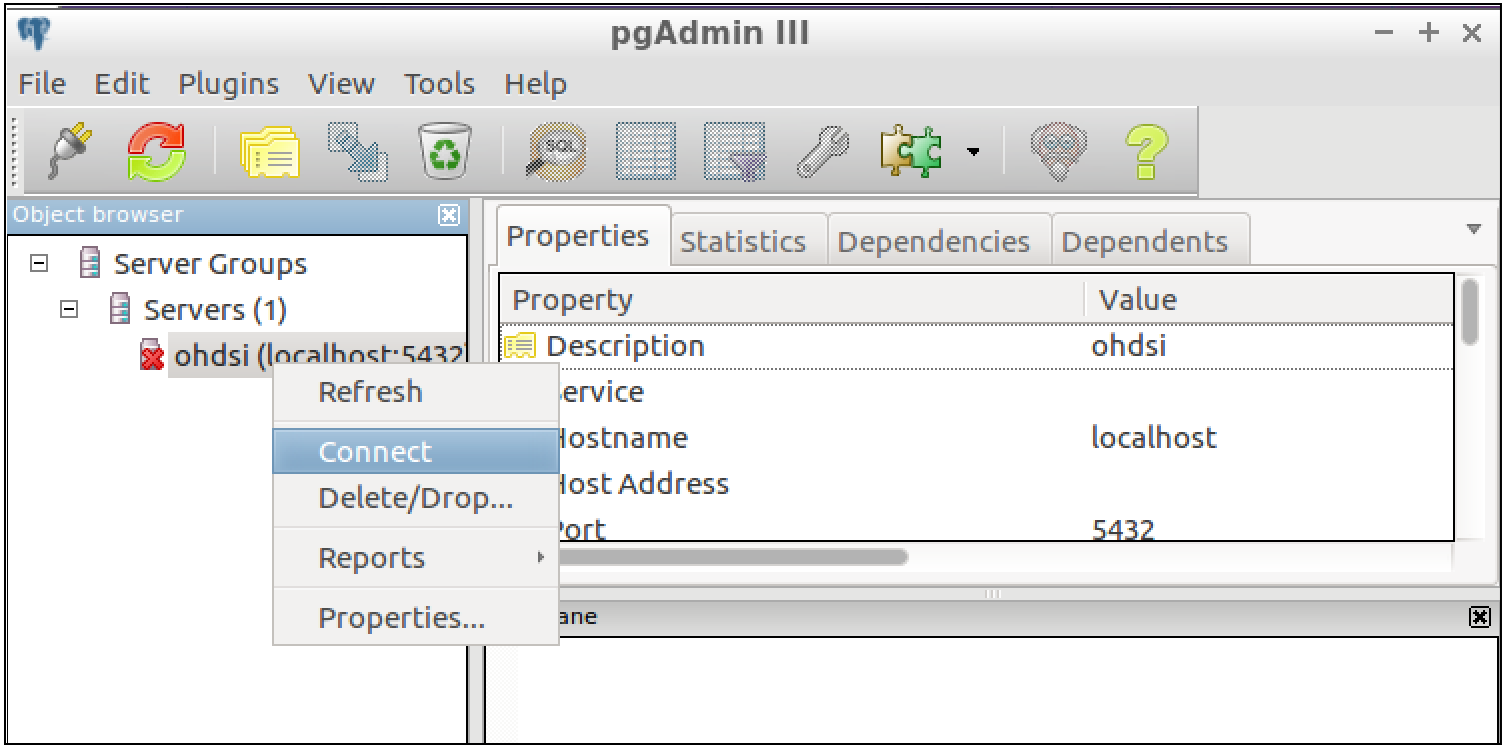
Adjust window size and position to best fit the guest display

CDM Database – pgAdmin III New Server



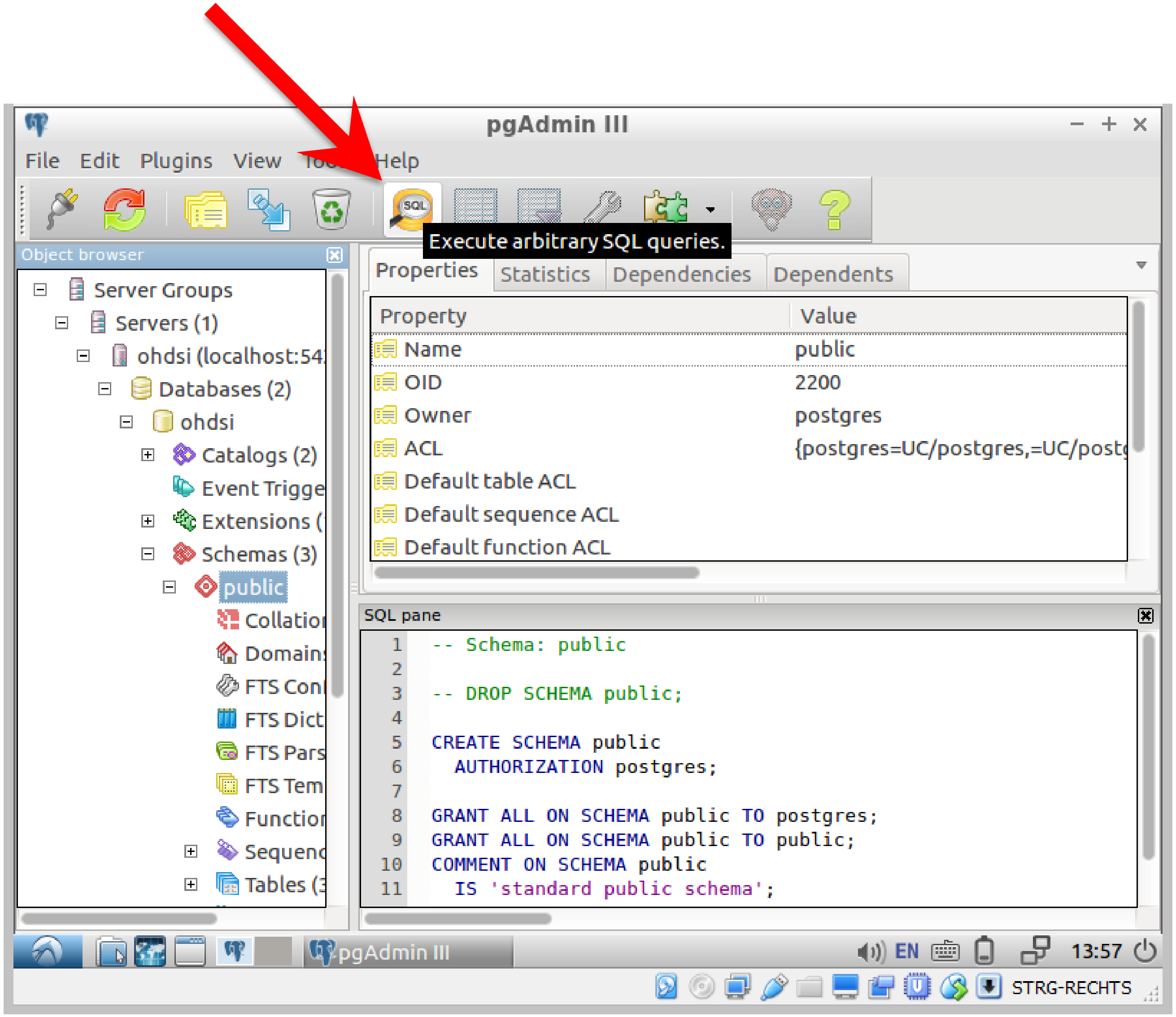
Name	ohdsi
Host	localhost
Port	5432
Service	
Maintenance DB	postgres
Username	ohdsi
Password	ohdsi
Store password	□
Colour	
Group	Servers

CDM Database – Connect



Property	Value
Description	ohdsi
ervice	
lostname	localhost
lost Address	
ort	5432

CDM Database – Open SQL Sheet



Property	Value
Name	public
OID	2200
Owner	postgres
ACL	{postgres=UC/postgres,=UC/postg
Default table ACL	
Default sequence ACL	
Default function ACL	

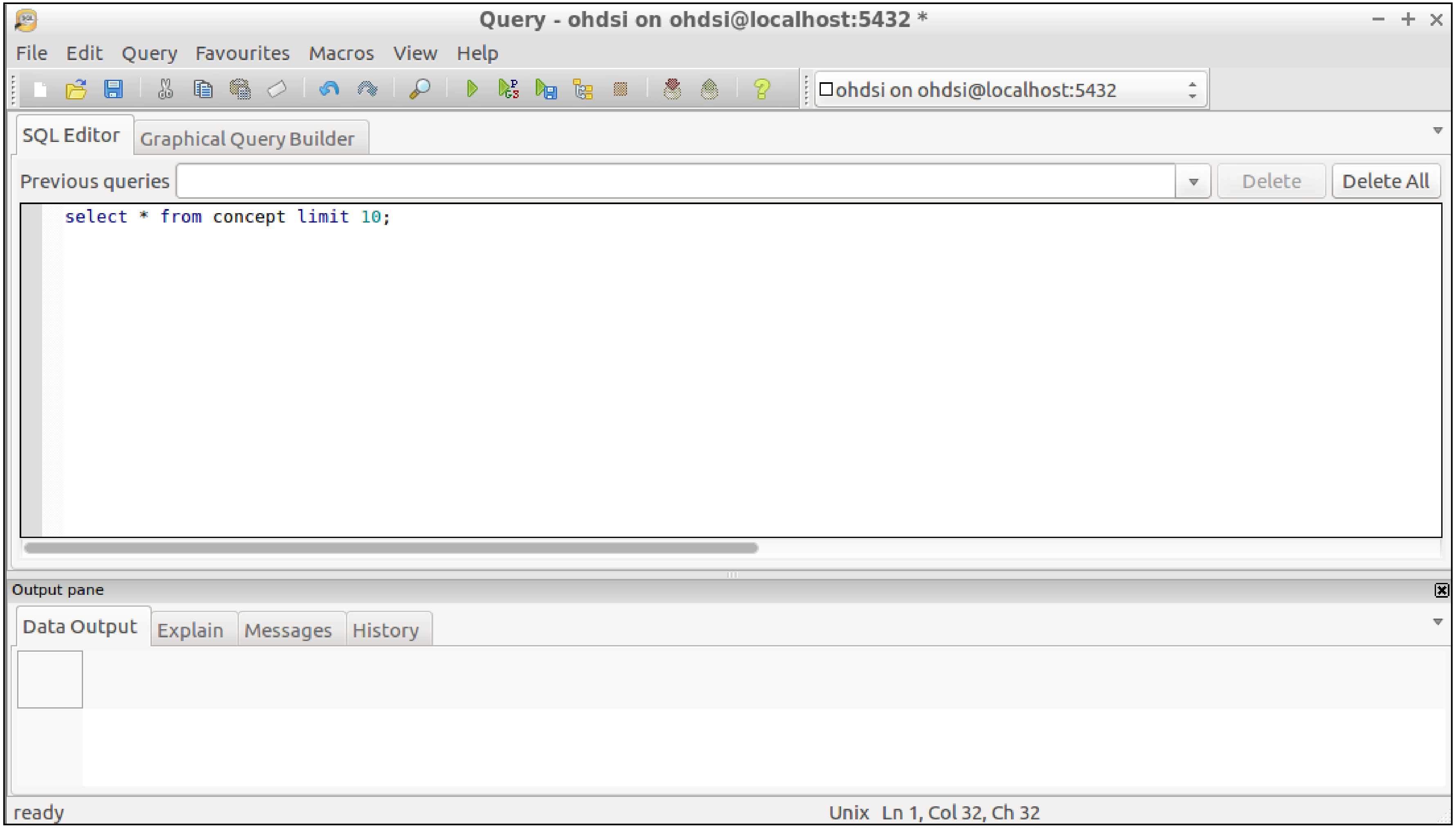
```
-- Schema: public

-- DROP SCHEMA public;

CREATE SCHEMA public
  AUTHORIZATION postgres;

GRANT ALL ON SCHEMA public TO postgres;
GRANT ALL ON SCHEMA public TO public;
COMMENT ON SCHEMA public
  IS 'standard public schema';
```

CDM Database – Ready



```
select * from concept limit 10;
```

Data Used for Demonstration

- Medicare Claims Synthetic Public Use Files (SynPUFs)
  - synthetic US Medicare insurance claims database
  - Medicare is a government based insurance program for primarily 65 and older but also individuals with disabilities
  - SynPUF not for research but rather demonstration/development purposes
  - Has been converted to the Common Data Model

CMS.gov
Centers for Medicare & Medicaid Services

https://www.cms.gov/research-statistics-data-and-systems/downloadable-public-use-files/synpufs/

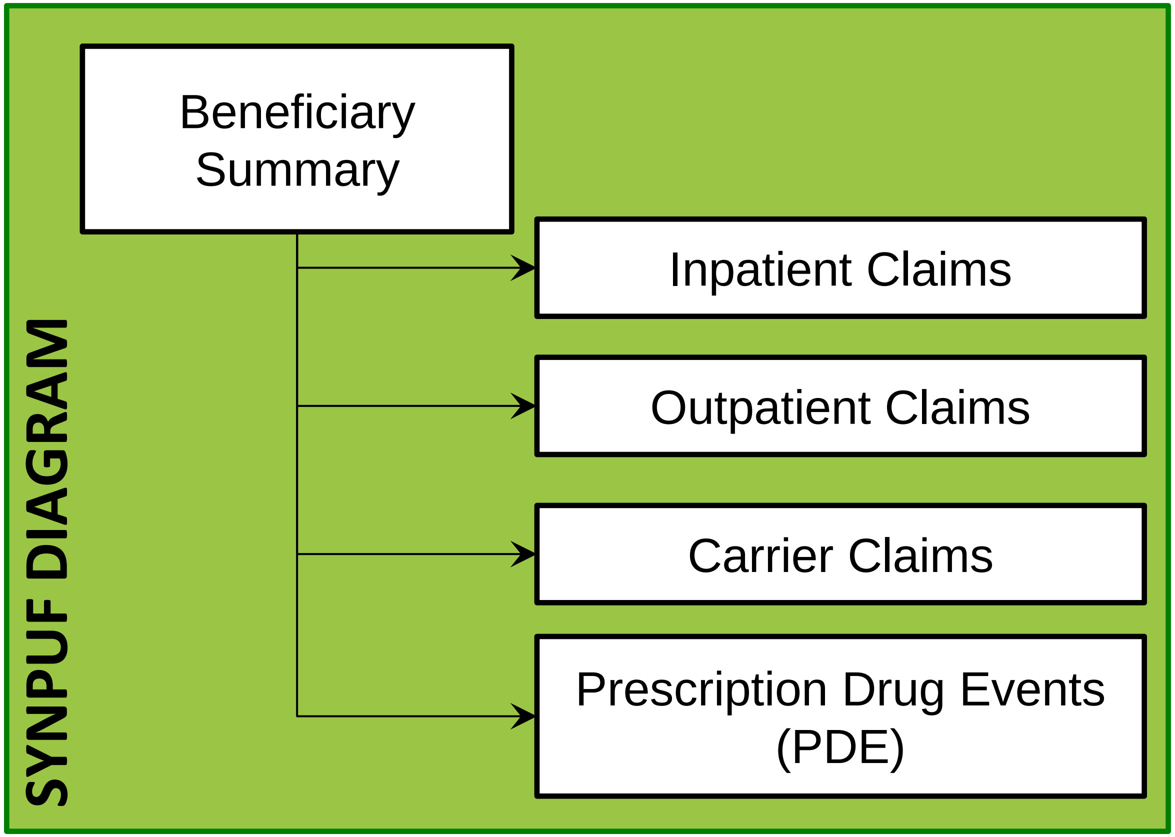
Data Used for Demonstration

- Five types of data:

	DE-SynPUF	Unit of record	Number of Records 2008	Number of Records 2009	Number of Records 2010
1	*Beneficiary Summary*	Benefi-ciary	2,326,856	2,291,320	2,255,098
2	*Inpatient Claims*	claim	547,800	504,941	280,081
3	*Outpatient Claims*	claim	5,673,808	6,519,340	3,633,839
4	*Carrier Claims*	claim	34,276,324	37,304,993	23,282,135
5	*Prescription Drug Events (PDE)*	event	39,927,827	43,379,293	27,778,849

https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html

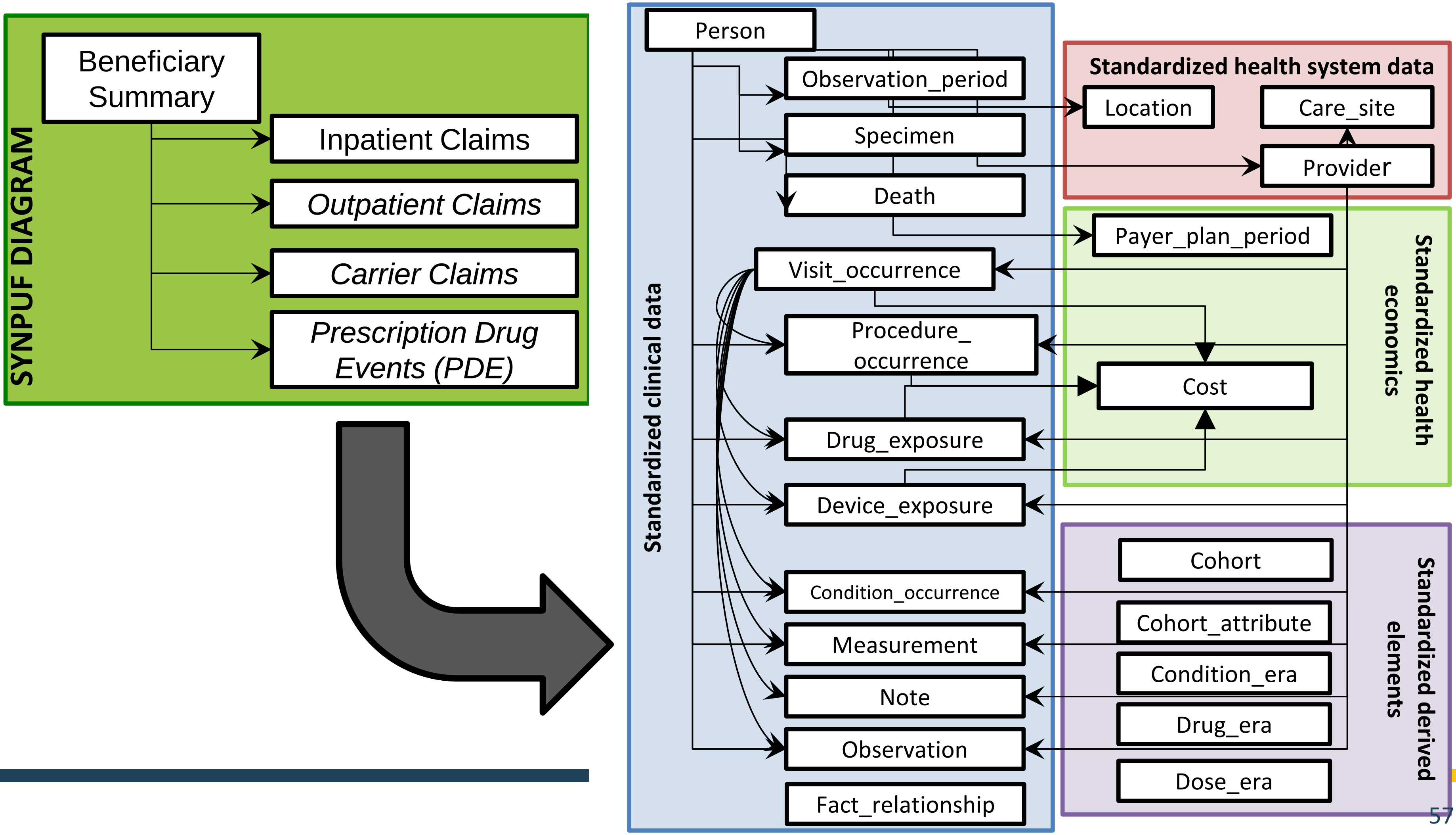
SynPUF High Level Diagram



Mapping SynPUF to CDM

SynPUF

CDM



SYNPUF DIAGRAM

Beneficiary Summary

Inpatient Claims

Outpatient Claims

Carrier Claims

Prescription Drug Events (PDE)

Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_ occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Some Example Questions

Ex 1

New Users of Warfarin

Ex 2

New Users of Warfarin
who are >=65?

Ex 3

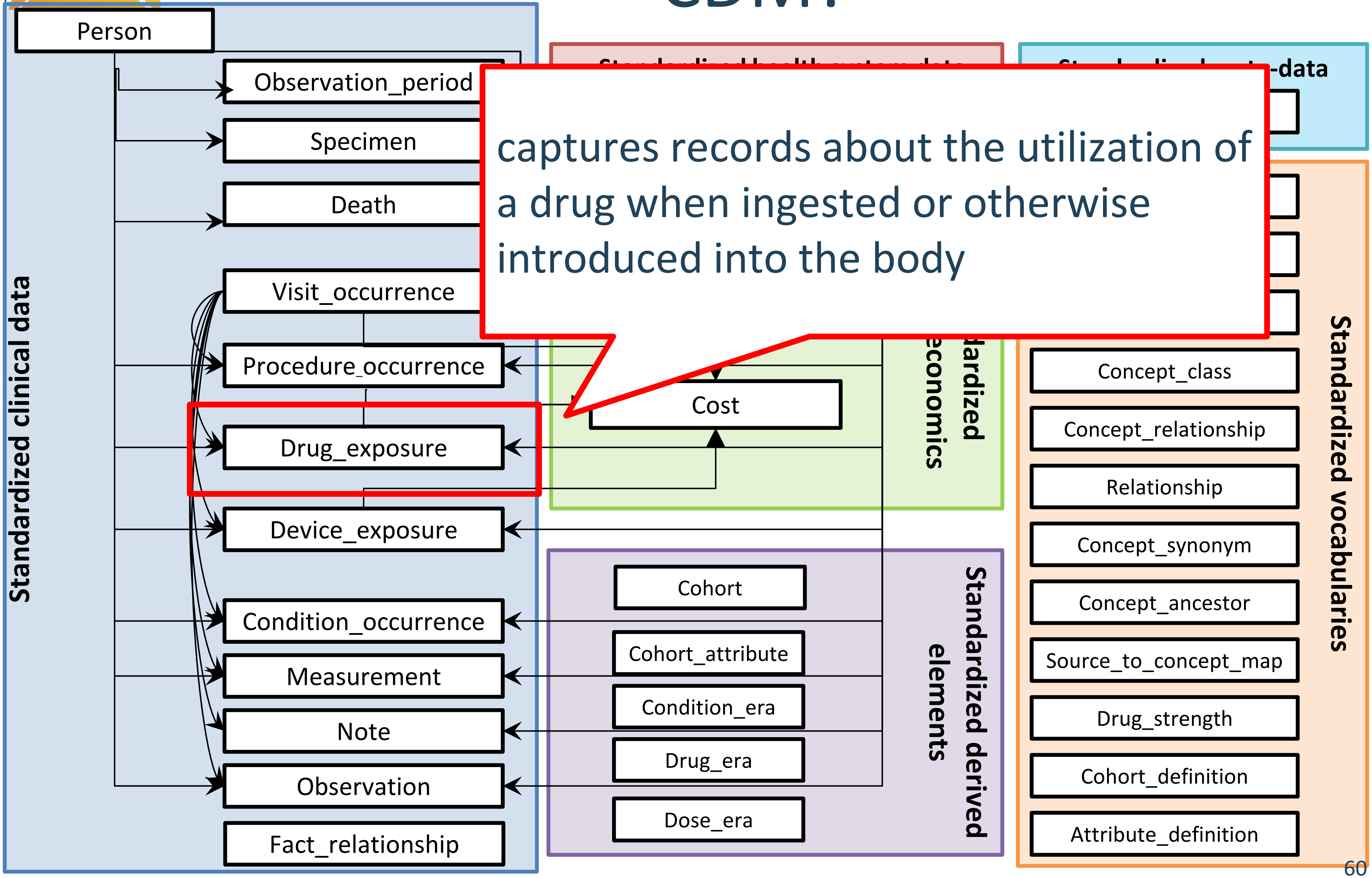
New Users of Warfarin
with prior Atrial Fibrillation?

Ex 1

New Users of Warfarin

- Warfarin is a blood thinner that is used to treat/prevent blood clots.
  - Where do you find drug data in the CDM?
  - What codes do I use to define drugs?
  - What does “New User” mean?

Where are Drug Exposures in the CDM?



captures records about the utilization of a drug when ingested or otherwise introduced into the body

Ex 1

How do I define Warfarin?

- When raw data is transformed into the CDM raw source codes are transformed into standard OMOP Vocabulary concepts

- In the CDM, we no longer care what source concepts existed in the raw data, we just need to use concept identifiers

- We can use the OMOP Vocabulary to identify all concepts that contain the ingredient warfarin

Ex 1

How do I define Warfarin?

SQL

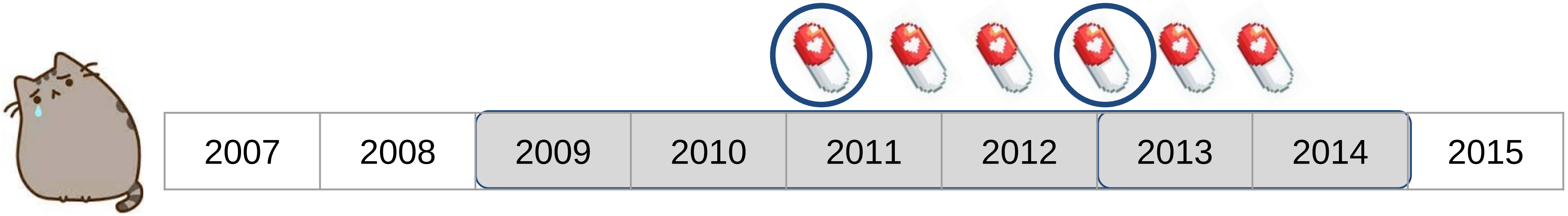
- Writing SQL Statement

- OHDSI Tool ATLAS

Ex 1

How do I define new users of a drug?

- someone who has recently started taking the drug, typically with a 6 or 12 month wash out

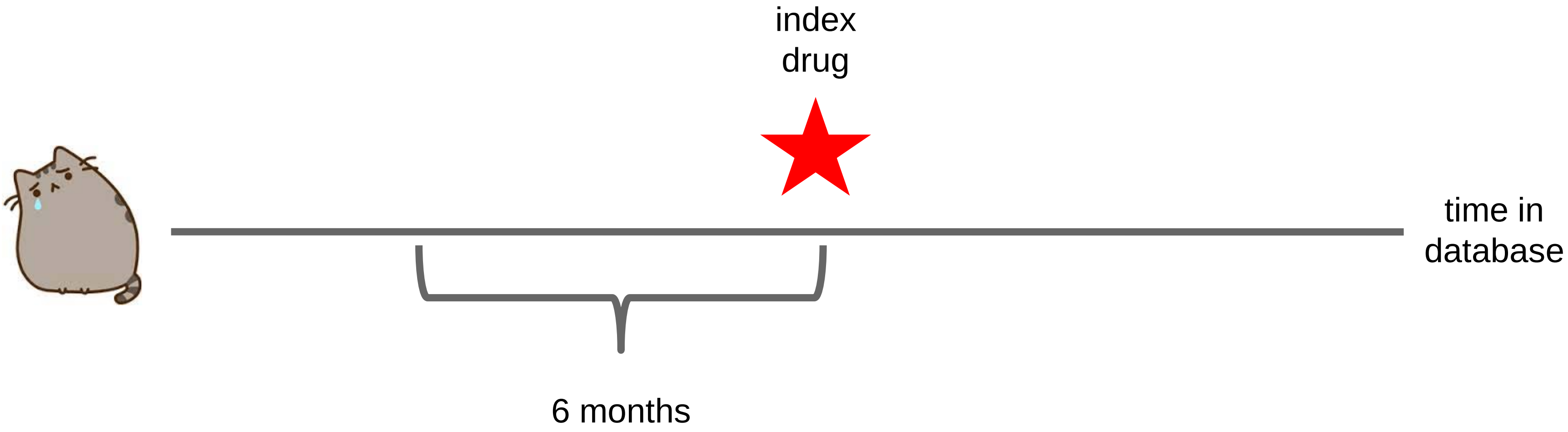


2007	2008	2009	2010	2011	2012	2013	2014	2015

Ex 1

How do I define new users of a drug?

- someone who has recently started taking the drug, typically with a 6 or 12 month wash out



Ex 1

What is Needed in the CDM?

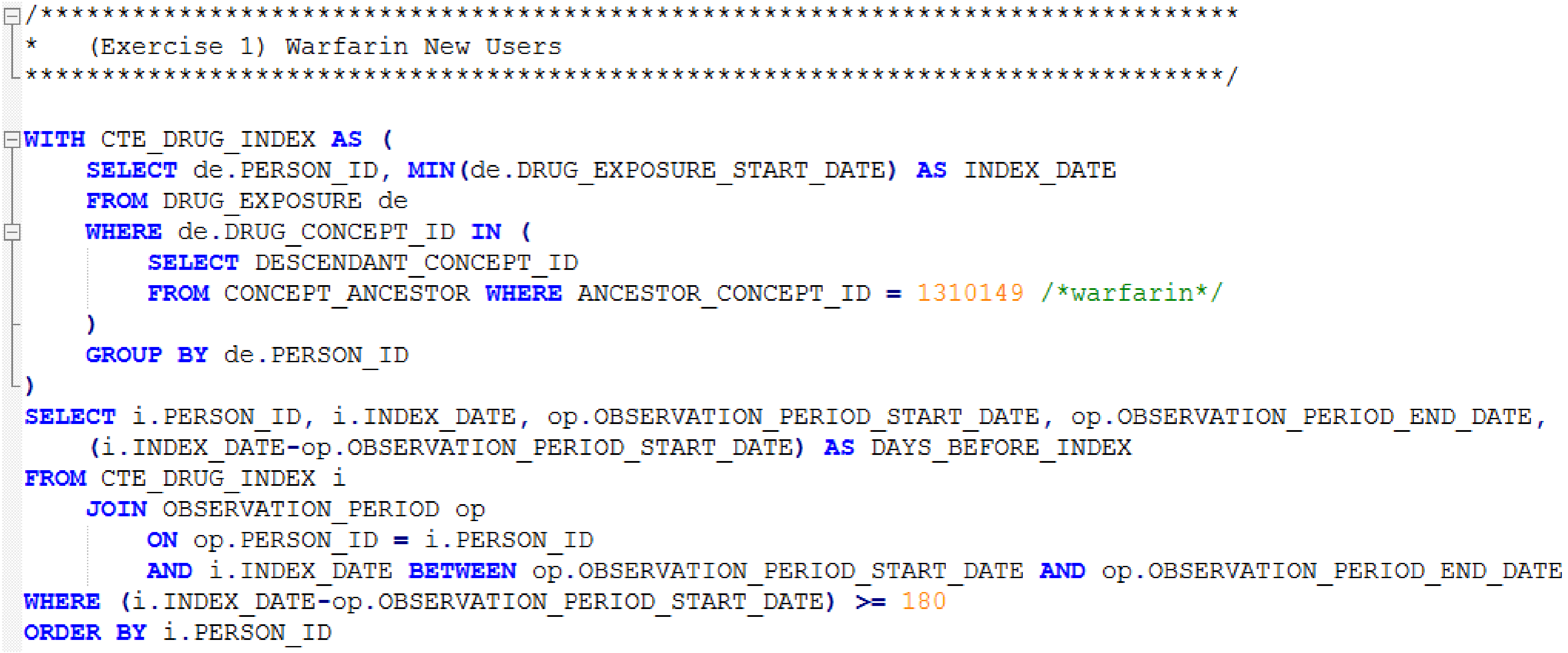
- **OMOP Vocabulary**
  to find the concepts

- **DRUG_EXPOSURE**
  to find individuals with exposure

- **OBSERVATION_PERIOD**
  to know people’s time within the database

Ex 1

New Users of Warfarin

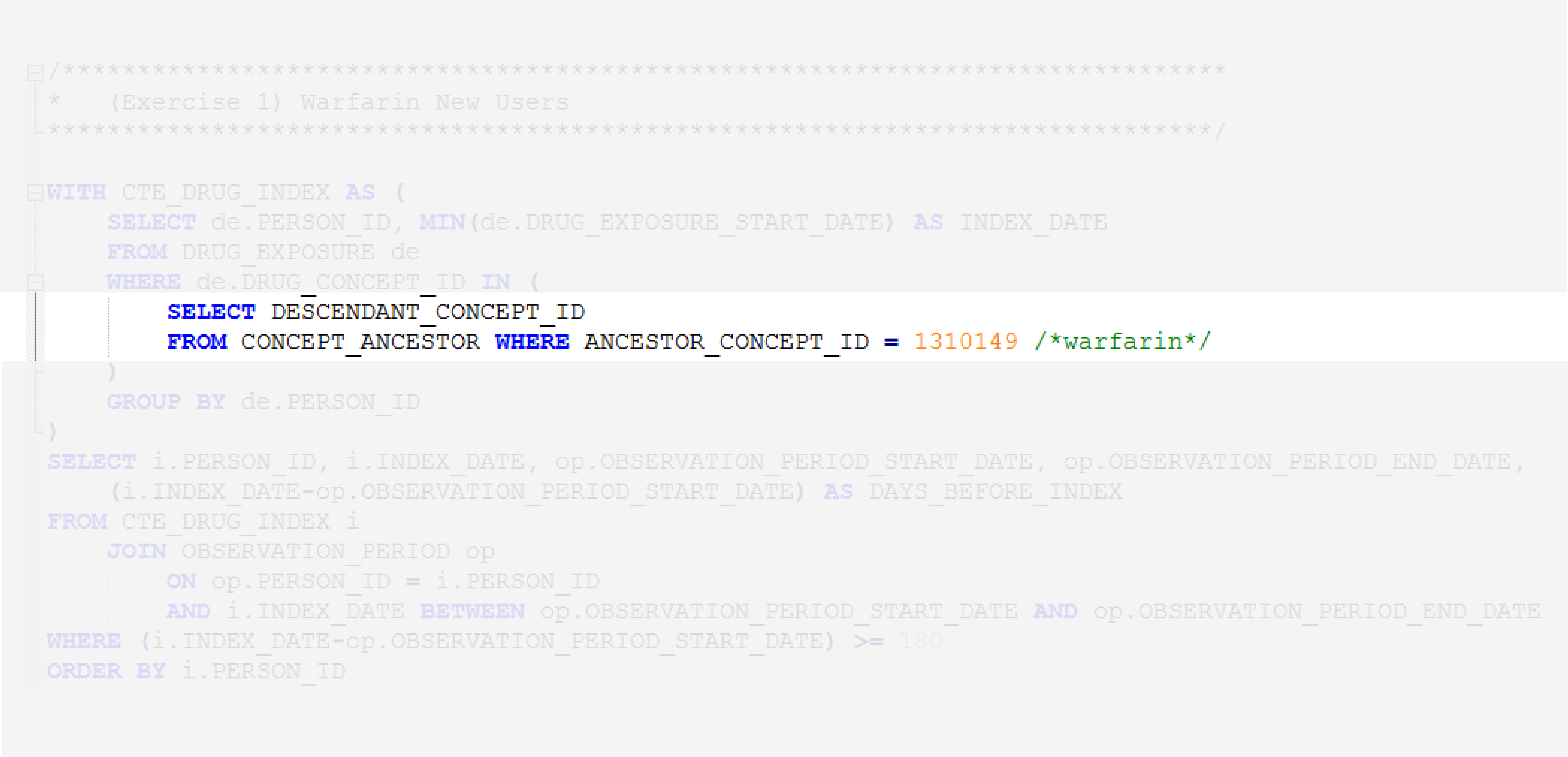


```sql
/******************************************************************************
*   (Exercise 1) Warfarin New Users
******************************************************************************/

WITH CTE_DRUG_INDEX AS (
    SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
    FROM DRUG_EXPOSURE de
    WHERE de.DRUG_CONCEPT_ID IN (
        SELECT DESCENDANT_CONCEPT_ID
        FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
    )
    GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
    (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
FROM CTE_DRUG_INDEX i
    JOIN OBSERVATION_PERIOD op
        ON op.PERSON_ID = i.PERSON_ID
        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
ORDER BY i.PERSON_ID
```

Ex 1

Step 1: Get the codes you need

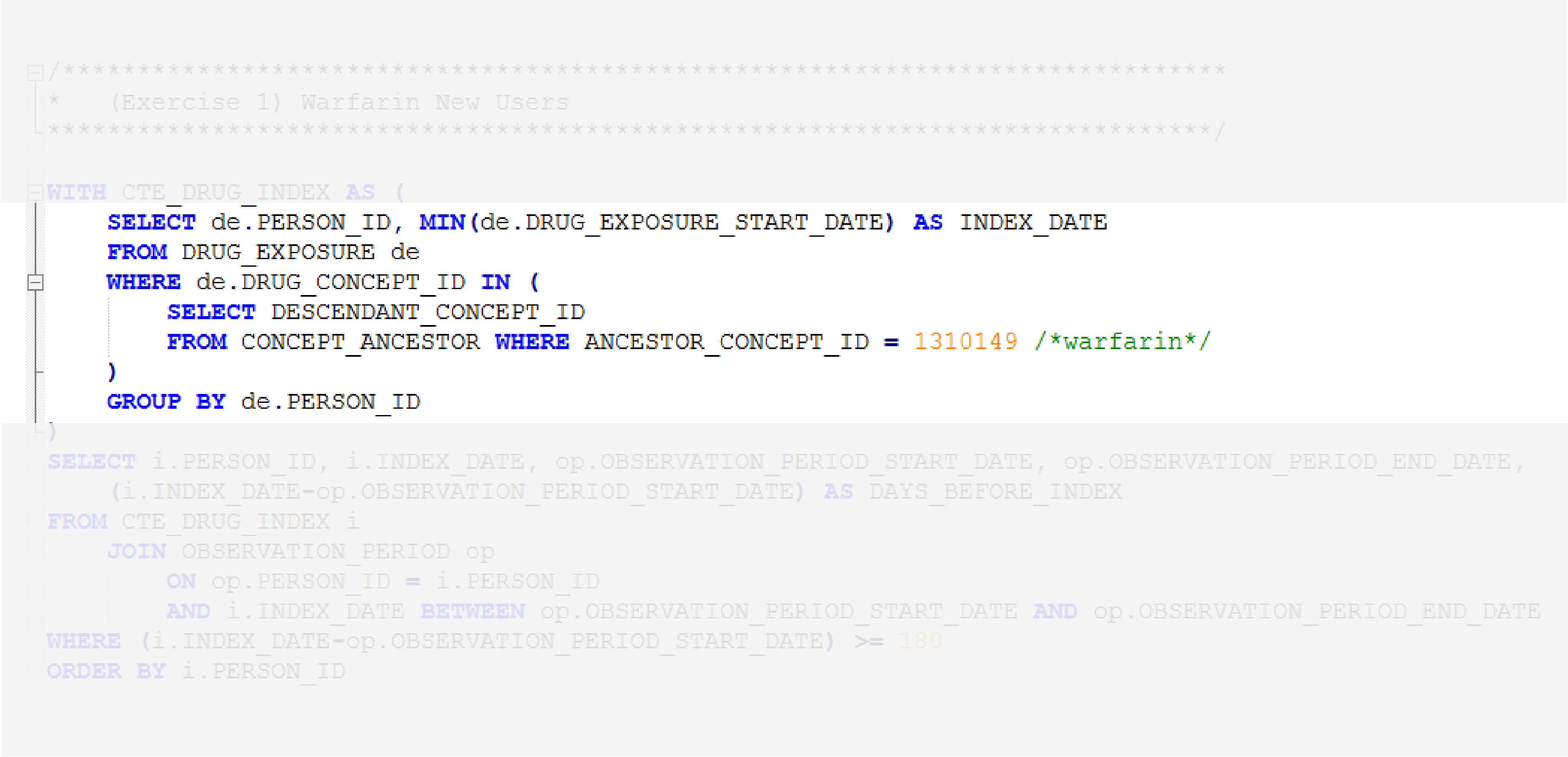


```
/*******************************************************************************
*    (Exercise 1) Warfarin New Users
*******************************************************************************/

WITH CTE_DRUG_INDEX AS (
    SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
    FROM DRUG_EXPOSURE de
    WHERE de.DRUG_CONCEPT_ID IN (
        SELECT DESCENDANT_CONCEPT_ID
        FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
    )
    GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
    (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
FROM CTE_DRUG_INDEX i
    JOIN OBSERVATION_PERIOD op
        ON op.PERSON_ID = i.PERSON_ID
        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
ORDER BY i.PERSON_ID
```

Ex 1

Step 2: Find Drug Exposures

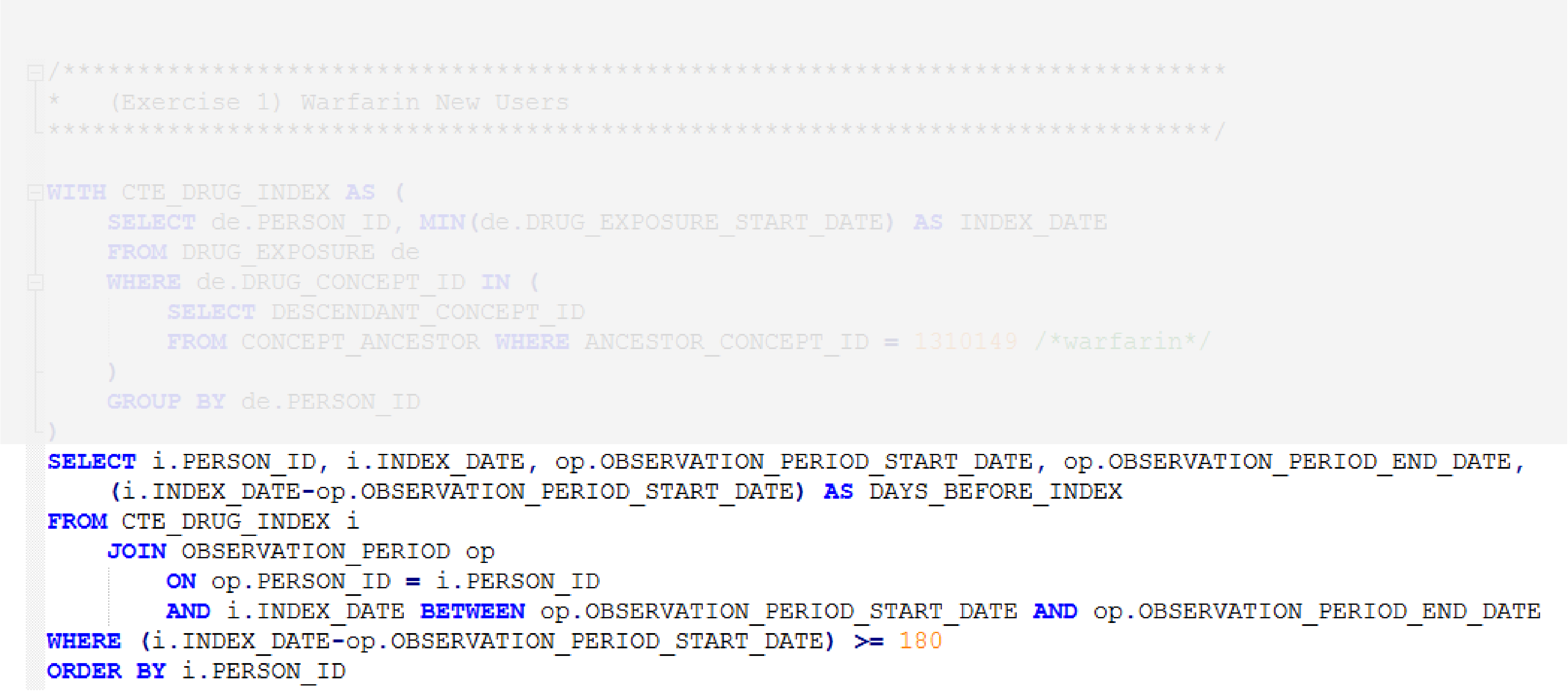


```
/****************************************************************************
*    (Exercise 1) Warfarin New Users
****************************************************************************/

WITH CTE_DRUG_INDEX AS (
    SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
    FROM DRUG_EXPOSURE de
    WHERE de.DRUG_CONCEPT_ID IN (
        SELECT DESCENDANT_CONCEPT_ID
        FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
    )
    GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
    (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
FROM CTE_DRUG_INDEX i
    JOIN OBSERVATION_PERIOD op
        ON op.PERSON_ID = i.PERSON_ID
        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
ORDER BY i.PERSON_ID
```

Ex 1

Step 3: Find New Users

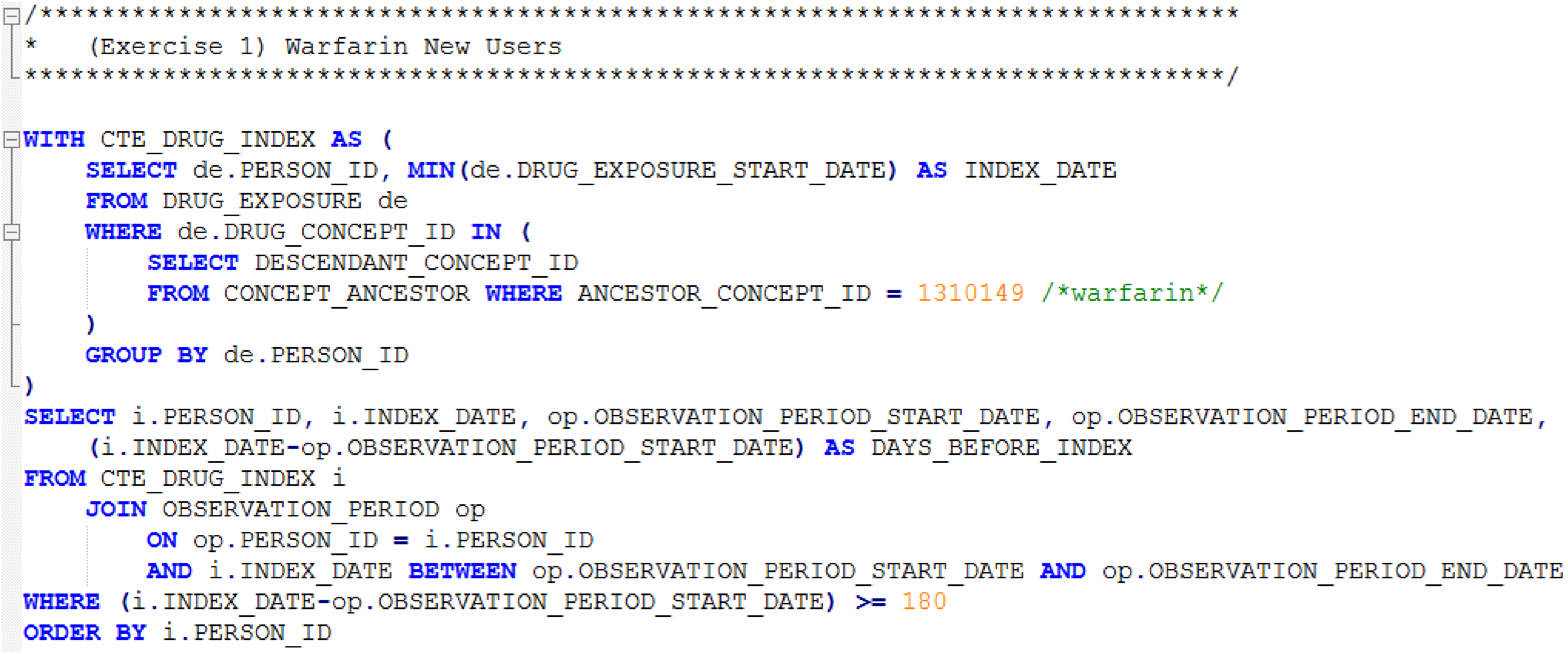


```sql
/***********************************************************************
*   (Exercise 1) Warfarin New Users
***********************************************************************/

WITH CTE_DRUG_INDEX AS (
    SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
    FROM DRUG_EXPOSURE de
    WHERE de.DRUG_CONCEPT_ID IN (
        SELECT DESCENDANT_CONCEPT_ID
        FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
    )
    GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
    (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
FROM CTE_DRUG_INDEX i
    JOIN OBSERVATION_PERIOD op
        ON op.PERSON_ID = i.PERSON_ID
        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
ORDER BY i.PERSON_ID
```

Ex 1

New Users of Warfarin



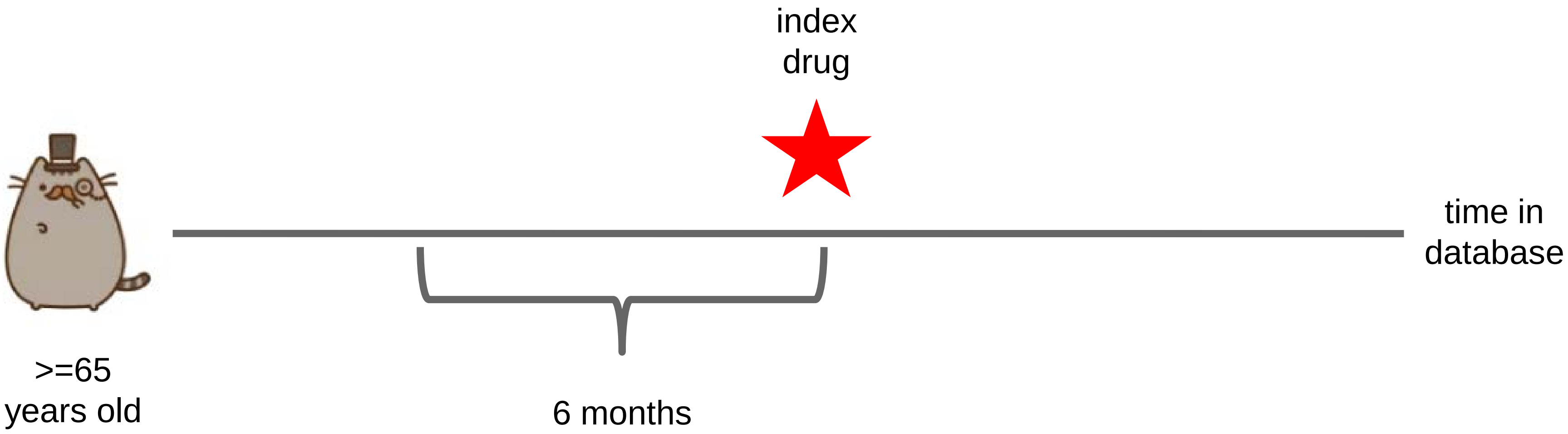
```sql
/*******************************************************************************
*   (Exercise 1) Warfarin New Users
*******************************************************************************/

WITH CTE_DRUG_INDEX AS (
    SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
    FROM DRUG_EXPOSURE de
    WHERE de.DRUG_CONCEPT_ID IN (
        SELECT DESCENDANT_CONCEPT_ID
        FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
    )
    GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
    (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
FROM CTE_DRUG_INDEX i
    JOIN OBSERVATION_PERIOD op
        ON op.PERSON_ID = i.PERSON_ID
        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
ORDER BY i.PERSON_ID
```

Ex 2

How do I define new users of warfarin who are >=65?

- someone who has recently started taking the drug, typically with a 6 or 12 month wash out



Ex 2

What is Needed in the CDM?

- **OMOP Vocabulary**
  to find the concepts

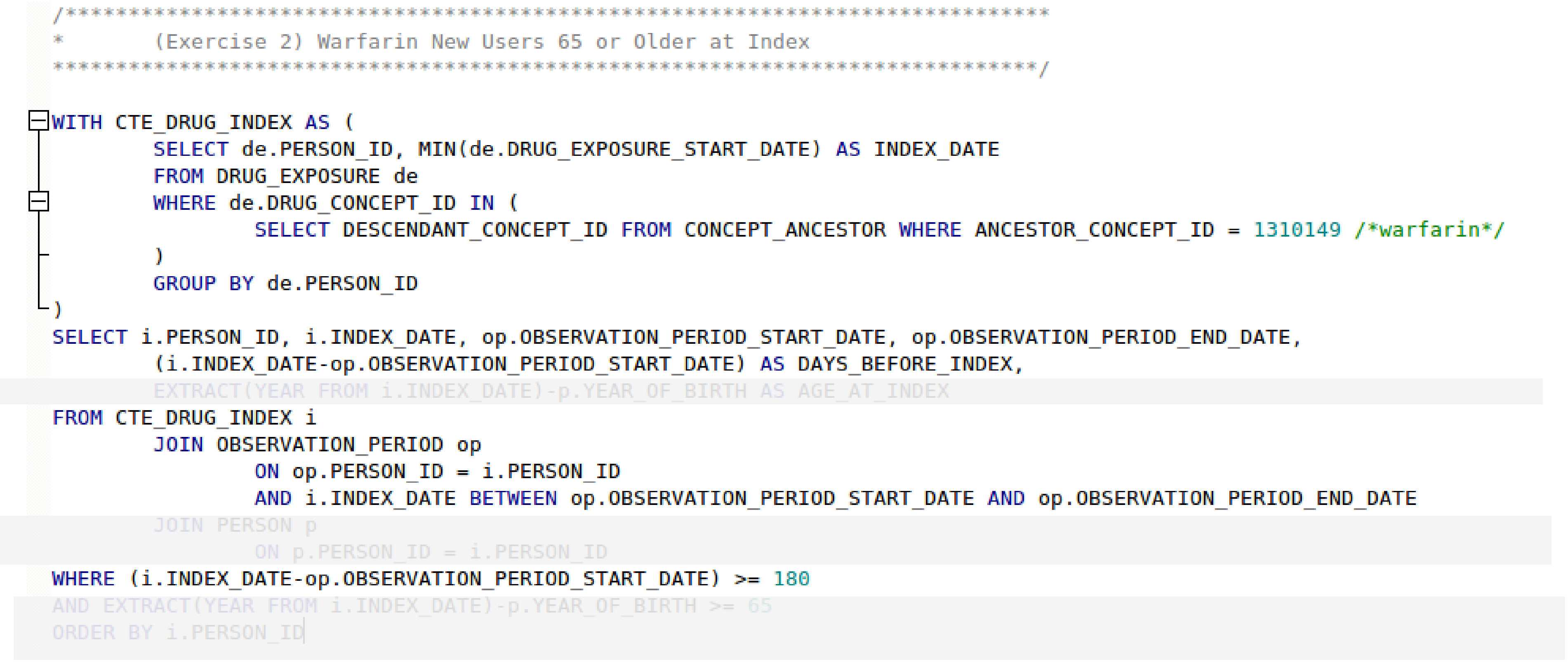
- **DRUG_EXPOSURE**
  to find individuals with exposure

- **OBSERVATION_PERIOD**
  to know people’s time within the database

- **PERSON**
  to know year of birth

Ex 2

Step 1: Start with the previous query

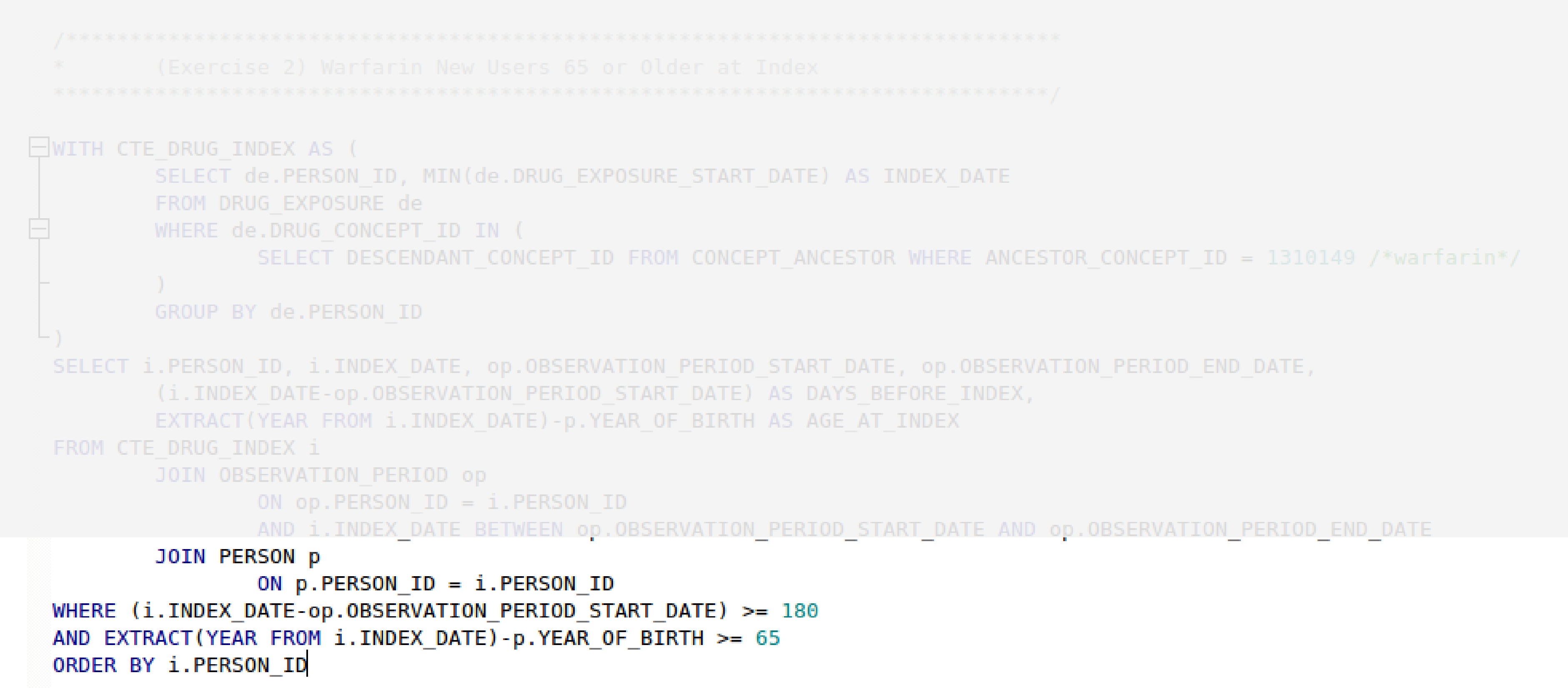


```
/*******************************************************************************
*       (Exercise 2) Warfarin New Users 65 or Older at Index
*******************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
        (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX,
        EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH AS AGE_AT_INDEX
FROM CTE_DRUG_INDEX i
        JOIN OBSERVATION_PERIOD op
                ON op.PERSON_ID = i.PERSON_ID
                AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        JOIN PERSON p
                ON p.PERSON_ID = i.PERSON_ID
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
AND EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH >= 65
ORDER BY i.PERSON_ID
```

Ex 2

Step 2: Add the Person Table to calculate age

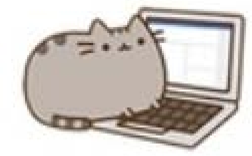


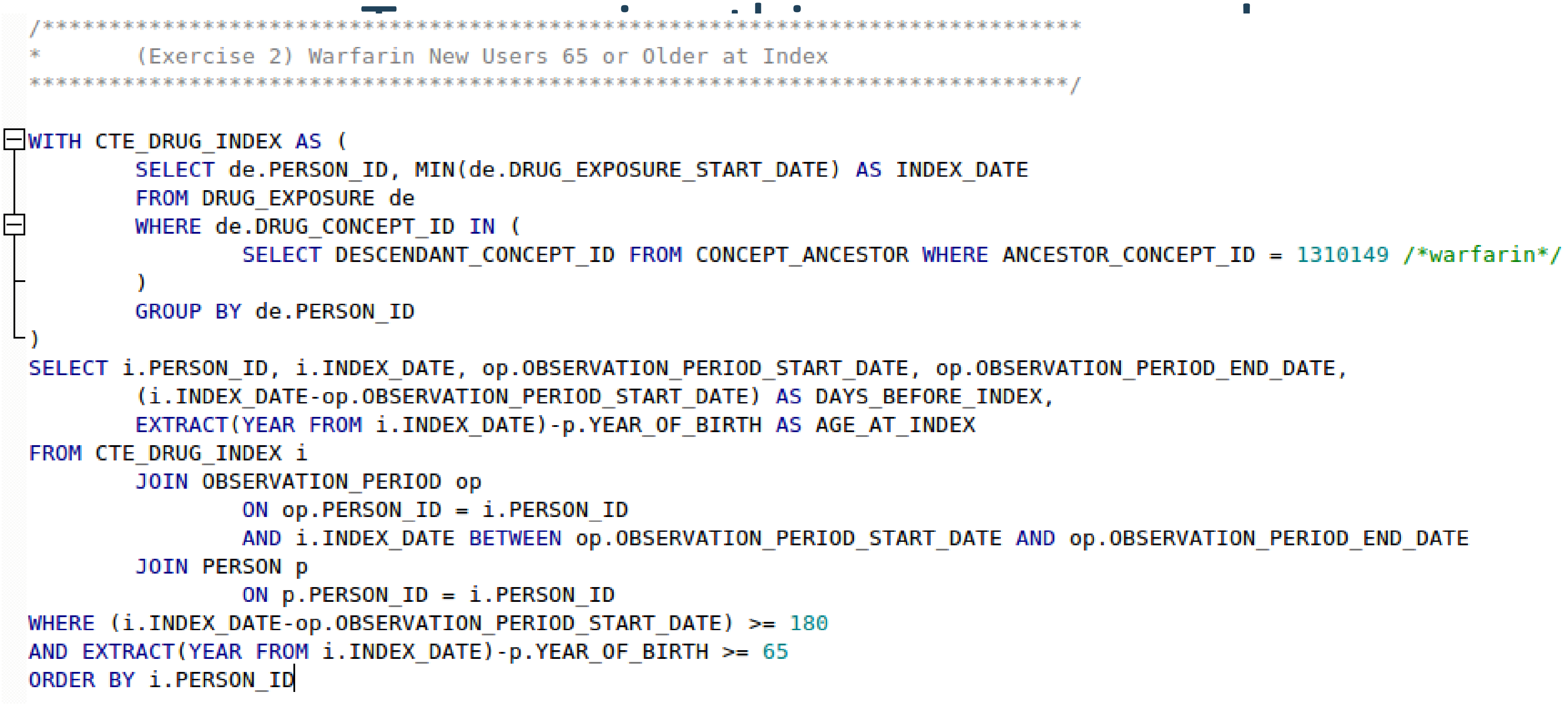
```
/*****************************************************************************
*       (Exercise 2) Warfarin New Users 65 or Older at Index
*****************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
        (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX,
        EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH AS AGE_AT_INDEX
FROM CTE_DRUG_INDEX i
        JOIN OBSERVATION_PERIOD op
                ON op.PERSON_ID = i.PERSON_ID
                AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        JOIN PERSON p
                ON p.PERSON_ID = i.PERSON_ID
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
AND EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH >= 65
ORDER BY i.PERSON_ID
```

Ex 2

New Users of Warfarin <= 65 years of age



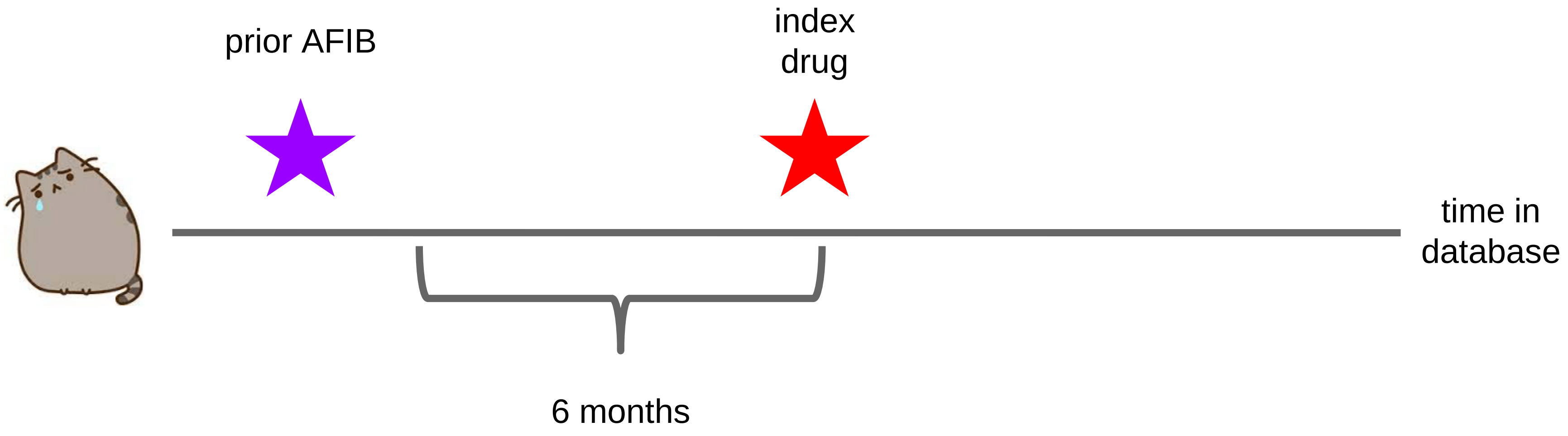


```
/******************************************************************************
*       (Exercise 2) Warfarin New Users 65 or Older at Index
******************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
)
SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
        (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX,
        EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH AS AGE_AT_INDEX
FROM CTE_DRUG_INDEX i
        JOIN OBSERVATION_PERIOD op
                ON op.PERSON_ID = i.PERSON_ID
                AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        JOIN PERSON p
                ON p.PERSON_ID = i.PERSON_ID
WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
AND EXTRACT(YEAR FROM i.INDEX_DATE)-p.YEAR_OF_BIRTH >= 65
ORDER BY i.PERSON_ID
```

Ex 3

How do I define new users of Warfarin with prior Atrial Fibrillation?



Ex 3

What is Needed in the CDM?

- **OMOP Vocabulary**
  to find the concepts

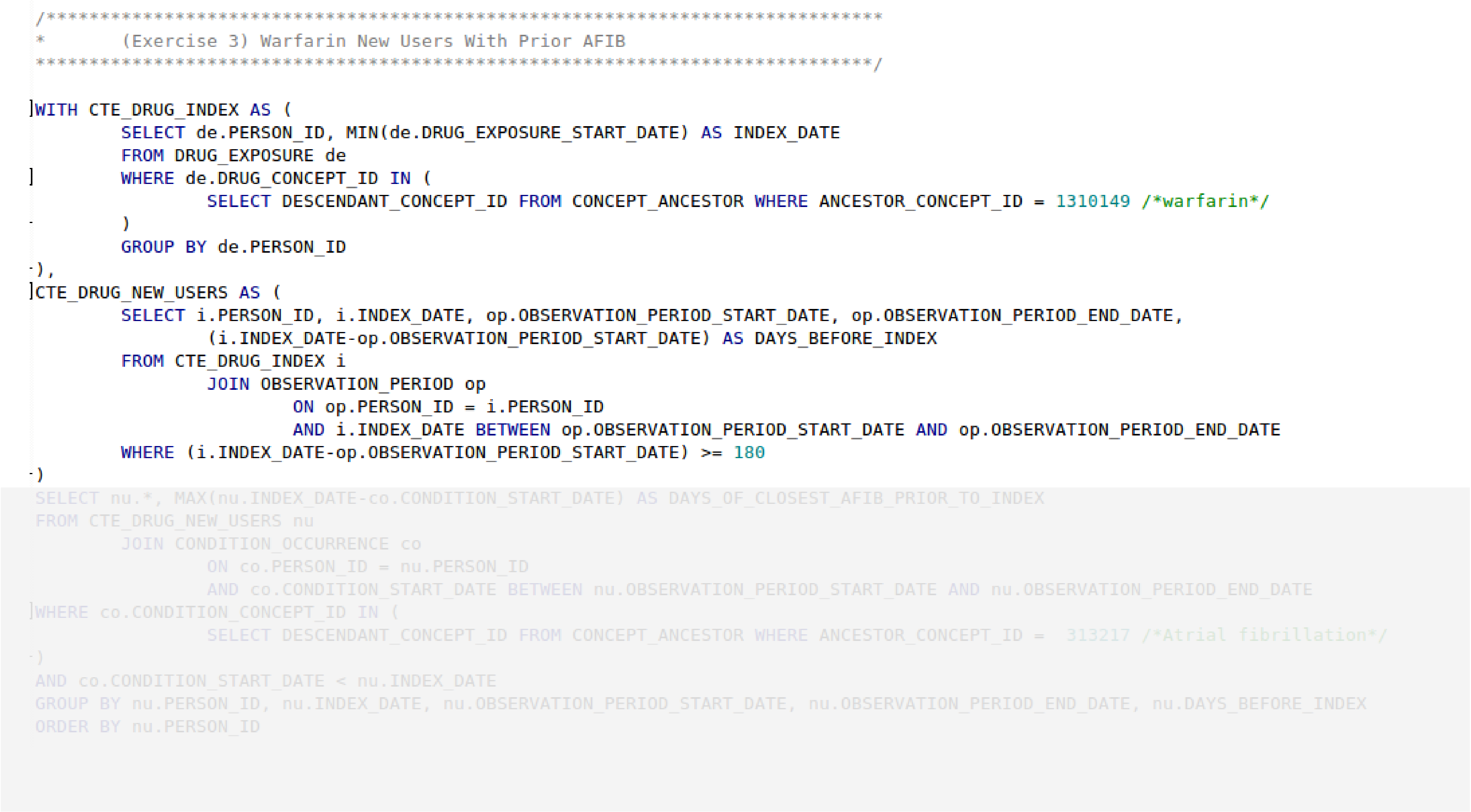
- **DRUG_EXPOSURE**
  to find individuals with exposure

- **OBSERVATION_PERIOD**
  to know people’s time within the database

- **CONDITION_OCCURRENCE**
  to find presence of a disease

Ex 3

Step 1: Start with the Ex 1 query

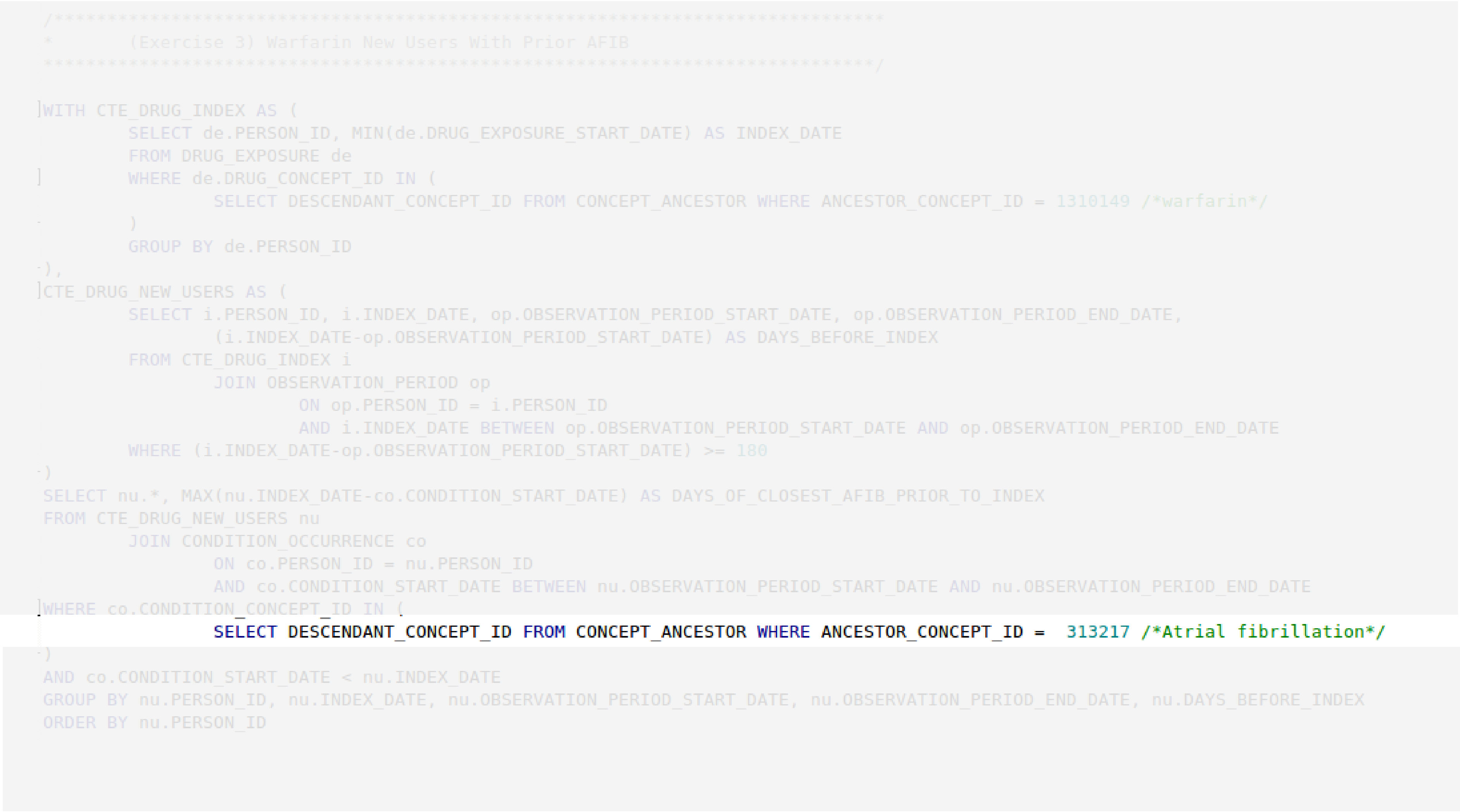


```
/*****************************************************************************
*       (Exercise 3) Warfarin New Users With Prior AFIB
*****************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
),
CTE_DRUG_NEW_USERS AS (
        SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
                (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
        FROM CTE_DRUG_INDEX i
                JOIN OBSERVATION_PERIOD op
                        ON op.PERSON_ID = i.PERSON_ID
                        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
)
SELECT nu.*, MAX(nu.INDEX_DATE-co.CONDITION_START_DATE) AS DAYS_OF_CLOSEST_AFIB_PRIOR_TO_INDEX
FROM CTE_DRUG_NEW_USERS nu
        JOIN CONDITION_OCCURRENCE co
                ON co.PERSON_ID = nu.PERSON_ID
                AND co.CONDITION_START_DATE BETWEEN nu.OBSERVATION_PERIOD_START_DATE AND nu.OBSERVATION_PERIOD_END_DATE
WHERE co.CONDITION_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID =  313217 /*Atrial fibrillation*/
)
AND co.CONDITION_START_DATE < nu.INDEX_DATE
GROUP BY nu.PERSON_ID, nu.INDEX_DATE, nu.OBSERVATION_PERIOD_START_DATE, nu.OBSERVATION_PERIOD_END_DATE, nu.DAYS_BEFORE_INDEX
ORDER BY nu.PERSON_ID
```

Ex 3

Step 2: Define Atrial Fibrillation

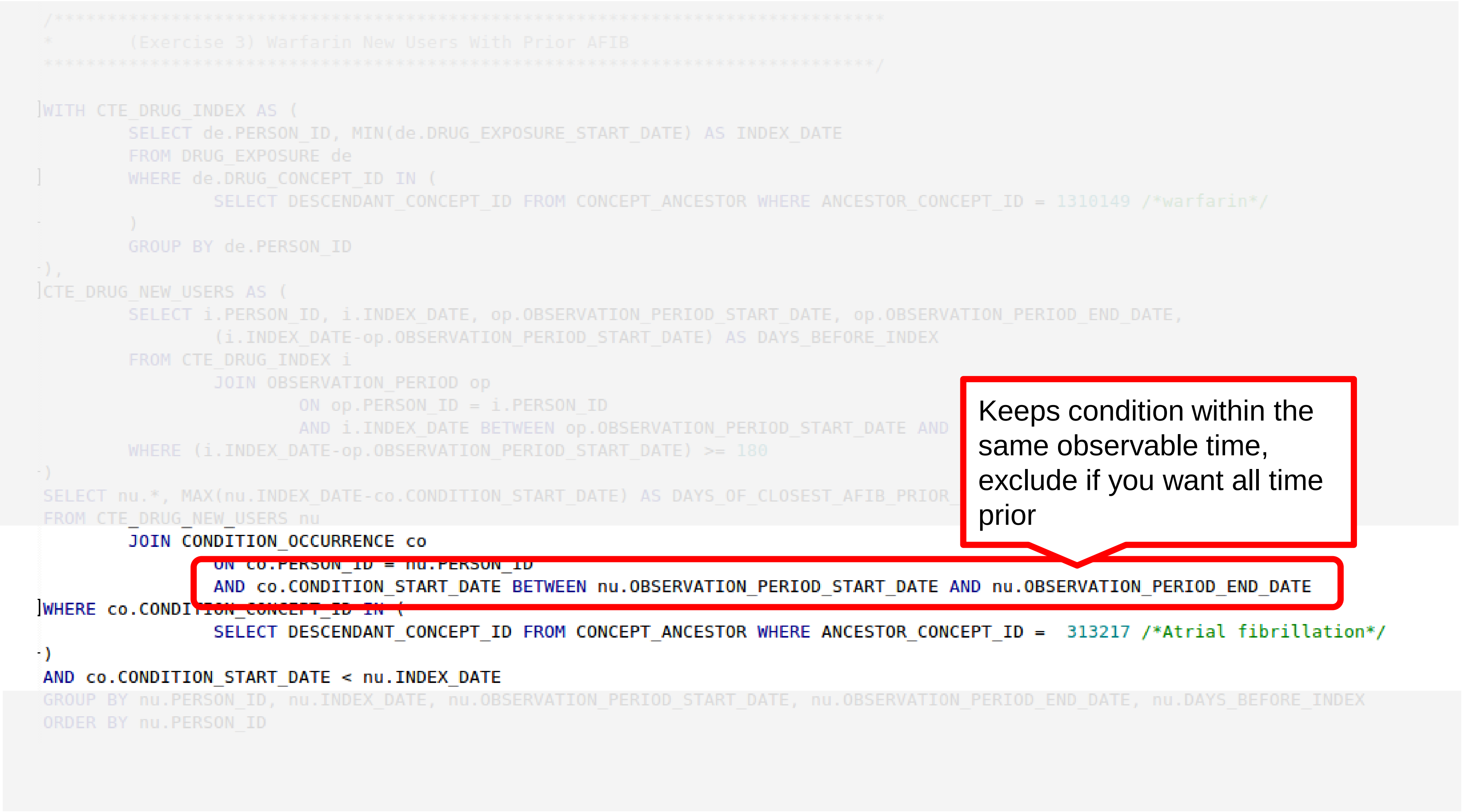


```sql
/*****************************************************************************
*       (Exercise 3) Warfarin New Users With Prior AFIB
*****************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
),
CTE_DRUG_NEW_USERS AS (
        SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
                (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
        FROM CTE_DRUG_INDEX i
                JOIN OBSERVATION_PERIOD op
                        ON op.PERSON_ID = i.PERSON_ID
                        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
)
SELECT nu.*, MAX(nu.INDEX_DATE-co.CONDITION_START_DATE) AS DAYS_OF_CLOSEST_AFIB_PRIOR_TO_INDEX
FROM CTE_DRUG_NEW_USERS nu
        JOIN CONDITION_OCCURRENCE co
                ON co.PERSON_ID = nu.PERSON_ID
                AND co.CONDITION_START_DATE BETWEEN nu.OBSERVATION_PERIOD_START_DATE AND nu.OBSERVATION_PERIOD_END_DATE
WHERE co.CONDITION_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID =  313217 /*Atrial fibrillation*/
)
AND co.CONDITION_START_DATE < nu.INDEX_DATE
GROUP BY nu.PERSON_ID, nu.INDEX_DATE, nu.OBSERVATION_PERIOD_START_DATE, nu.OBSERVATION_PERIOD_END_DATE, nu.DAYS_BEFORE_INDEX
ORDER BY nu.PERSON_ID
```

Ex 3

Step 3: Prior Atrial Fibrillation



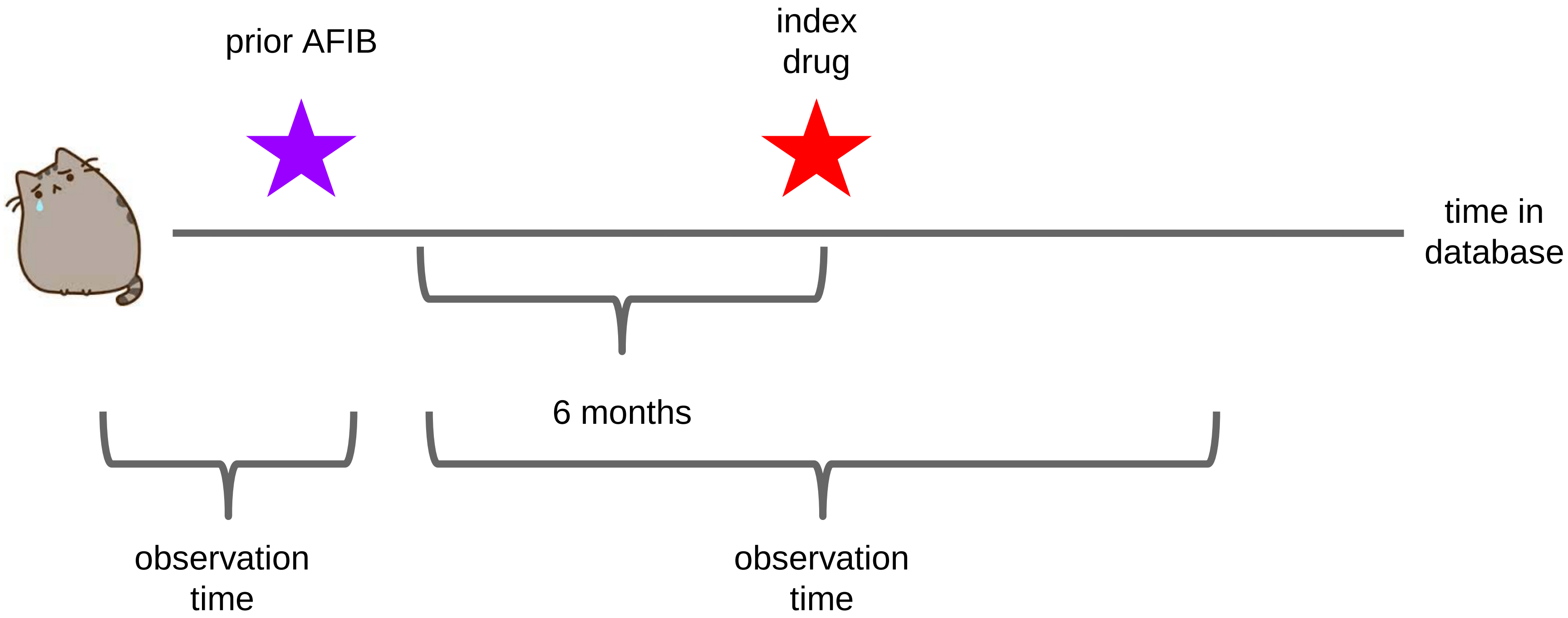
```
/***********************************************************************
*       (Exercise 3) Warfarin New Users With Prior AFIB
***********************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
),
CTE_DRUG_NEW_USERS AS (
        SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
                (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
        FROM CTE_DRUG_INDEX i
                JOIN OBSERVATION_PERIOD op
                        ON op.PERSON_ID = i.PERSON_ID
                        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND
        WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
)
SELECT nu.*, MAX(nu.INDEX_DATE-co.CONDITION_START_DATE) AS DAYS_OF_CLOSEST_AFIB_PRIOR_
FROM CTE_DRUG_NEW_USERS nu
        JOIN CONDITION_OCCURRENCE co
                ON co.PERSON_ID = nu.PERSON_ID
                AND co.CONDITION_START_DATE BETWEEN nu.OBSERVATION_PERIOD_START_DATE AND nu.OBSERVATION_PERIOD_END_DATE
WHERE co.CONDITION_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID =  313217 /*Atrial fibrillation*/
)
AND co.CONDITION_START_DATE < nu.INDEX_DATE
GROUP BY nu.PERSON_ID, nu.INDEX_DATE, nu.OBSERVATION_PERIOD_START_DATE, nu.OBSERVATION_PERIOD_END_DATE, nu.DAYS_BEFORE_INDEX
ORDER BY nu.PERSON_ID
```

Keeps condition within the same observable time, exclude if you want all time prior

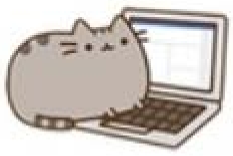
Ex 3

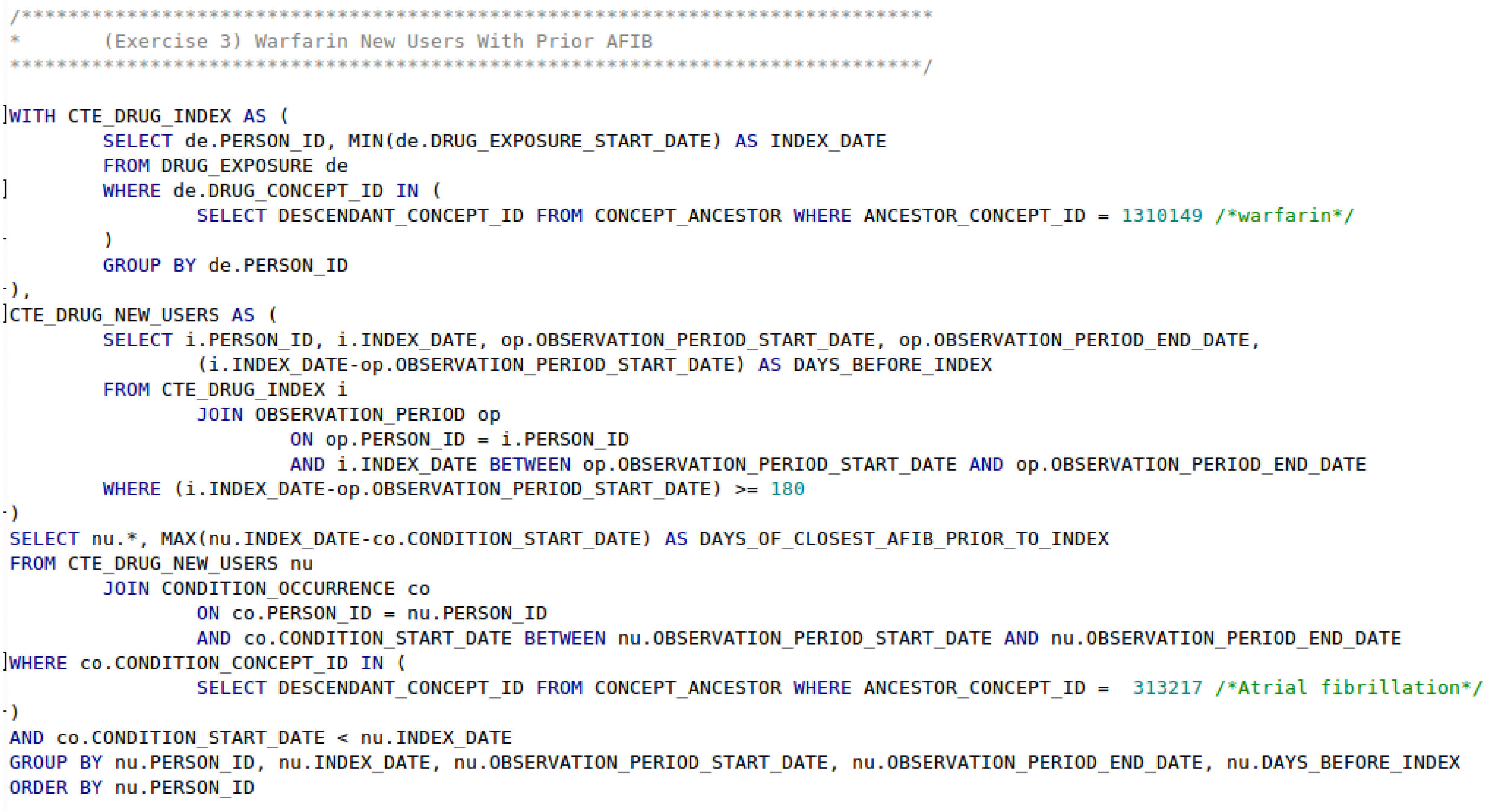
How do I define new users of Warfarin with prior Atrial Fibrillation?



Ex 3

New Users of Warfarin with prior Atrial Fibrillation





```
/******************************************************************************
*       (Exercise 3) Warfarin New Users With Prior AFIB
******************************************************************************/

WITH CTE_DRUG_INDEX AS (
        SELECT de.PERSON_ID, MIN(de.DRUG_EXPOSURE_START_DATE) AS INDEX_DATE
        FROM DRUG_EXPOSURE de
        WHERE de.DRUG_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID = 1310149 /*warfarin*/
        )
        GROUP BY de.PERSON_ID
),
CTE_DRUG_NEW_USERS AS (
        SELECT i.PERSON_ID, i.INDEX_DATE, op.OBSERVATION_PERIOD_START_DATE, op.OBSERVATION_PERIOD_END_DATE,
                (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) AS DAYS_BEFORE_INDEX
        FROM CTE_DRUG_INDEX i
                JOIN OBSERVATION_PERIOD op
                        ON op.PERSON_ID = i.PERSON_ID
                        AND i.INDEX_DATE BETWEEN op.OBSERVATION_PERIOD_START_DATE AND op.OBSERVATION_PERIOD_END_DATE
        WHERE (i.INDEX_DATE-op.OBSERVATION_PERIOD_START_DATE) >= 180
)
SELECT nu.*, MAX(nu.INDEX_DATE-co.CONDITION_START_DATE) AS DAYS_OF_CLOSEST_AFIB_PRIOR_TO_INDEX
FROM CTE_DRUG_NEW_USERS nu
        JOIN CONDITION_OCCURRENCE co
                ON co.PERSON_ID = nu.PERSON_ID
                AND co.CONDITION_START_DATE BETWEEN nu.OBSERVATION_PERIOD_START_DATE AND nu.OBSERVATION_PERIOD_END_DATE
WHERE co.CONDITION_CONCEPT_ID IN (
                SELECT DESCENDANT_CONCEPT_ID FROM CONCEPT_ANCESTOR WHERE ANCESTOR_CONCEPT_ID =  313217 /*Atrial fibrillation*/
)
AND co.CONDITION_START_DATE < nu.INDEX_DATE
GROUP BY nu.PERSON_ID, nu.INDEX_DATE, nu.OBSERVATION_PERIOD_START_DATE, nu.OBSERVATION_PERIOD_END_DATE, nu.DAYS_BEFORE_INDEX
ORDER BY nu.PERSON_ID
```

Try on your own!

- Warfarin New Users 65 or Older at Index with Prior Atrial Fibrillation

8,207 individuals

- Bonus: Clipidogrel New Users 65 or Older at Index with Prior Atrial Fibrillation

3,148 individuals

Queries Can Be Automated

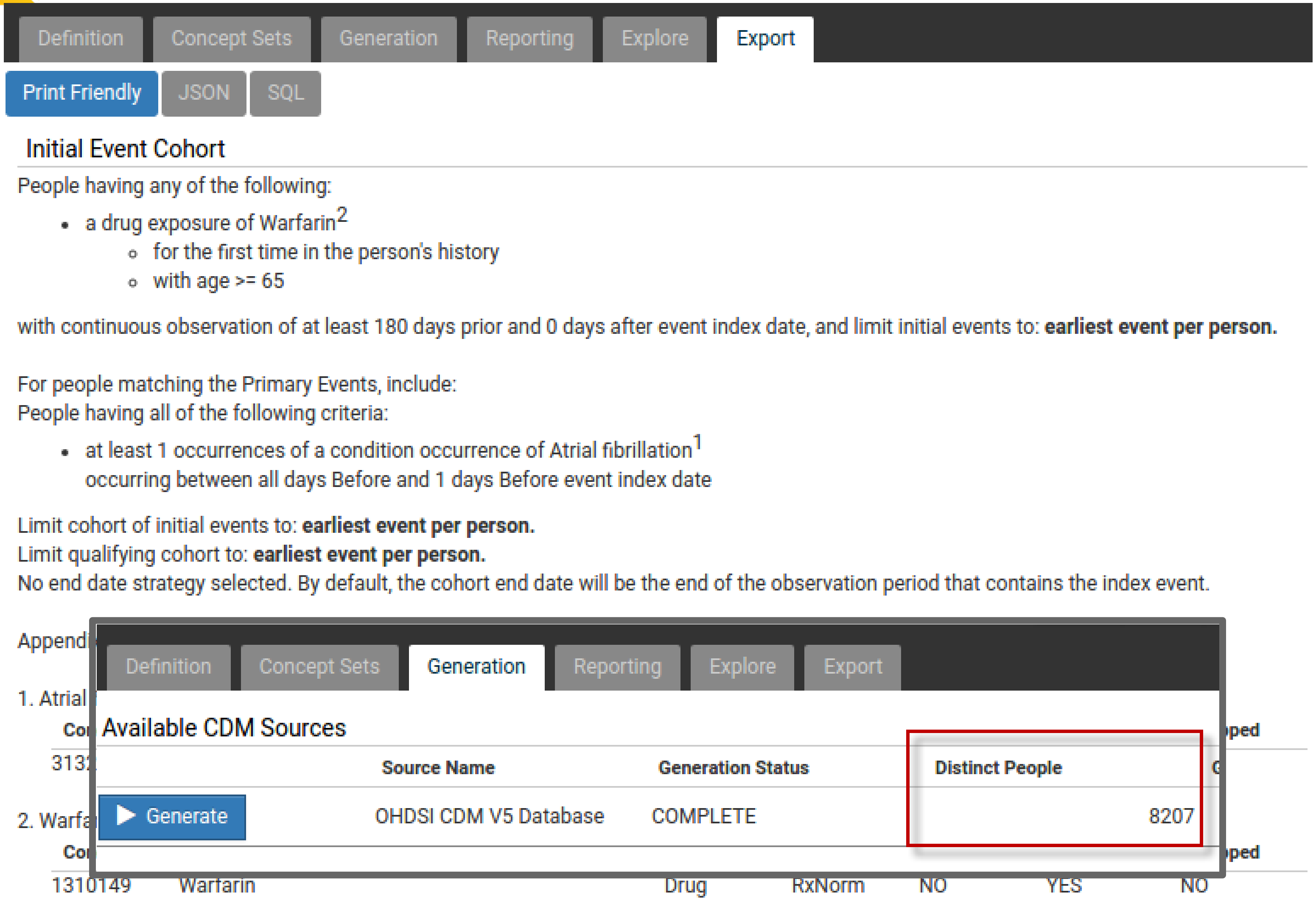
- Open up Google Chrome

- Navigate to:
  http://localhost:8080/atlas/#/home

- In Atlas navigate to Cohorts

- There should be a pre-existing cohort called "Warfarin New Users 65 or Older at Index with Prior Atrial Fibrillation."

Queries Can Be Automated



Initial Event Cohort

People having any of the following:

- a drug exposure of Warfarin[2]
  - for the first time in the person's history
  - with age >= 65

with continuous observation of at least 180 days prior and 0 days after event index date, and limit initial events to: **earliest event per person.**

For people matching the Primary Events, include:
People having all of the following criteria:

- at least 1 occurrences of a condition occurrence of Atrial fibrillation[1]
  occurring between all days Before and 1 days Before event index date

Limit cohort of initial events to: **earliest event per person.**
Limit qualifying cohort to: **earliest event per person.**
No end date strategy selected. By default, the cohort end date will be the end of the observation period that contains the index event.

Append

1. Atrial

Co | | pped
313 | |

2. Warfa

Co | | pped

1310149 Warfarin Drug RxNorm NO YES NO

Available CDM Sources

	Source Name	Generation Status	Distinct People
▶ Generate	OHDSI CDM V5 Database	COMPLETE	8207

Break

Please return in 15 minutes

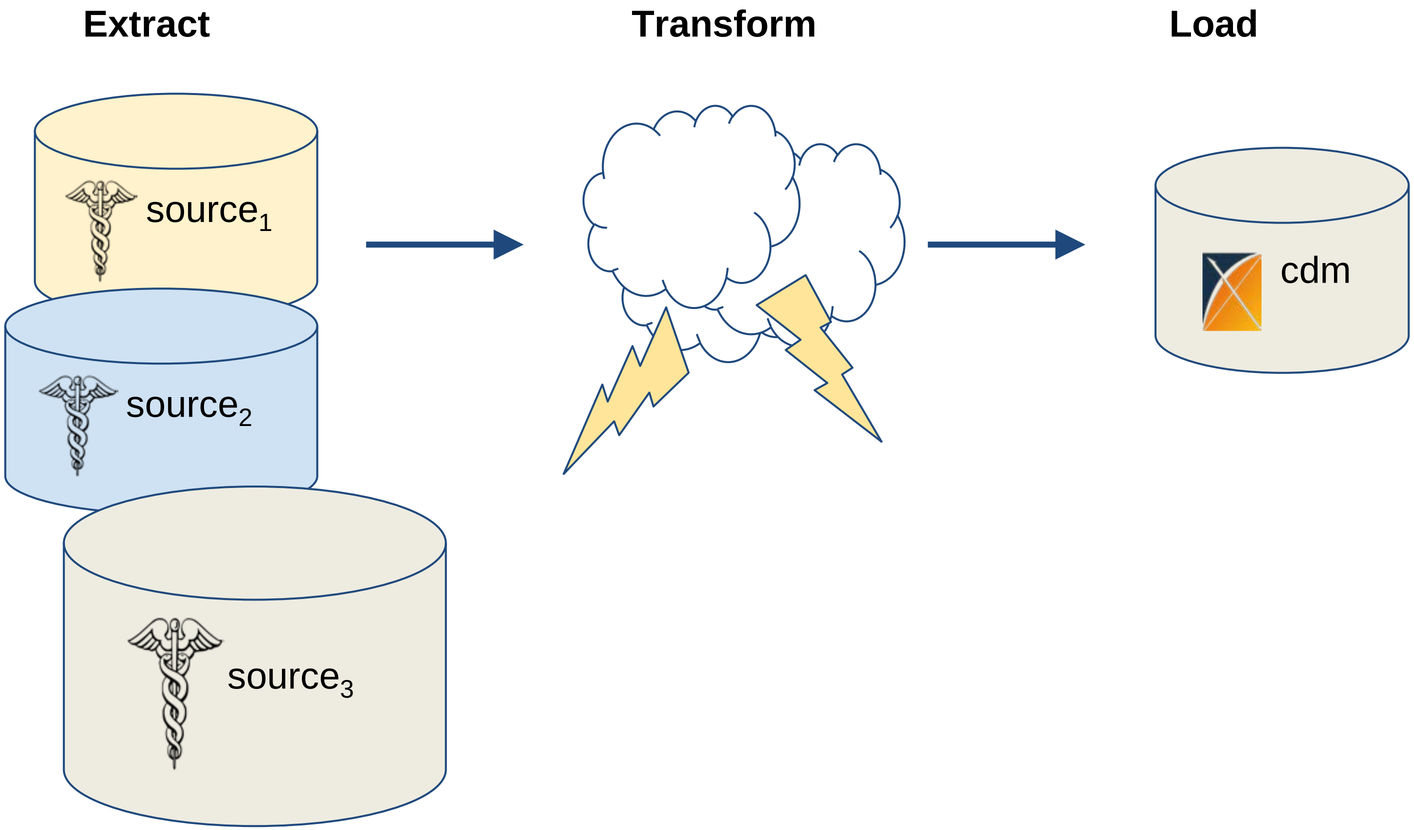
Implementation

Setup and Performing of an Extract Transform and Load process into the CDM

Brief Review

- Foundational
  - OHDSI - Why and how
  - OMOP CDM - Standardizing structure & content
  - Real-world examples (SQL and Atlas)

How do we create our own OMOP CDM instance?



ETL: Real world scenario

Truven MarketScan Commercial Claims and Encounters (CCAE)

INPATIENT_SERVICES

enrolid	admdate	pdx	dx1	dx2	dx3
1570337021	5/31/2000	41071	41071	4241	V5881

Optum Extended SES

MEDICAL_CLAIMS

patid	fst_dt	diag1	diag2	diag3	diag4
259000476532	5/30/2000	41071	27800	4019	2724

Premier

PATICD_DIAG

pat_key	period	icd_code	icd_p
-17197140	1/1/2000	410.71	P
-17197140	1/1/2000	414.01	S
-17197140	1/1/2000	427.31	S
-17197140	1/1/2000	496	S

Japan Medical Data Center

DIAGNOSIS

member_id	admission_date	icd10_level4_code
M0041437	4/11/2013	I214
M0041437	4/11/2013	A539
M0041437	4/11/2013	B182
M0041437	4/11/2013	E14-

4 real observational databases, all containing an inpatient admission for a patient with a diagnosis of 'acute subendocardial infarction'

- Not a single table name the same…
- Not a single variable name the same….
- Different table structures (rows vs. columns)
- Different conventions (with and without decimal points)
- Different coding schemes (ICD9 vs. ICD10)

What does it mean to ETL to OMOP CDM?
Standardize **structure** and **content**

Truven MarketScan Commerical Claims and Encounters (CCAE)

INPATIENT_SERVICES

enrolid	admdate	pdx	dx1	dx2	dx3
1570337021	5/31/2000	41071	41071	4241	V5881

Structure optimized for large-scale analysis for clinical characterization, population-level estimation, and patient-level prediction

Truven MarketScan Commerical Claims and Encounters (CCAE)

CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
157033702	5/31/2000	41071	Inpatient claims - primary position
157033702	5/31/2000	41071	Inpatient claims - 1st position
157033702	5/31/2000	4241	Inpatient claims - 2nd position
157033702	5/31/2000	V5881	Inpatient claims - 3rd position

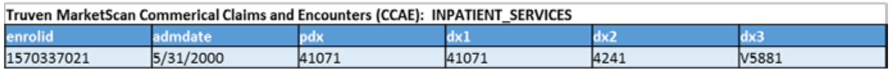
Content using international vocabulary standards that can be applied to any data source

Truven MarketScan Commerical Claims and Encounters (CCAE)

CONDITION_OCCURRENCE

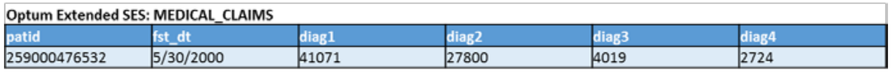
PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION _TYPE_CONCEPT_ID	CONDITION_SOURCE_CONCEPT_ID	CONDITION_CONCEPT_ID
157033702	5/31/2000	41071	Inpatient claims - primary position	44825429	444406

OMOP CDM = Standardized structure: same tables, same fields, same datatypes, same conventions across disparate sources



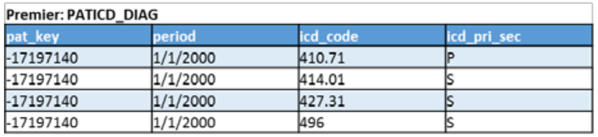
Truven MarketScan Commerical Claims and Encounters (CCAE): INPATIENT_SERVICES

enrolid	admdate	pdx	dx1	dx2	dx3
1570337021	5/31/2000	41071	41071	4241	V5881



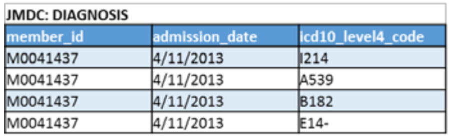
Optum Extended SES: MEDICAL_CLAIMS

patid	fst_dt	diag1	diag2	diag3	diag4
259000476532	5/30/2000	41071	27800	4019	2724



Premier: PATICD_DIAG

pat_key	period	icd_code	icd_pri_sec
-17197140	1/1/2000	410.71	P
-17197140	1/1/2000	414.01	S
-17197140	1/1/2000	427.31	S
-17197140	1/1/2000	496	S



JMDC: DIAGNOSIS

member_id	admission_date	icd10_level4_code
M0041437	4/11/2013	I214
M0041437	4/11/2013	A539
M0041437	4/11/2013	B182
M0041437	4/11/2013	E14-

- Consistent structure optimized for large-scale analysis
- Structure preserves all source content and provenance

Truven CCAE: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
157033702	5/31/2000	41071	Inpatient claims - primary position
157033702	5/31/2000	41071	Inpatient claims - 1st position
157033702	5/31/2000	4241	Inpatient claims - 2nd position
157033702	5/31/2000	V5881	Inpatient claims - 3rd position

Optum Extended SES: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
259000474406532	5/30/2000	41071	Inpatient claims - 1st position
259000474406532	5/30/2000	27800	Inpatient claims - 2nd position
259000474406532	5/30/2000	4019	Inpatient claims - 3rd position
259000474406532	5/30/2000	2724	Inpatient claims - 4th position

Premier : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
-171971409	1/1/2000	410.71	Hospital record - primary
-171971409	1/1/2000	414.01	Hospital record - secondary
-171971409	1/1/2000	427.31	Hospital record - secondary
-171971409	1/1/2000	496	Hospital record - secondary

JMDC : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
4149337	4/11/2013	I214	Inpatient claims
4149337	4/11/2013	A539	Inpatient claims
4149337	4/11/2013	B182	Inpatient claims
4149337	4/11/2013	E14-	Inpatient claims

OMOP CDM = Standardized content: common vocabularies across disparate sources

Truven CCAE: CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
157033702	5/31/2000	41071	Inpatient claims - primary position	44825429	444406

Optum Extended SES: CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
259000474406532	5/30/2000	41071	Inpatient claims - 1st position	44825429	444406

Premier : CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
-171971409	1/1/2000	410.71	Hospital record - primary	44825429	444406

JMDC : CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
4149337	4/11/2013	I214	Inpatient claims	45572081	444406

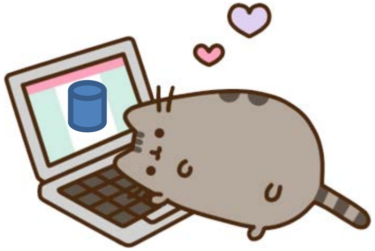
- Standardize source codes to be uniquely defined across all vocabularies
- No more worries about formatting or code overlap

- Standardize across vocabularies to a common referent standard (ICD9/10→SNOMED)
- Source codes mapped into each domain standard so that now you can talk across different languages

ETL Process: Roles

- Members of the team

  - CDM expert



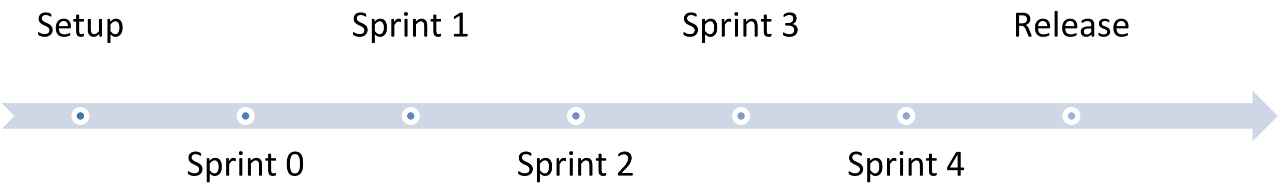
  - Local data expert

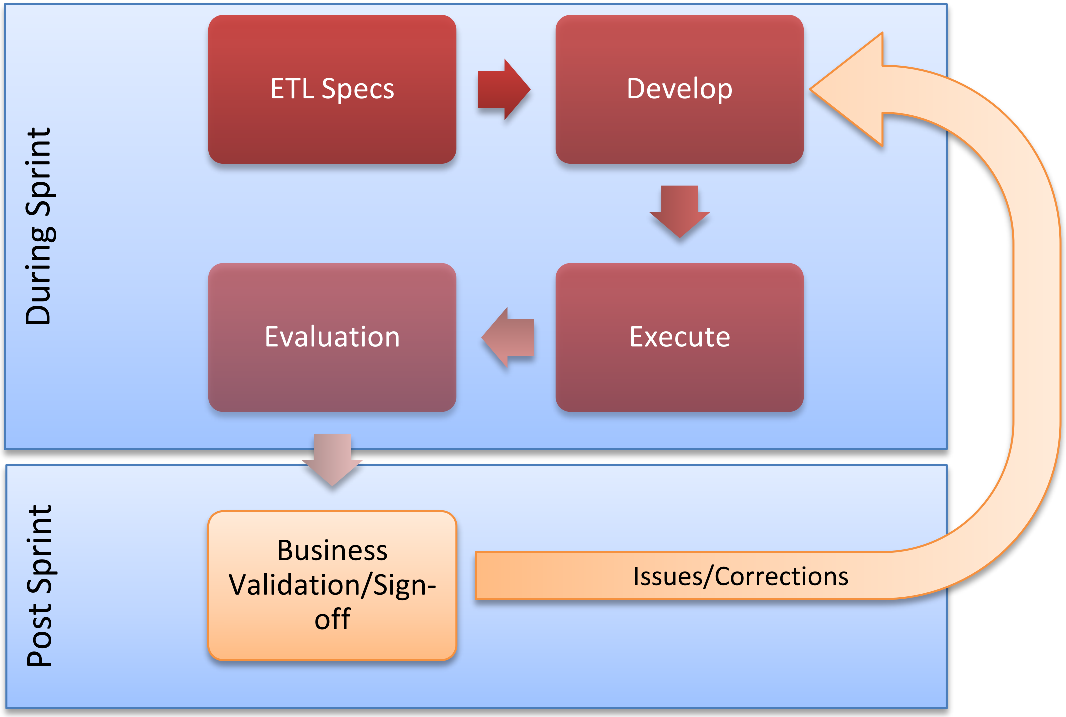
  - Data engineer

  - Person with medical knowledge

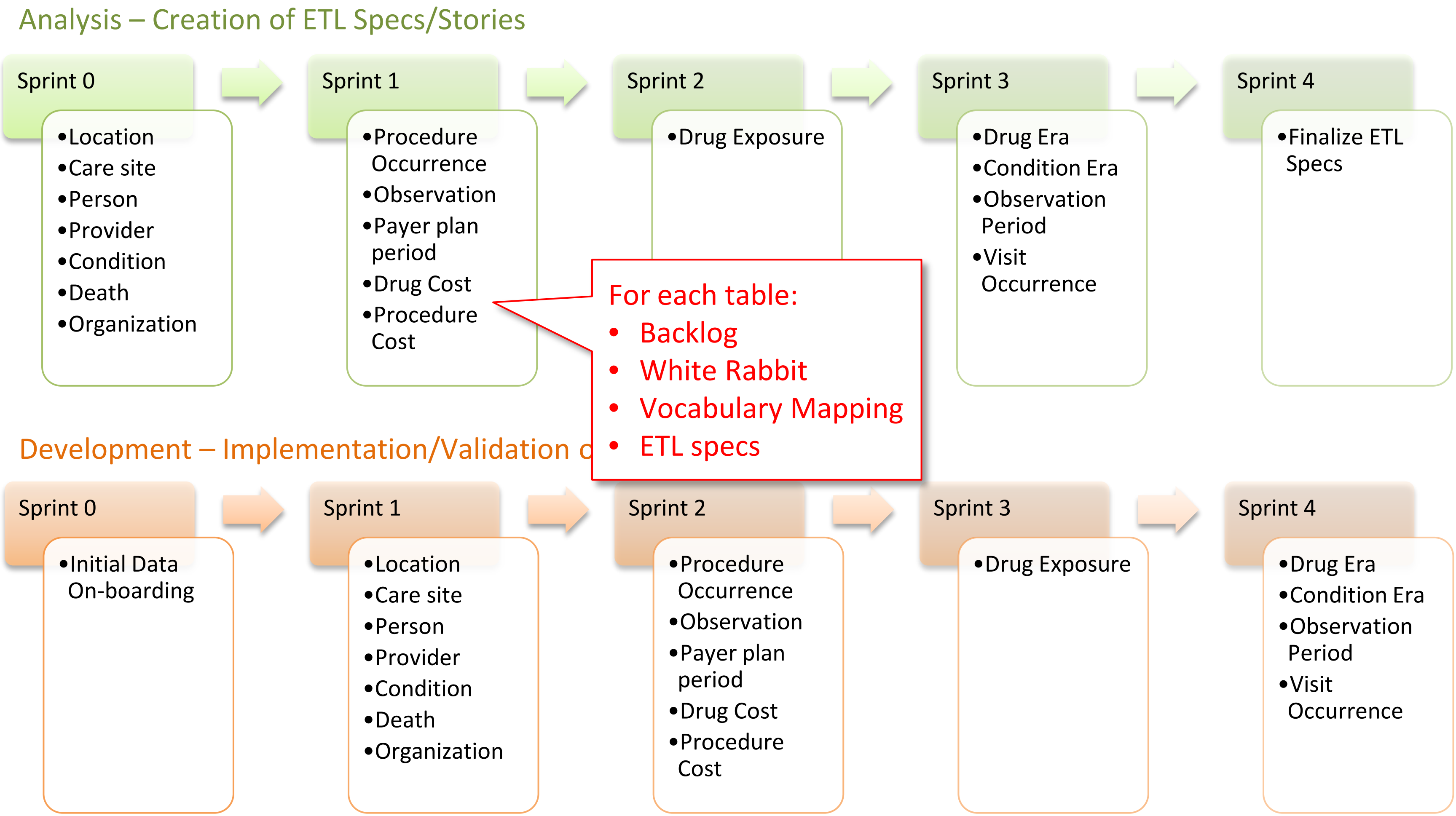
  - Business stakeholder

ETL Process: Agile





Example OHDSI ETL Process



Analysis – Creation of ETL Specs/Stories

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

Sprint 1
- Procedure Occurrence
- Observation
- Payer plan period
- Drug Cost
- Procedure Cost

Sprint 2
- Drug Exposure

Sprint 3
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence

Sprint 4
- Finalize ETL Specs

For each table:
- Backlog
- White Rabbit
- Vocabulary Mapping
- ETL specs

Development – Implementation/Validation o

Sprint 0
- Initial Data On-boarding

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

Sprint 2
- Procedure Occurrence
- Observation
- Payer plan period
- Drug Cost
- Procedure Cost

Sprint 3
- Drug Exposure

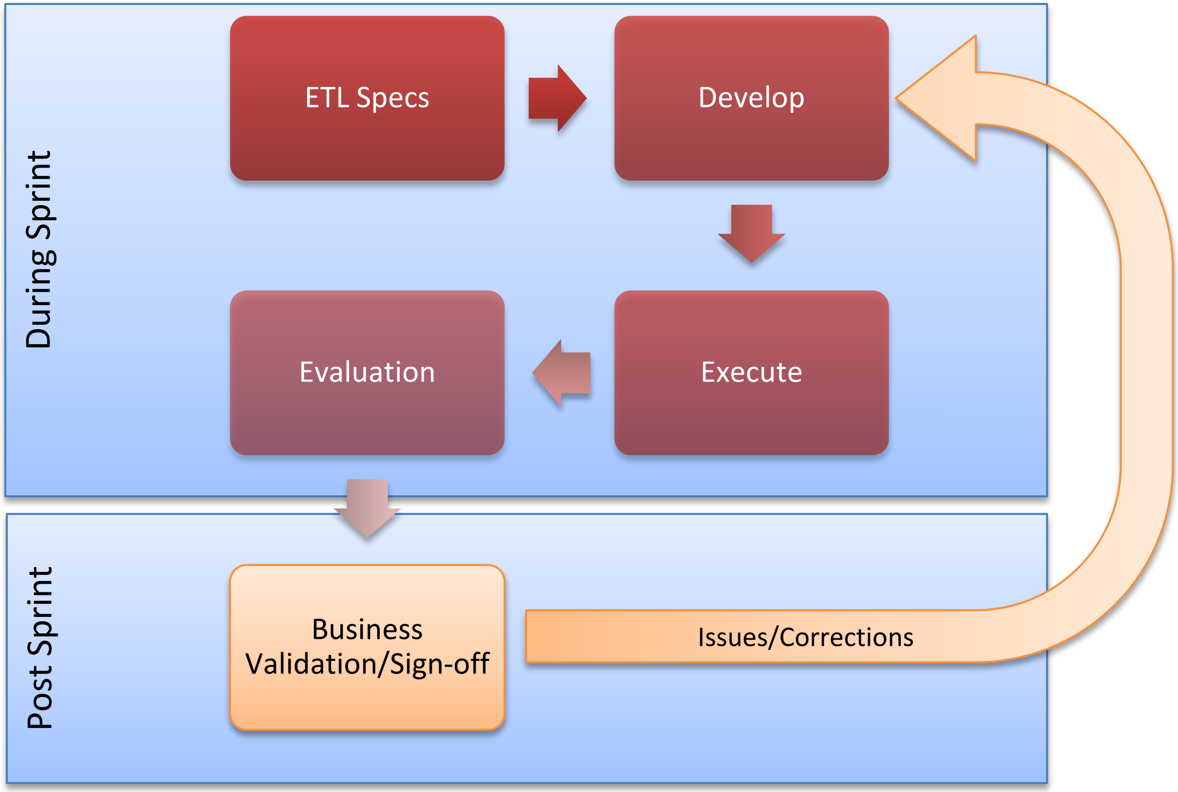
Sprint 4
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence

OHDSI Resources for ETL

Rabbits

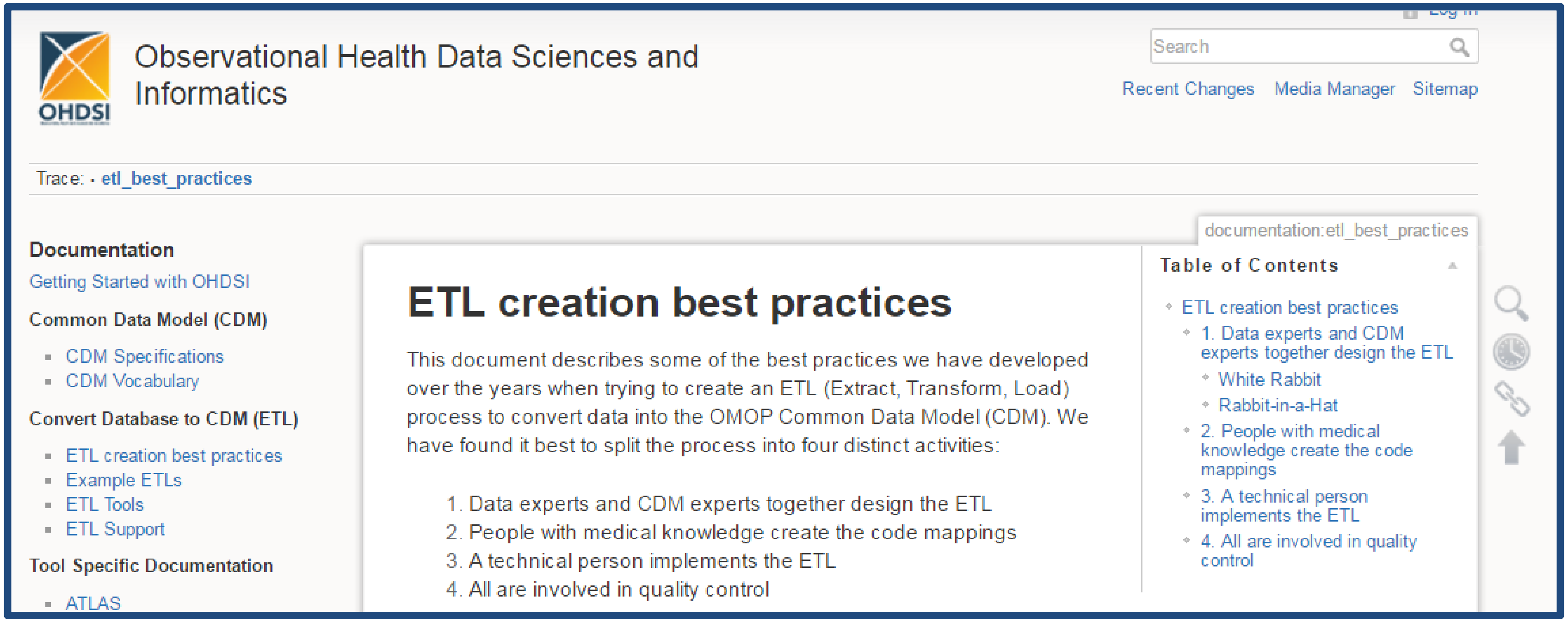
Usagi

Atlas Data Sources (Achilles)



Best Practices Documented

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



Log In
OHDSI
Observational Health Data Sciences and Informatics
Search
Recent Changes
Media Manager
Sitemap
Trace: · etl_best_practices
documentation: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

ETL creation 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

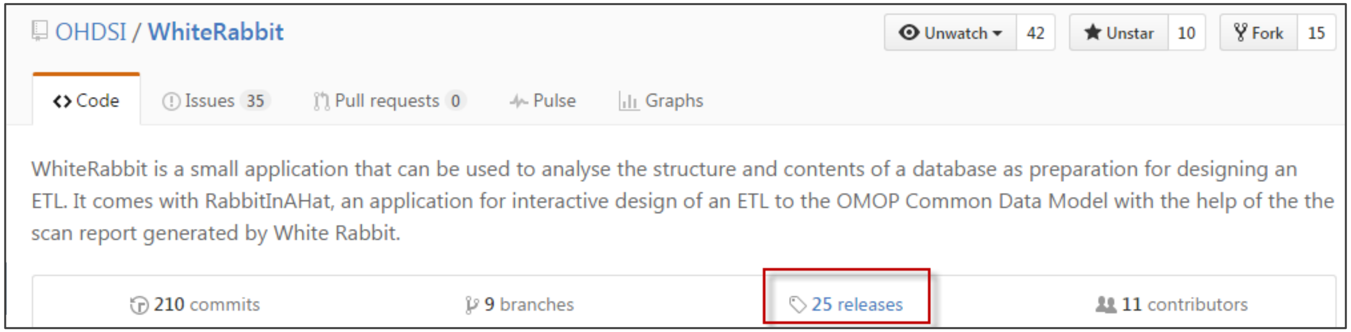
Table of Contents

Implementation

Using WhiteRabbit and Rabbit-In-A-Hat to Build an ETL

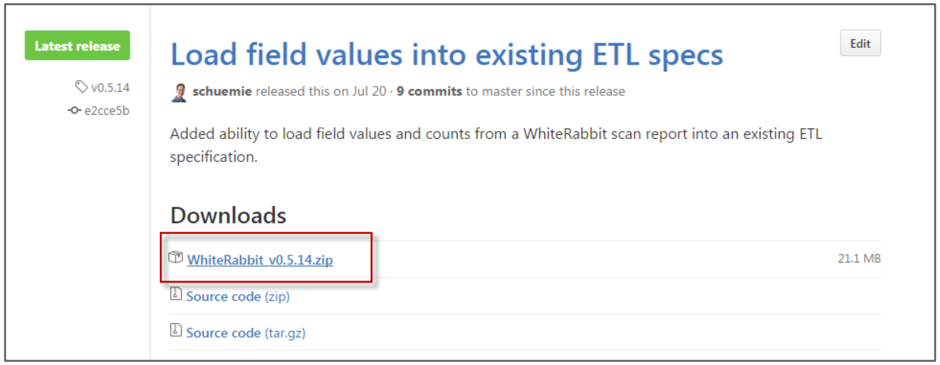
Getting WhiteRabbit

- https://github.com/OHDSI/WhiteRabbit
- Click on “releases”



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.

- “Latest Release” and download the WhiteRabbit zip file



Load field values into existing ETL specs

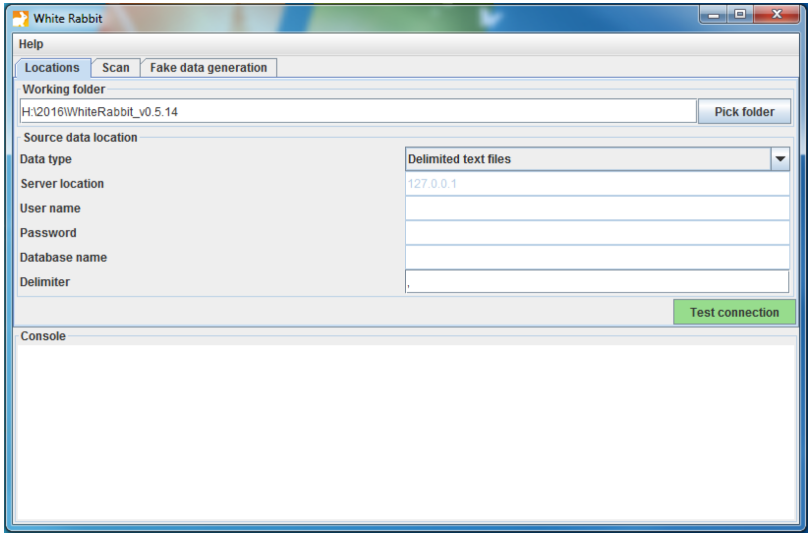
schuemie released this on Jul 20 · 9 commits to master since this release

Added ability to load field values and counts from a WhiteRabbit scan report into an existing ETL specification.

Downloads

Getting WhiteRabbit

- Save the ZIP file somewhere and extract the files
- Double-click on the WhiteRabbit.jar to start the application.



Working folder: H:\2016\WhiteRabbit_v0.5.14

Source data location

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

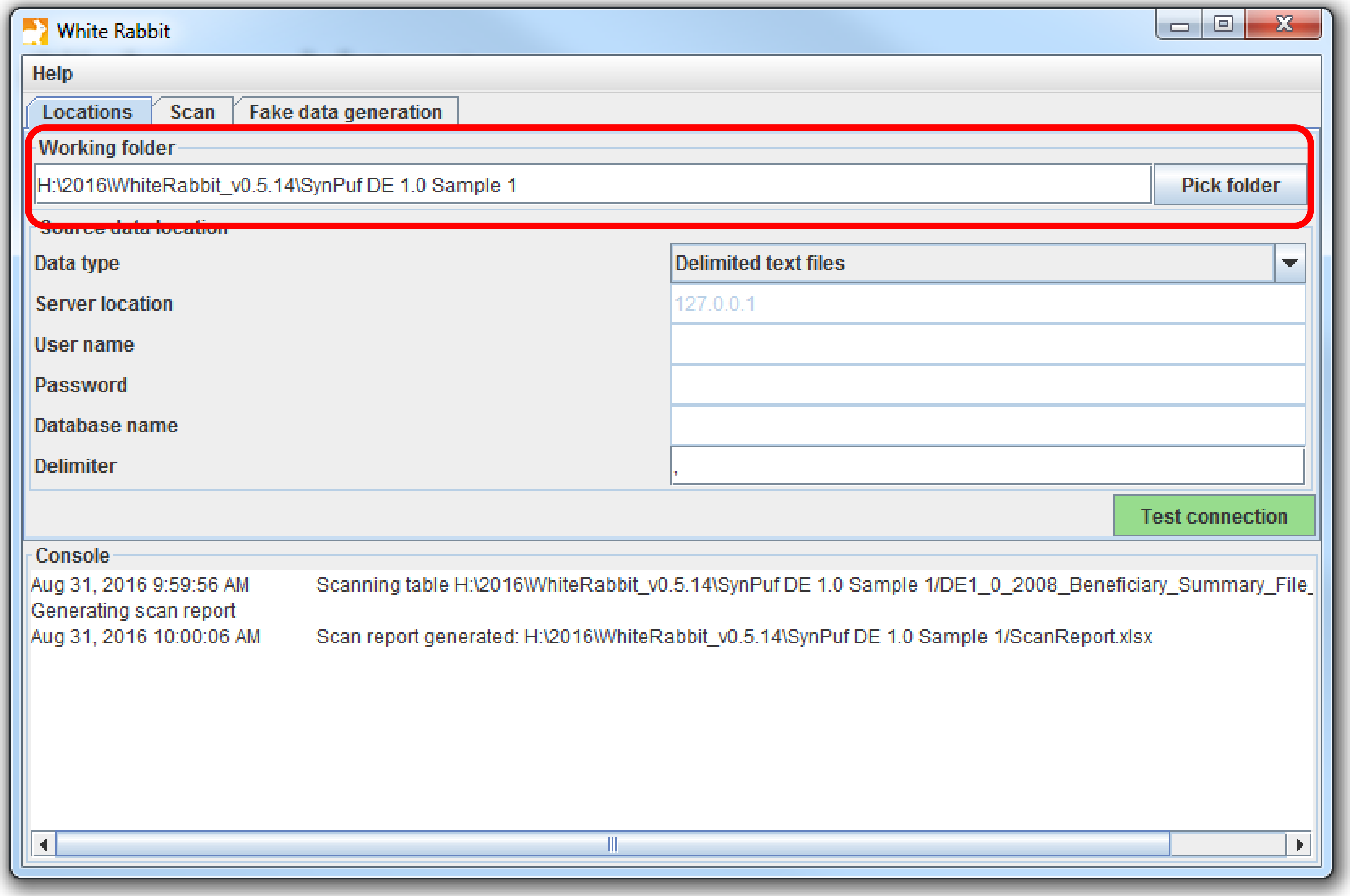
Working with WhiteRabbit

- Wiki: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:whiterabbit

1. Specify the location of your data
   - Supports database connections as well as text files

2. Scanning your Database
   - Characterizes your data

Specify the Location of Data



Working folder

H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

User name

Password

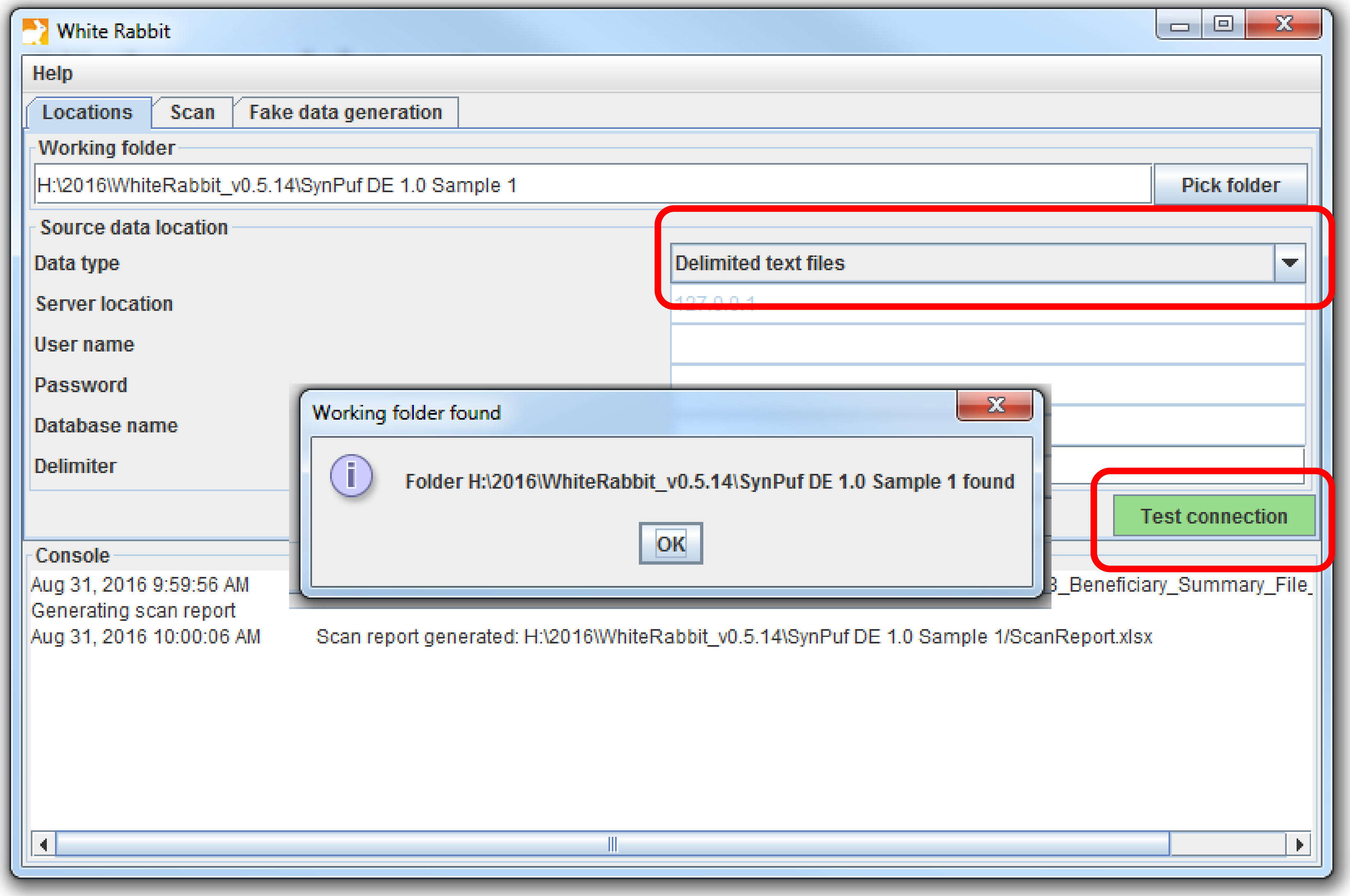
Database name

Delimiter: ,

Console

```
Aug 31, 2016 9:59:56 AM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/DE1_0_2008_Beneficiary_Summary_File_
Generating scan report
Aug 31, 2016 10:00:06 AM       Scan report generated: H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/ScanReport.xlsx
```

Specify the Location of Data



Working folder: H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

User name

Password

Database name

Delimiter

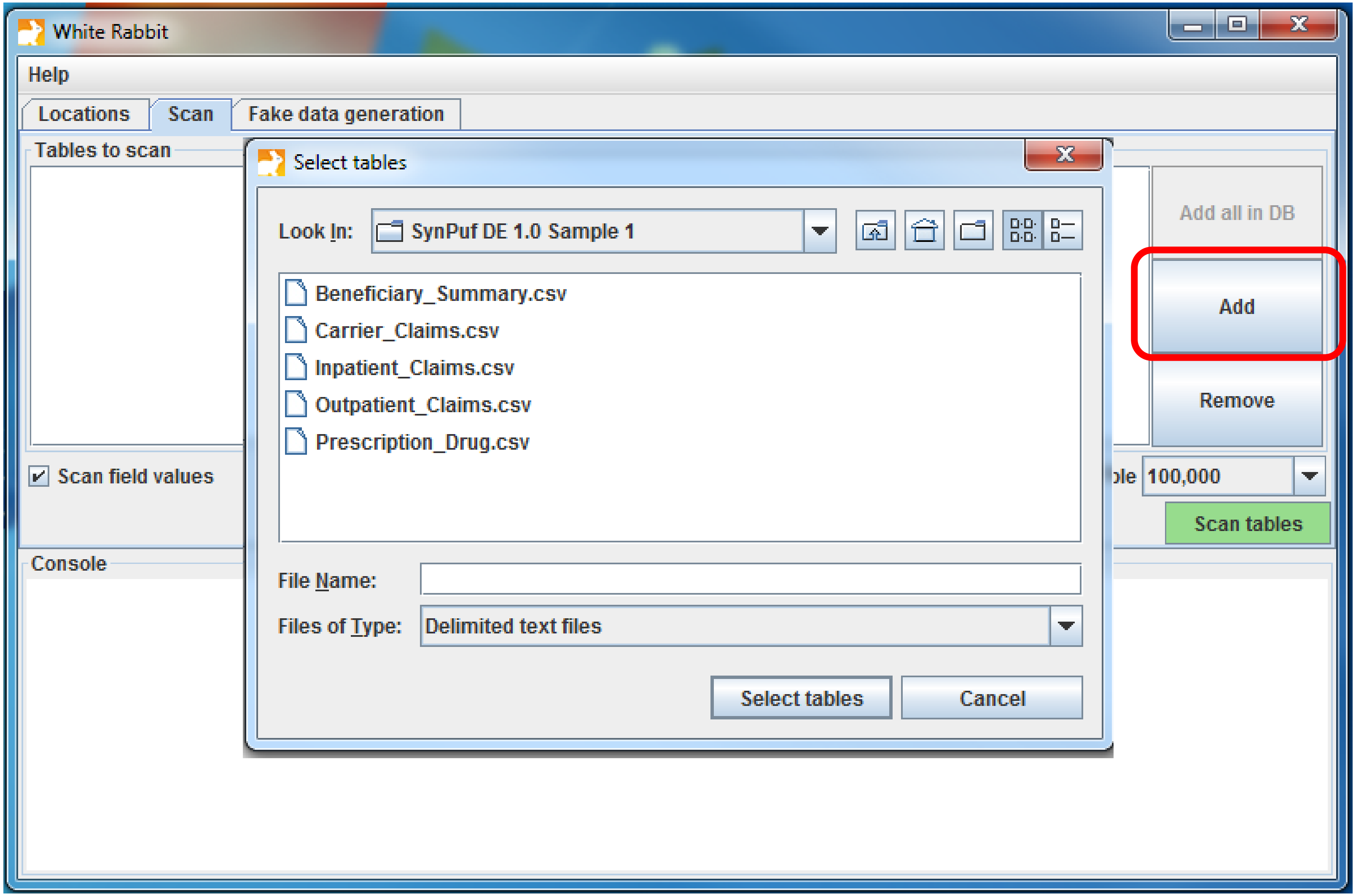
Working folder found

Folder H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1 found

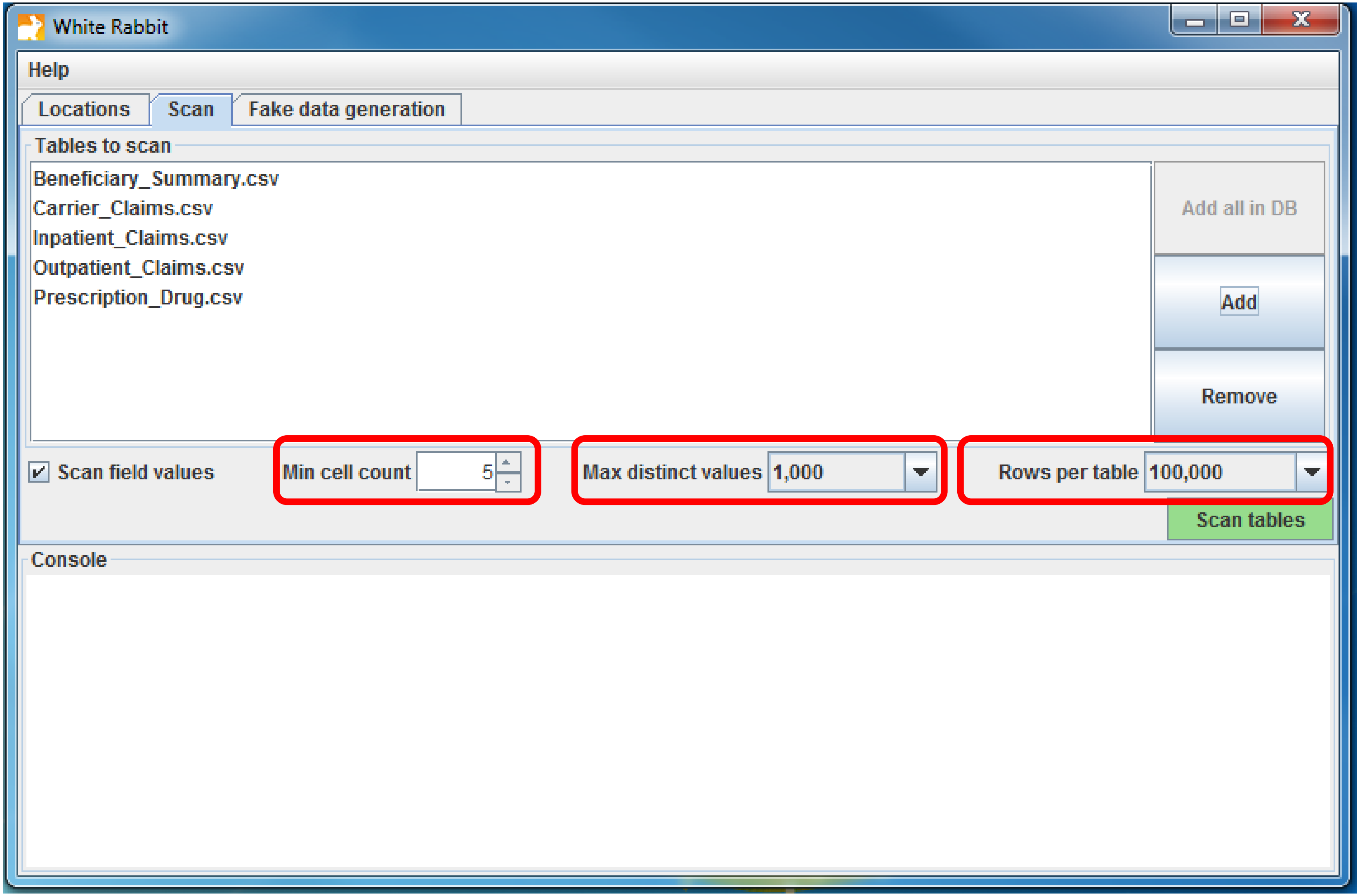
Console

```
Aug 31, 2016 9:59:56 AM                                        3_Beneficiary_Summary_File_
Generating scan report
Aug 31, 2016 10:00:06 AM      Scan report generated: H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/ScanReport.xlsx
```

Scanning your Data



Scanning your Data



White Rabbit

Tables to scan

Beneficiary_Summary.csv
Carrier_Claims.csv
Inpatient_Claims.csv
Outpatient_Claims.csv
Prescription_Drug.csv

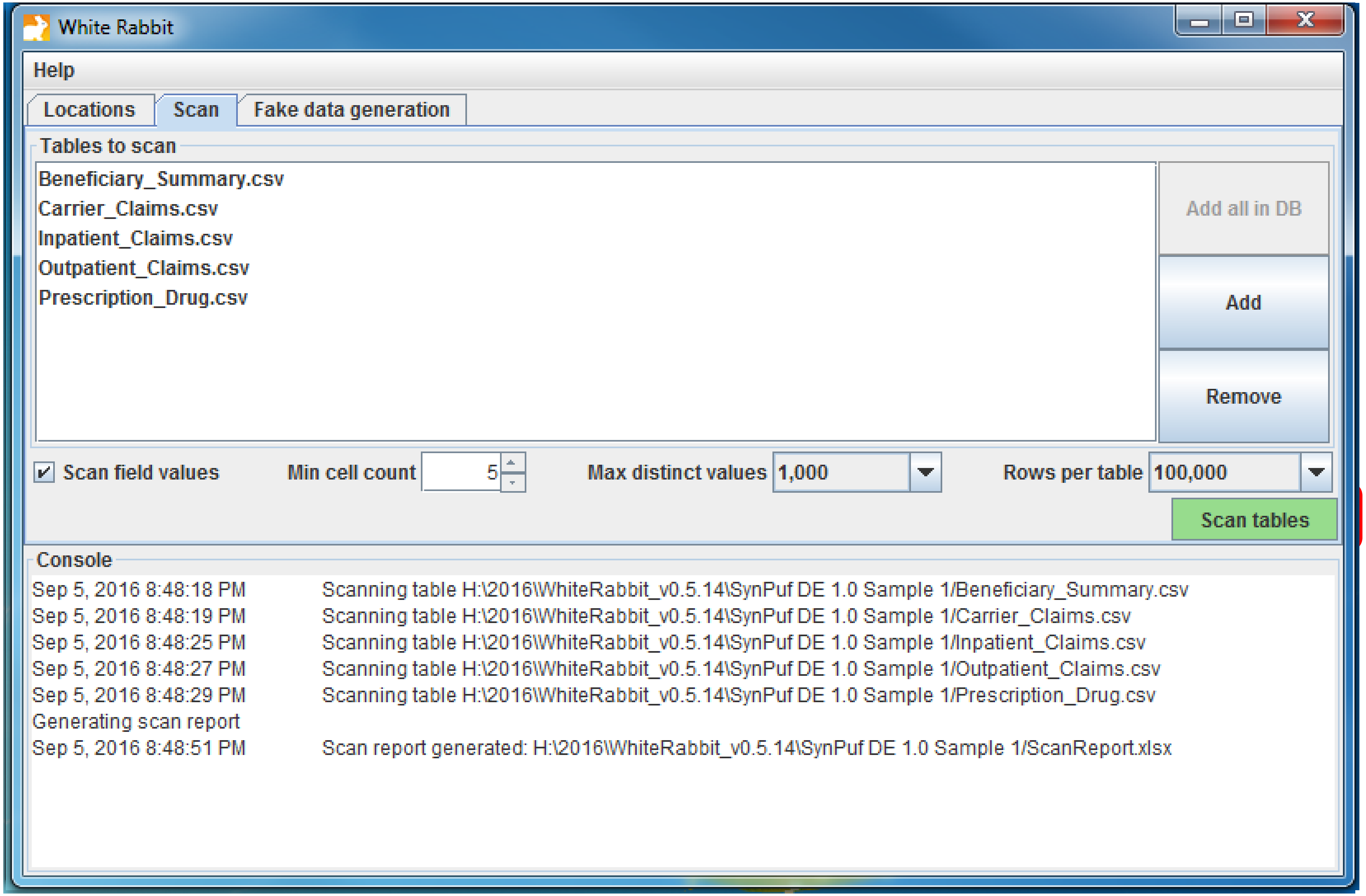
☑ Scan field values

Min cell count 5

Max distinct values 1,000

Rows per table 100,000

Scanning your Data



Tables to scan

Beneficiary_Summary.csv
Carrier_Claims.csv
Inpatient_Claims.csv
Outpatient_Claims.csv
Prescription_Drug.csv

☑ Scan field values

Min cell count 5

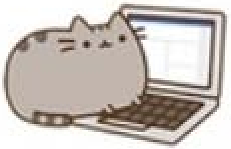
Max distinct values 1,000

Rows per table 100,000

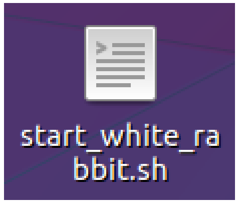
Console

```
Sep 5, 2016 8:48:18 PM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/Beneficiary_Summary.csv
Sep 5, 2016 8:48:19 PM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/Carrier_Claims.csv
Sep 5, 2016 8:48:25 PM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/Inpatient_Claims.csv
Sep 5, 2016 8:48:27 PM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/Outpatient_Claims.csv
Sep 5, 2016 8:48:29 PM        Scanning table H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/Prescription_Drug.csv
Generating scan report
Sep 5, 2016 8:48:51 PM        Scan report generated: H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1/ScanReport.xlsx
```

Run the Scan Report on Your Data!

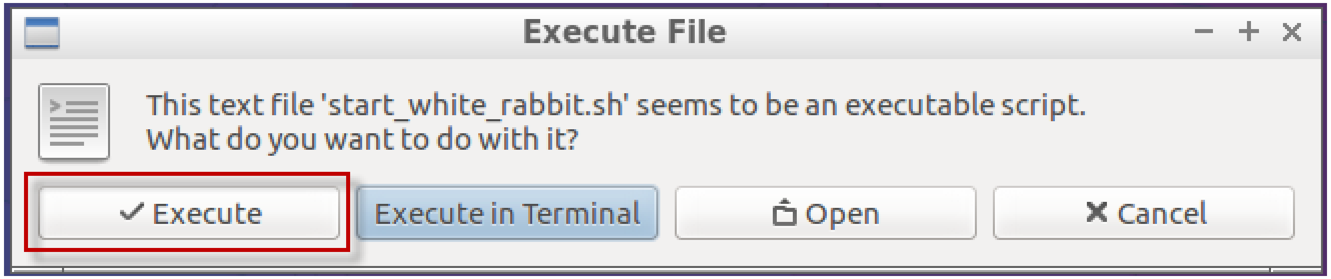


- Link on desktop



start_white_rabbit.sh

- Execute



Execute File

This text file 'start_white_rabbit.sh' seems to be an executable script.
What do you want to do with it?

- WhiteRabbit appears

Run the Scan Report on Your Data!

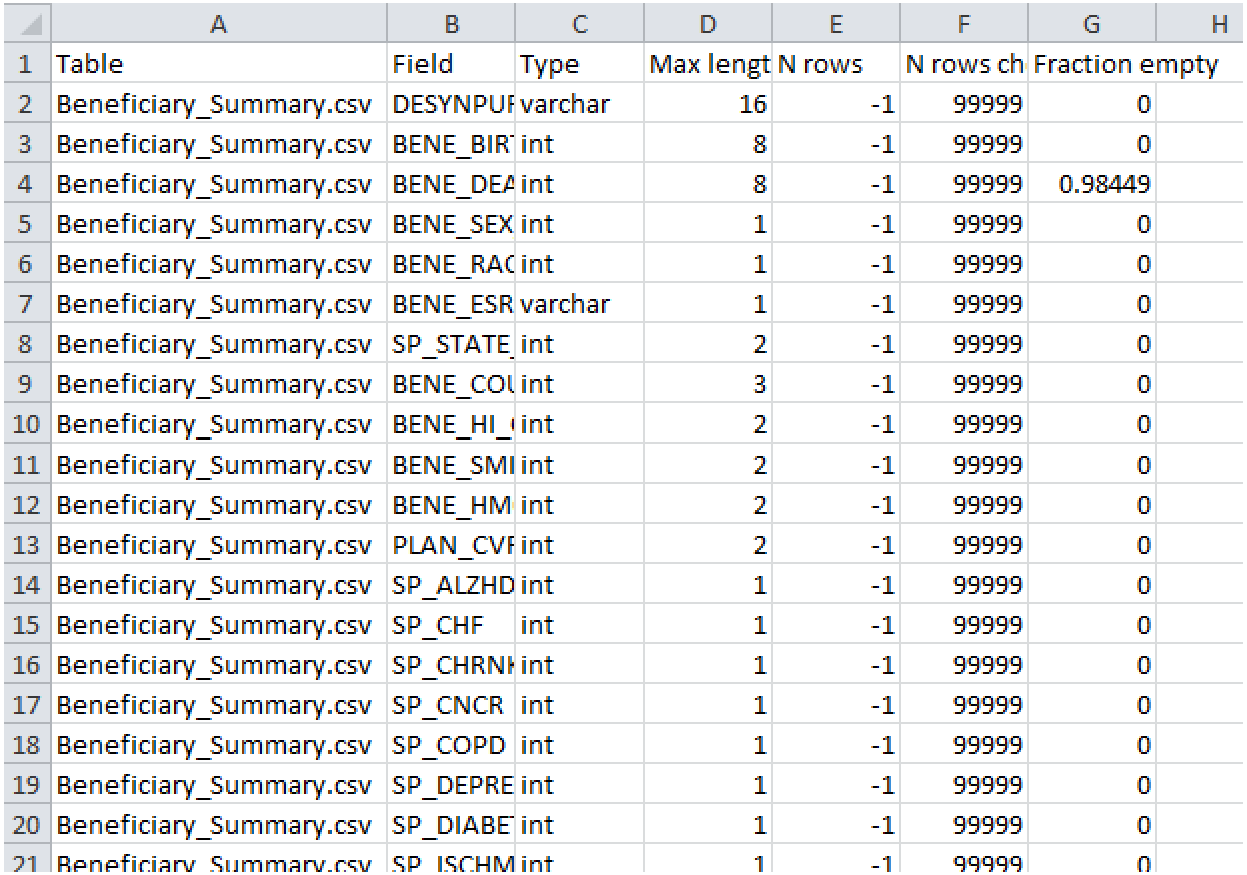
- Set the “Working Folder” to /home/ohdsi/whiterabbit/SynPUFSmall
- Press “Test connection”
- Move over to the “Scan” tab, and hit the “Add” button. Select the CSVs in the folder.
- Keep the default settings and press “Scan tables”.
- Scan report is created in the folder you specified on the “Locations” tab as “ScanReport.xlsx”.

Reading the Scan

- Series of tabs in an XLSX file
  - **Overview Tab**
    provides the definition of each table analyzed, there will only be one tab of this type
  - **Table Tab(s)**
    a summary column for every column, there will be as many tabs as tables selected to analyze

Overview Tab

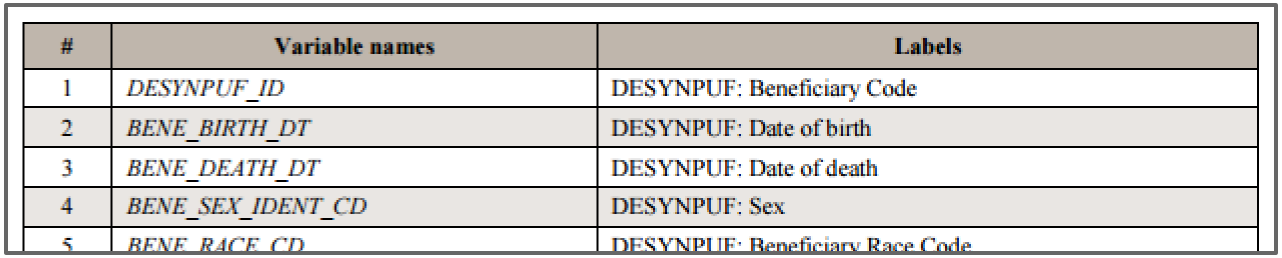
- defines the tables you scanned



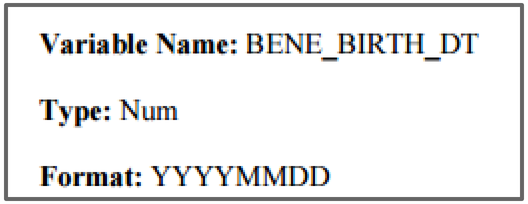
	A	B	C	D	E	F	G	H
1	Table	Field	Type	Max lengt	N rows	N rows ch	Fraction empty	
2	Beneficiary_Summary.csv	DESYNPUF	varchar	16	-1	99999	0	
3	Beneficiary_Summary.csv	BENE_BIRT	int	8	-1	99999	0	
4	Beneficiary_Summary.csv	BENE_DEA	int	8	-1	99999	0.98449	
5	Beneficiary_Summary.csv	BENE_SEX	int	1	-1	99999	0	
6	Beneficiary_Summary.csv	BENE_RAC	int	1	-1	99999	0	
7	Beneficiary_Summary.csv	BENE_ESR	varchar	1	-1	99999	0	
8	Beneficiary_Summary.csv	SP_STATE_	int	2	-1	99999	0	
9	Beneficiary_Summary.csv	BENE_COU	int	3	-1	99999	0	
10	Beneficiary_Summary.csv	BENE_HI_(	int	2	-1	99999	0	
11	Beneficiary_Summary.csv	BENE_SMI	int	2	-1	99999	0	
12	Beneficiary_Summary.csv	BENE_HM	int	2	-1	99999	0	
13	Beneficiary_Summary.csv	PLAN_CVF	int	2	-1	99999	0	
14	Beneficiary_Summary.csv	SP_ALZHD	int	1	-1	99999	0	
15	Beneficiary_Summary.csv	SP_CHF	int	1	-1	99999	0	
16	Beneficiary_Summary.csv	SP_CHRNK	int	1	-1	99999	0	
17	Beneficiary_Summary.csv	SP_CNCR	int	1	-1	99999	0	
18	Beneficiary_Summary.csv	SP_COPD	int	1	-1	99999	0	
19	Beneficiary_Summary.csv	SP_DEPRE	int	1	-1	99999	0	
20	Beneficiary_Summary.csv	SP_DIABET	int	1	-1	99999	0	
21	Beneficiary_Summary.csv	SP_ISCHM	int	1	-1	99999	0	

Table Tabs

- Definition of the Beneficiary_Summary.csv table and each record pertains to a synthetic medicare beneficiary



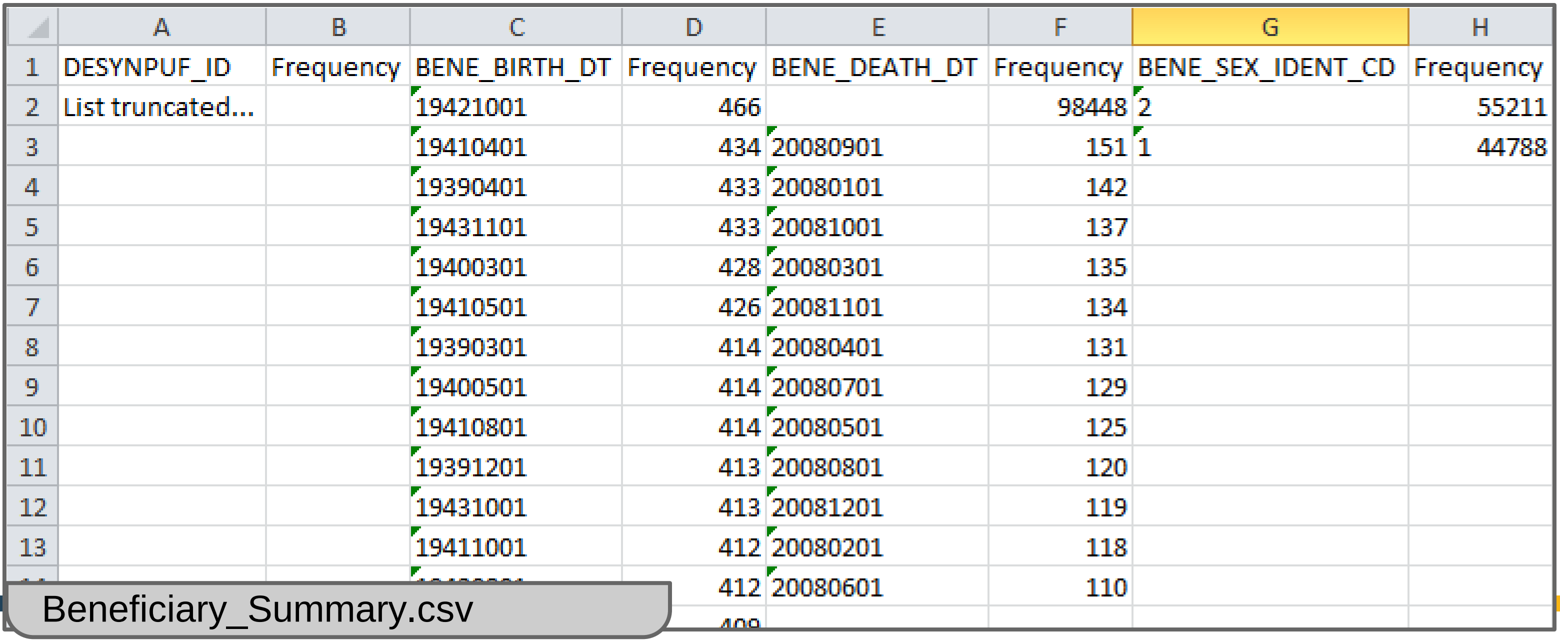
#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*BENE_BIRTH_DT*	DESYNPUF: Date of birth
3	*BENE_DEATH_DT*	DESYNPUF: Date of death
4	*BENE_SEX_IDENT_CD*	DESYNPUF: Sex
5	*BENE_RACE_CD*	DESYNPUF: Beneficiary Race Code



Variable Name: BENE_BIRTH_DT

Type: Num

Format: YYYYMMDD



	A	B	C	D	E	F	G	H
1	DESYNPUF_ID	Frequency	BENE_BIRTH_DT	Frequency	BENE_DEATH_DT	Frequency	BENE_SEX_IDENT_CD	Frequency
2	List truncated...		19421001	466		98448	2	55211
3			19410401	434	20080901	151	1	44788
4			19390401	433	20080101	142		
5			19431101	433	20081001	137		
6			19400301	428	20080301	135		
7			19410501	426	20081101	134		
8			19390301	414	20080401	131		
9			19400501	414	20080701	129		
10			19410801	414	20080501	125		
11			19391201	413	20080801	120		
12			19431001	413	20081201	119		
13			19411001	412	20080201	118		
14				412	20080601	110		
				409				

Beneficiary_Summary.csv

Read the Scan Report

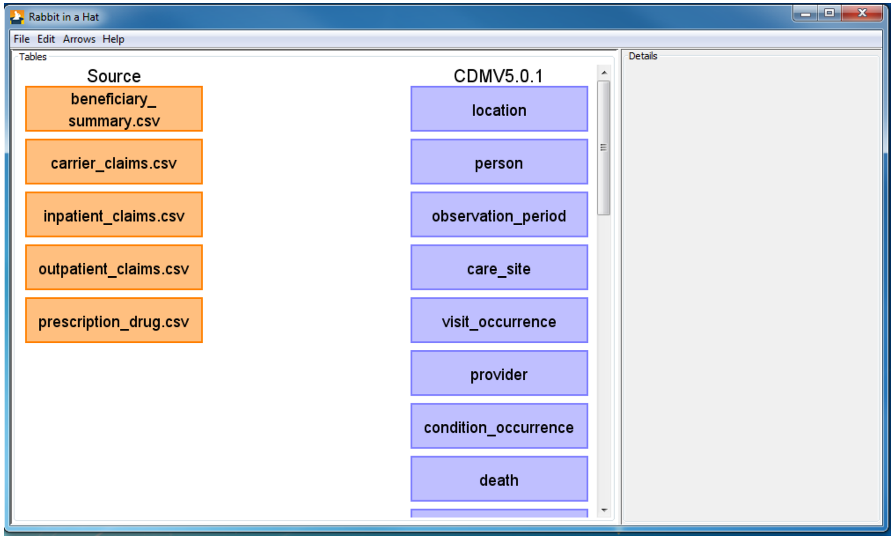
- Open up the scan report generated
- Go to the “Inpatient_Claims.csv” tab which represents claims processed from inpatient setting.
- What is the most common admitting diagnosis code?
- Hints:
  - ADMTNG_ICD9_DGNS_CD
  - ICD9 codes are in ###.## format
  - You can use ATLAS to look it up

Rabbit In a Hat

- Already part of the WhiteRabbit download
- Uses the information from WhiteRabbit to help you produce documentation for the ETL process
- Helps you define the logic in a consistent way does not generate code to create ETL

Getting Started

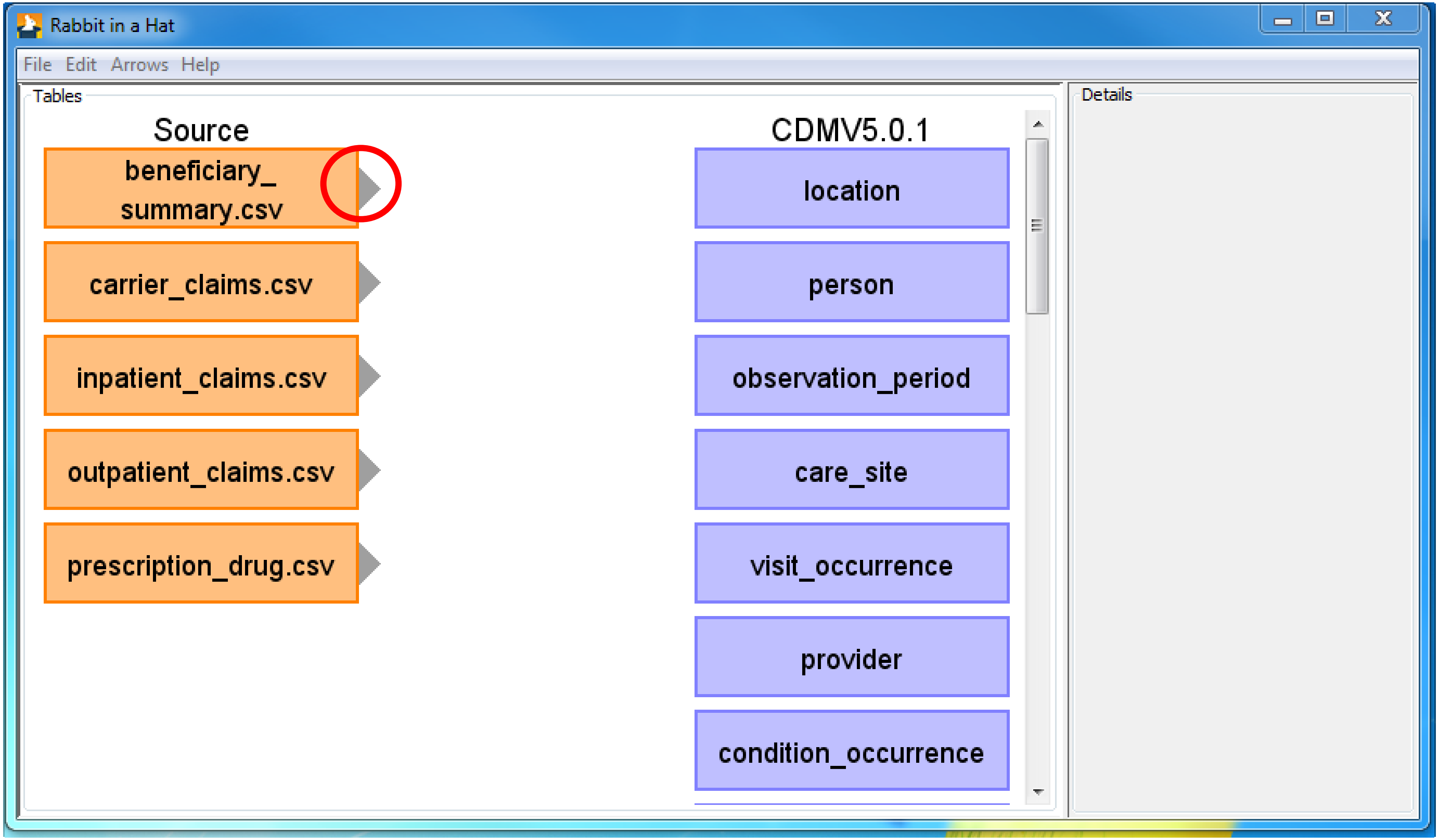
- Double-click on the RabbitInAHat.jar to start the application.
- File → Open Scan Report and navigate to the scan report that was just created.



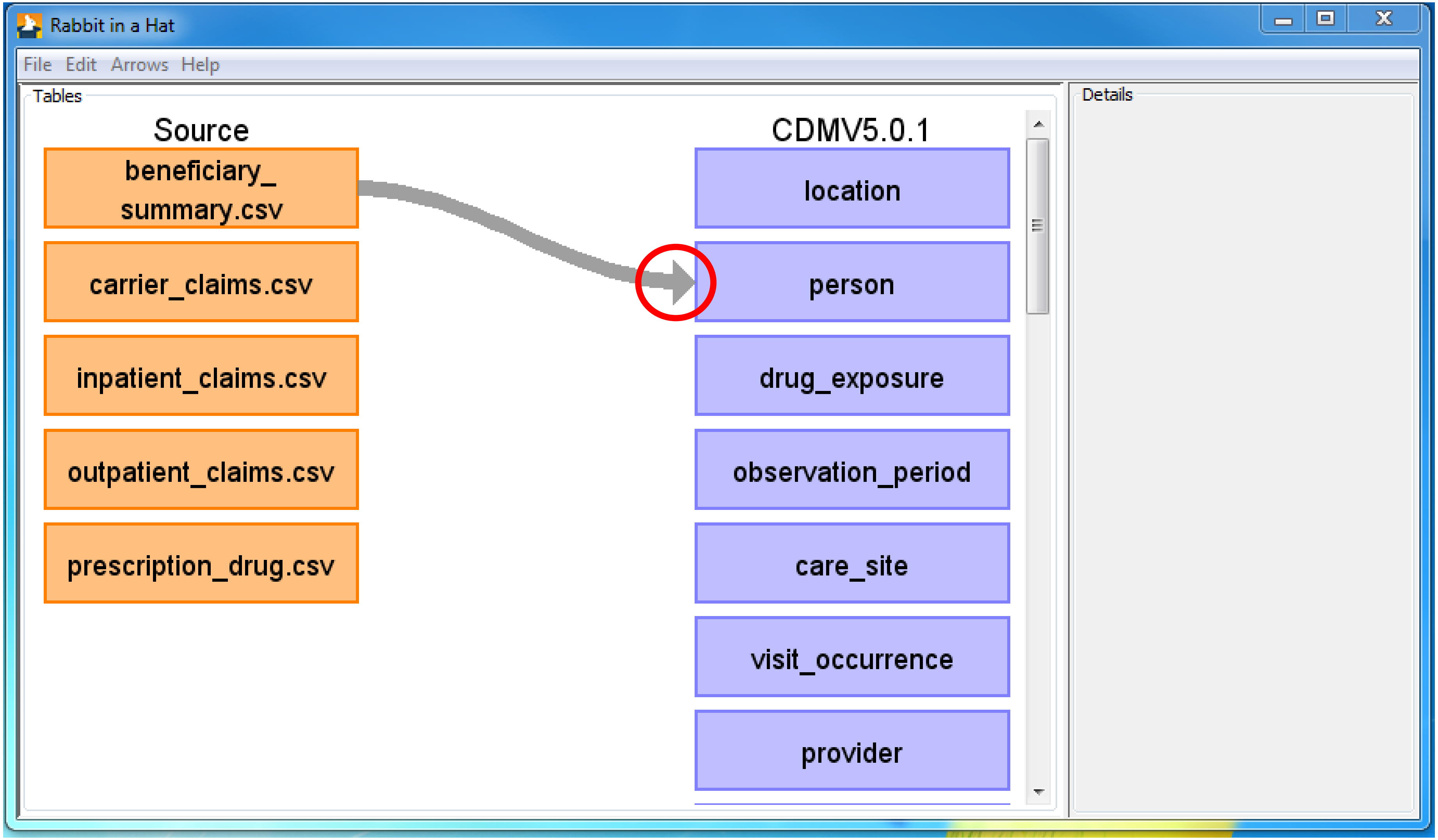
Process for Developing ETL

- Get the right people in the room
- Block off time necessary
- Map all the raw data tables to CDM tables
- Then go back through and provide detailed mapping information for each raw data table to CDM table
- Generate final ETL document

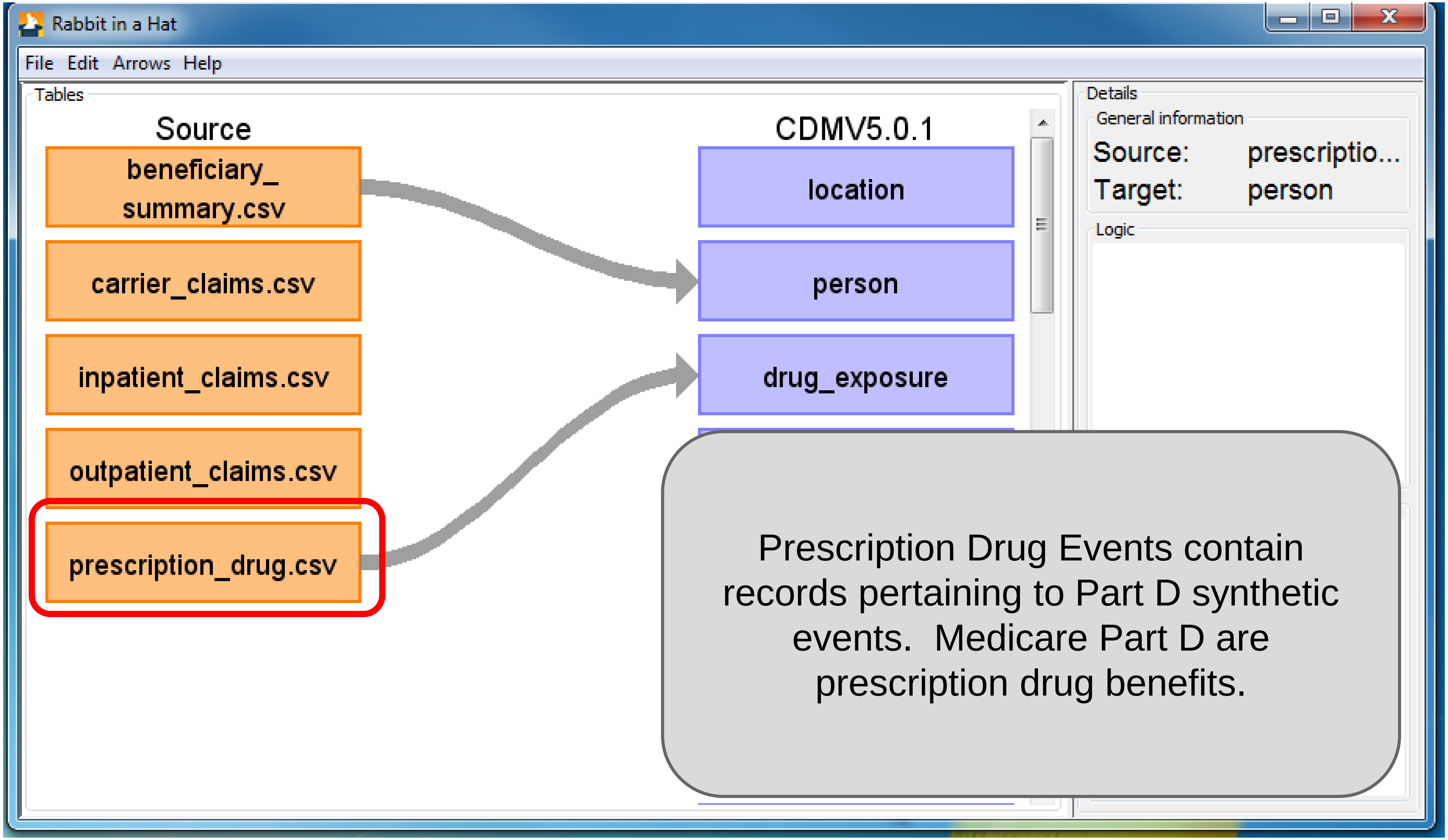
Map Raw Tables to CDM Tables



Map Raw Tables to CDM Tables



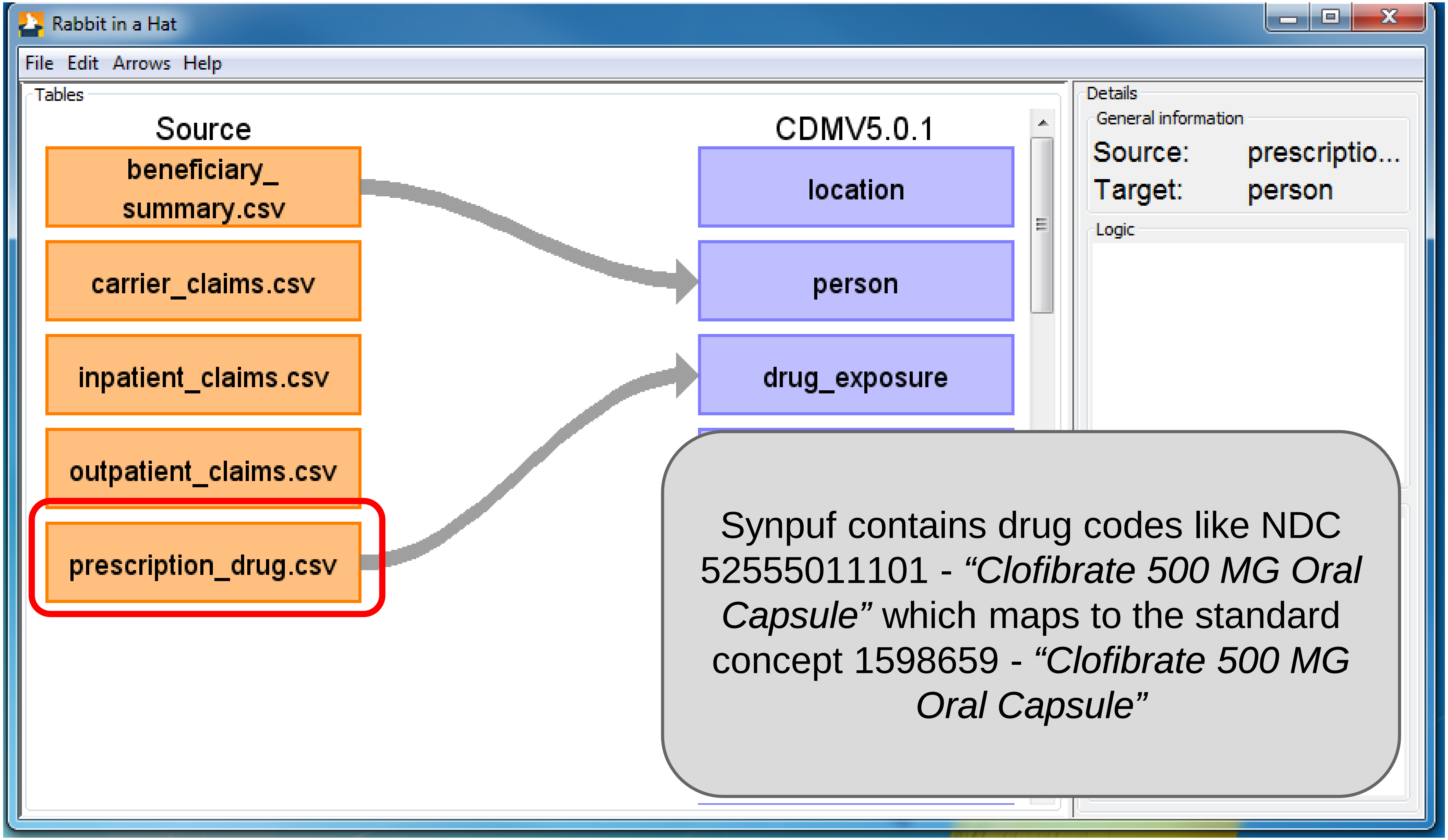
Map Raw Tables to CDM Tables



Source: prescriptio...
Target: person

Prescription Drug Events contain records pertaining to Part D synthetic events. Medicare Part D are prescription drug benefits.

Map Raw Tables to CDM Tables

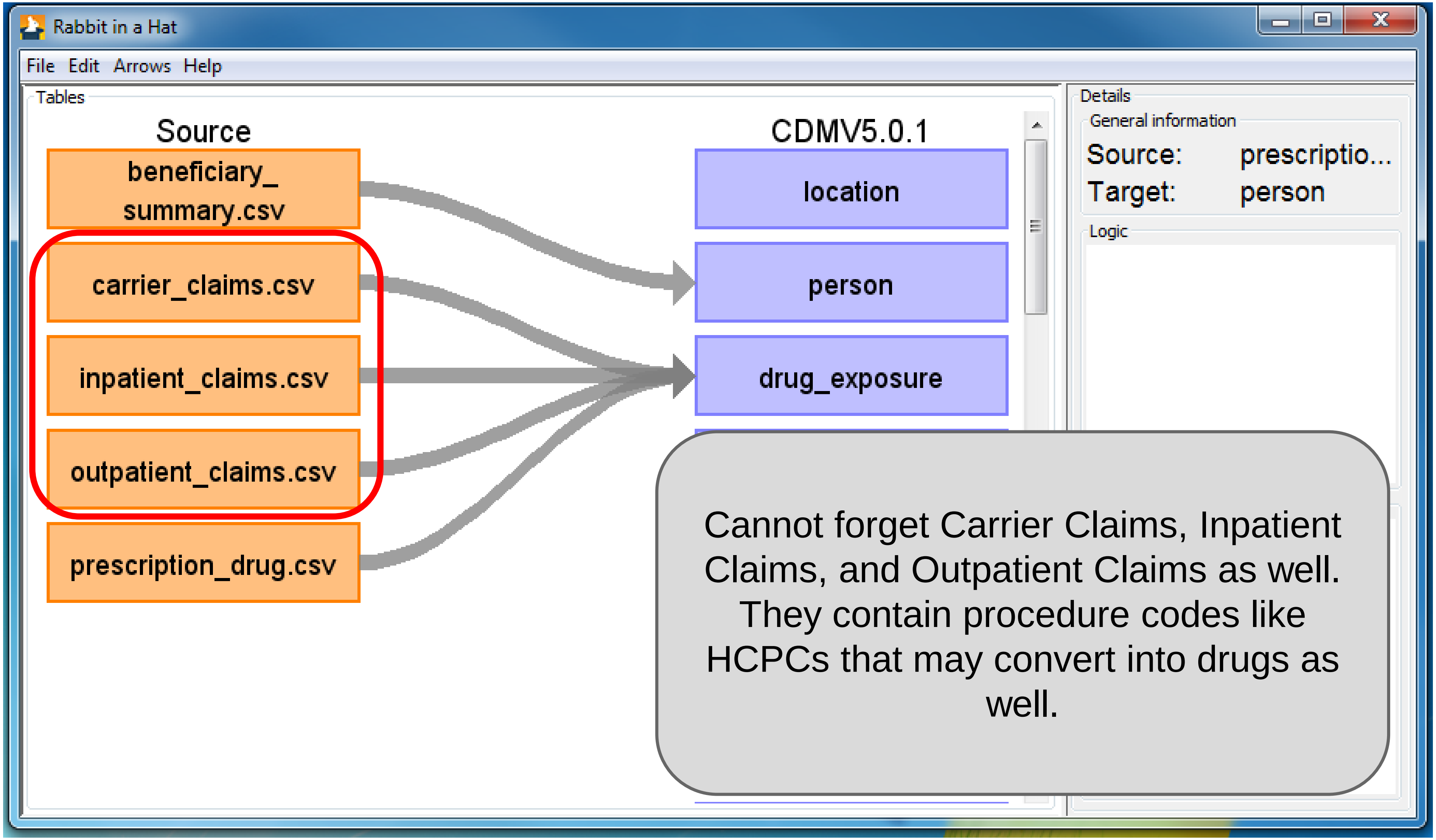


Source: prescriptio...

Target: person

Synpuf contains drug codes like NDC 52555011101 - *"Clofibrate 500 MG Oral Capsule"* which maps to the standard concept 1598659 - *"Clofibrate 500 MG Oral Capsule"*

Map Raw Tables to CDM Tables

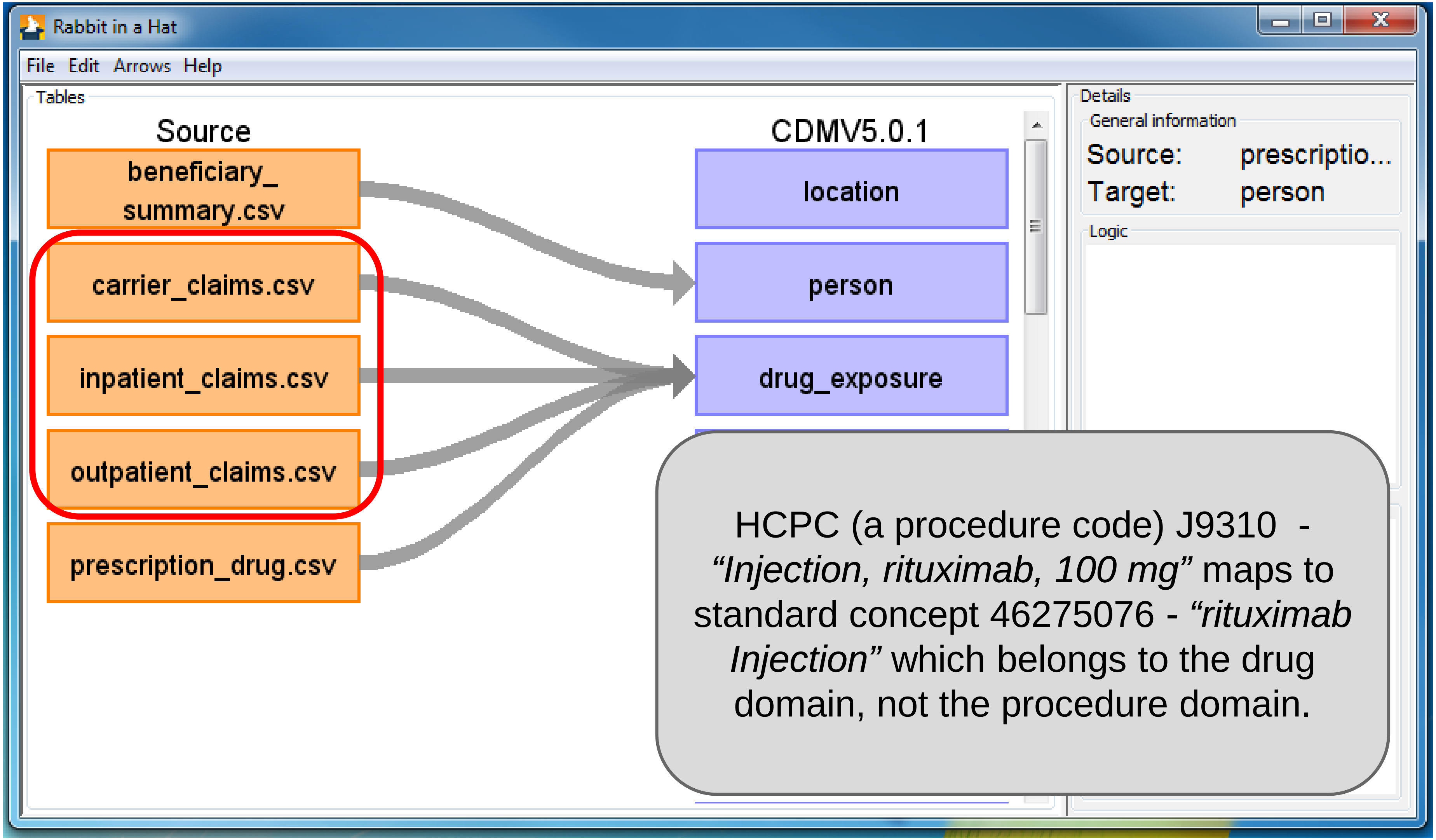


Source: prescriptio...

Target: person

Cannot forget Carrier Claims, Inpatient Claims, and Outpatient Claims as well. They contain procedure codes like HCPCs that may convert into drugs as well.

Map Raw Tables to CDM Tables

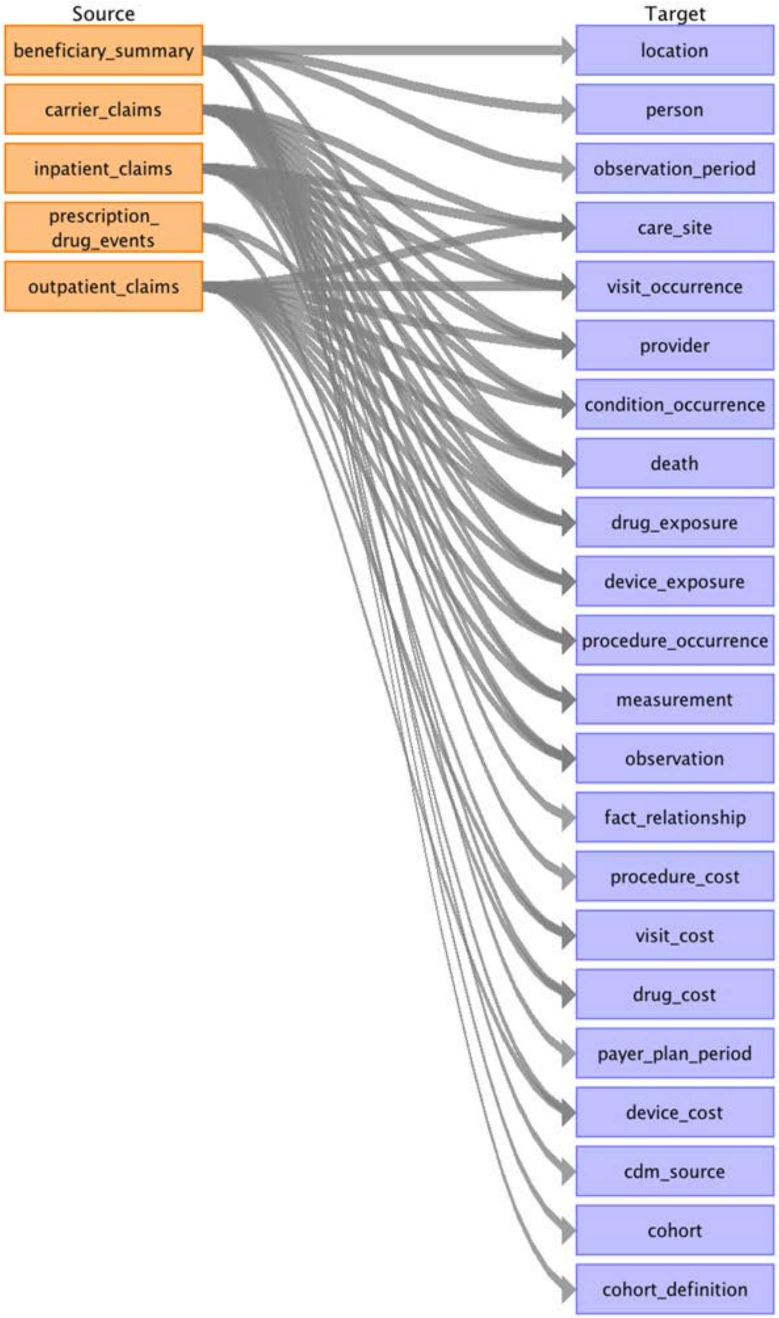


Source: prescriptio...
Target: person

HCPC (a procedure code) J9310 - *"Injection, rituximab, 100 mg"* maps to standard concept 46275076 - *"rituximab Injection"* which belongs to the drug domain, not the procedure domain.

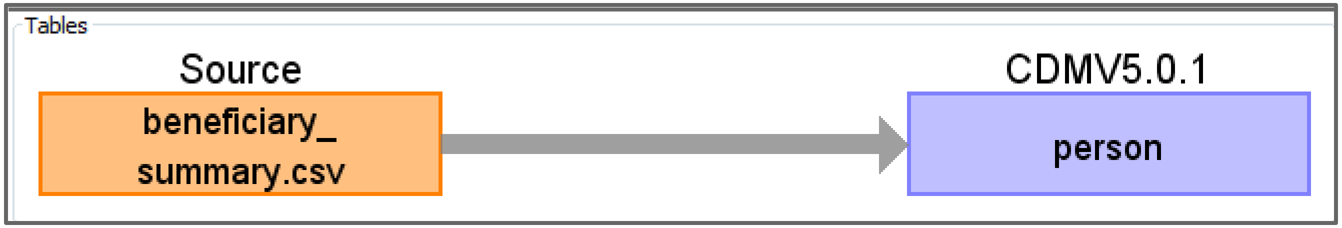
Map Raw Tables to CDM Tables

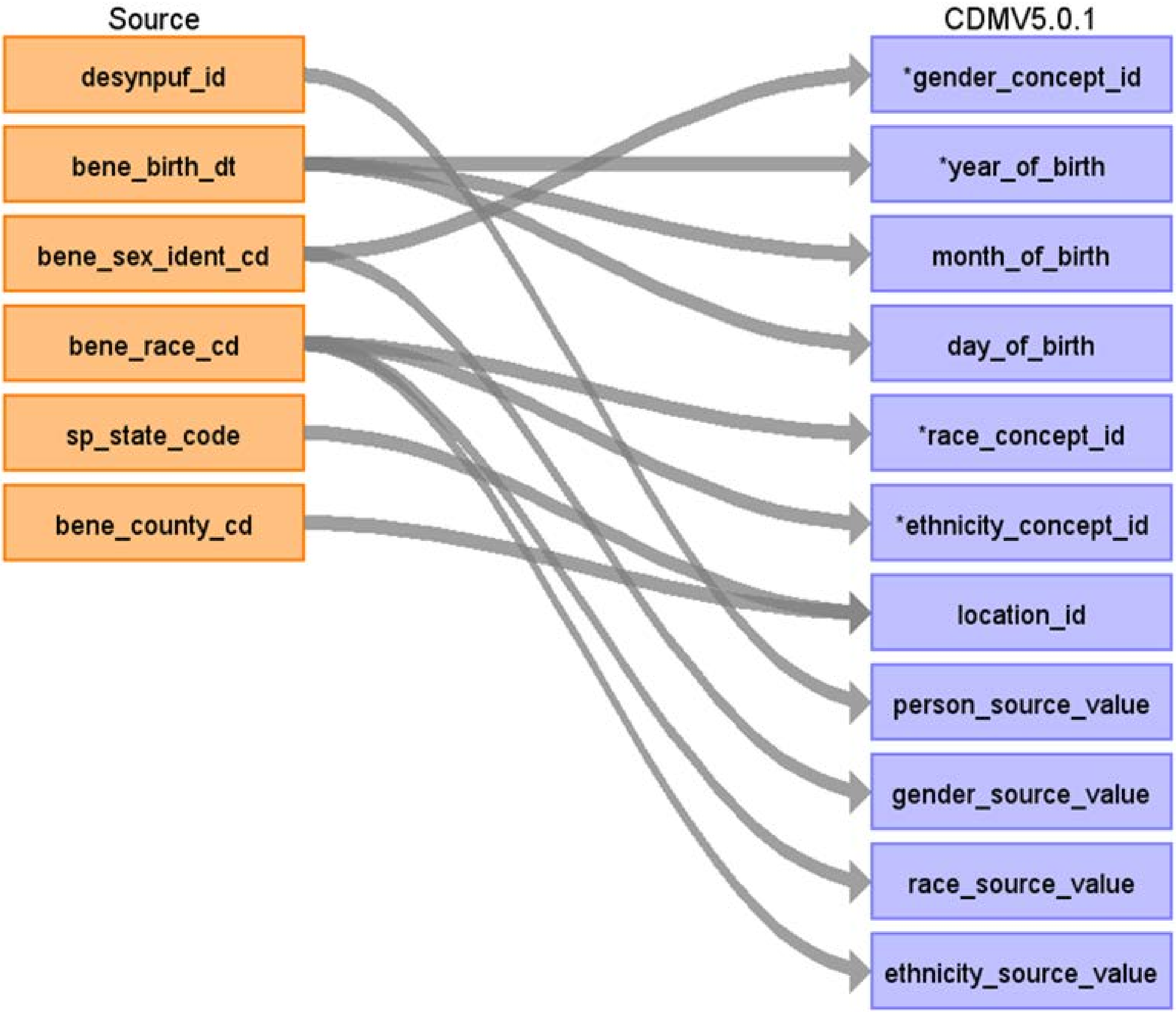
Continue mapping raw tables to CDM tables until you feel confident you are bringing over as much raw data as necessary



PERSON

- For today's example we'll start with the PERSON table





Source

desynpuf_id

bene_birth_dt

bene_sex_ident_cd

bene_race_cd

sp_state_code

bene_county_cd

CDMV5.0.1

*gender_concept_id

*year_of_birth

month_of_birth

day_of_birth

*race_concept_id

*ethnicity_concept_id

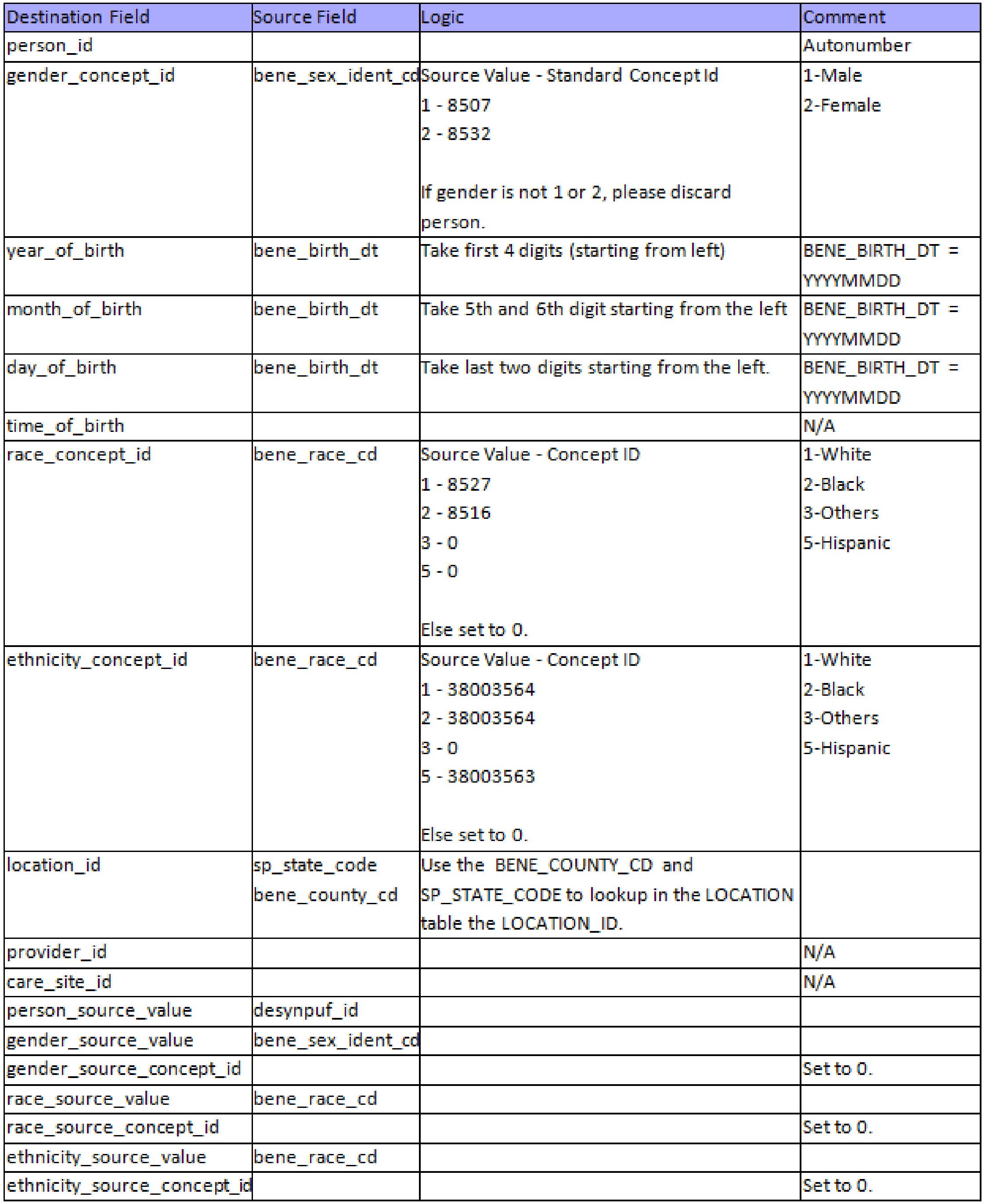
location_id

person_source_value

gender_source_value

race_source_value

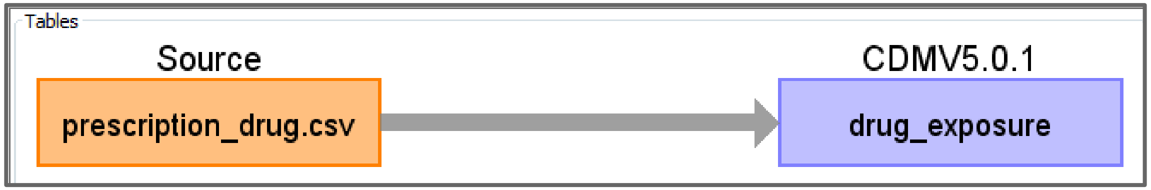
ethnicity_source_value



Destination Field	Source Field	Logic	Comment
person_id			Autonumber
gender_concept_id	bene_sex_ident_cd	Source Value - Standard Concept Id 1 - 8507 2 - 8532 If gender is not 1 or 2, please discard person.	1-Male 2-Female
year_of_birth	bene_birth_dt	Take first 4 digits (starting from left)	BENE_BIRTH_DT = YYYYMMDD
month_of_birth	bene_birth_dt	Take 5th and 6th digit starting from the left	BENE_BIRTH_DT = YYYYMMDD
day_of_birth	bene_birth_dt	Take last two digits starting from the left.	BENE_BIRTH_DT = YYYYMMDD
time_of_birth			N/A
race_concept_id	bene_race_cd	Source Value - Concept ID 1 - 8527 2 - 8516 3 - 0 5 - 0 Else set to 0.	1-White 2-Black 3-Others 5-Hispanic
ethnicity_concept_id	bene_race_cd	Source Value - Concept ID 1 - 38003564 2 - 38003564 3 - 0 5 - 38003563 Else set to 0.	1-White 2-Black 3-Others 5-Hispanic
location_id	sp_state_code bene_county_cd	Use the BENE_COUNTY_CD and SP_STATE_CODE to lookup in the LOCATION table the LOCATION_ID.	
provider_id			N/A
care_site_id			N/A
person_source_value	desynpuf_id		
gender_source_value	bene_sex_ident_cd		
gender_source_concept_id			Set to 0.
race_source_value	bene_race_cd		
race_source_concept_id			Set to 0.
ethnicity_source_value	bene_race_cd		
ethnicity_source_concept_id			Set to 0.

DRUG_EXPOSURE

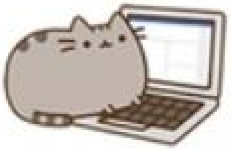
- Try drawing arrows from PRESCRIPTION_DRUG_EVENTS columns to DRUG_EXPOSURE columns

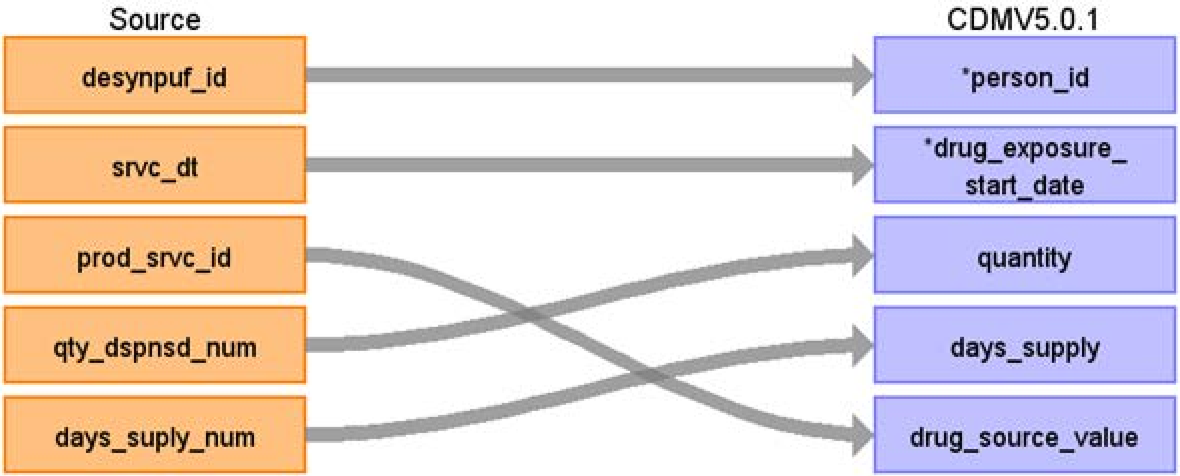


Focus on:

- PERSON_ID
- DRUG_EXPOSURE_START_DATE
- QUANTITY
- DAYS_SUPPLY
- DRUG_SOURCE_VALUE

DRUG_EXPOSURE

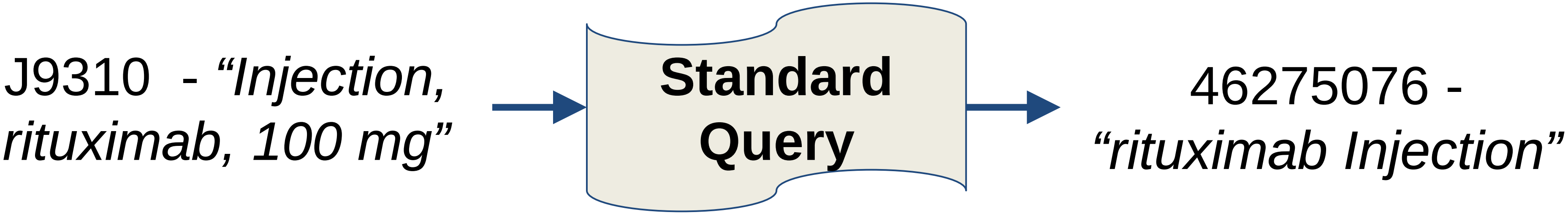




DRUG_EXPOSURE

- Mapping source codes to standard terminology
  - –Source to Source
  - –Source to Standard

- Use standard query for both, just define filters needed



J9310 - *“Injection, rituximab, 100 mg”*

Standard Query

46275076 - *“rituximab Injection”*

Standard Query: Source to Standard

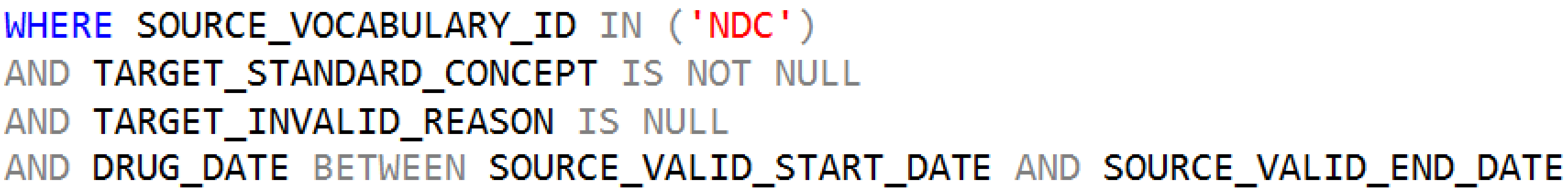
```
WITH CTE_VOCAB_MAP AS (
    SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.concept_name AS
   SOURCE_CODE_DESCRIPTION, c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS
   SOURCE_DOMAIN_ID, c.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID, c.VALID_START_DATE AS
   SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.INVALID_REASON AS
   SOURCE_INVALID_REASON,c1.concept_id AS TARGET_CONCEPT_ID, c1.concept_name AS
   TARGET_CONCEPT_NAME, c1.VOCABULARY_ID AS TARGET_VOCABUALRY_ID, c1.domain_id AS
   TARGET_DOMAIN_ID, c1.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c1.INVALID_REASON AS
   TARGET_INVALID_REASON, c1.standard_concept AS TARGET_STANDARD_CONCEPT
    FROM CONCEPT C
         JOIN CONCEPT_RELATIONSHIP CR ON C.CONCEPT_ID = CR.CONCEPT_ID_1  AND CR.invalid_reason IS NULL
                  AND cr.relationship_id = 'Maps to'
          JOIN CONCEPT C1 ON CR.CONCEPT_ID_2 = C1.CONCEPT_ID AND C1.INVALID_REASON IS NULL
   UNION
   SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id
   AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID, c1.VALID_START_DATE AS
   SOURCE_VALID_START_DATE, c1.VALID_END_DATE AS SOURCE_VALID_END_DATE, stcm.INVALID_REASON AS
   SOURCE_INVALID_REASON,target_concept_id, c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME,
   target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID, c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID,
   c2.INVALID_REASON AS TARGET_INVALID_REASON, c2.standard_concept AS TARGET_STANDARD_CONCEPT
      FROM source_to_concept_map stcm
           LEFT OUTER JOIN CONCEPT c1 ON c1.concept_id = stcm.source_concept_id
           LEFT OUTER JOIN CONCEPT c2 ON c2.CONCEPT_ID = stcm.target_concept_id
      WHERE stcm.INVALID_REASON IS NULL
)
SELECT *
FROM CTE_VOCAB_MAP
/*EXAMPLE FILTERS*/
WHERE SOURCE_VOCABULARY_ID IN ('NDC')
AND TARGET_STANDARD_CONCEPT IN ('S')
```

Standard Query:
Source to Source

```
WITH CTE_VOCAB_MAP AS (
    SELECT c.concept_code AS SOURCE_CODE, c.concept_id AS SOURCE_CONCEPT_ID, c.CONCEPT_NAME AS
   SOURCE_CODE_DESCRIPTION, c.vocabulary_id AS SOURCE_VOCABULARY_ID, c.domain_id AS
   SOURCE_DOMAIN_ID, c.concept_class_id AS SOURCE_CONCEPT_CLASS_ID, c.VALID_START_DATE AS
   SOURCE_VALID_START_DATE, c.VALID_END_DATE AS SOURCE_VALID_END_DATE, c.invalid_reason AS
   SOURCE_INVALID_REASON, c.concept_ID as TARGET_CONCEPT_ID, c.concept_name AS
   TARGET_CONCEPT_NAME, c.vocabulary_id AS TARGET_VOCABULARY_ID, c.domain_id AS TARGET_DOMAIN_ID,
   c.concept_class_id AS TARGET_CONCEPT_CLASS_ID, c.INVALID_REASON AS
   TARGET_INVALID_REASON,c.STANDARD_CONCEPT AS TARGET_STANDARD_CONCEPT
    FROM CONCEPT c
    UNION
    SELECT source_code, SOURCE_CONCEPT_ID, SOURCE_CODE_DESCRIPTION, source_vocabulary_id, c1.domain_id
   AS SOURCE_DOMAIN_ID, c2.CONCEPT_CLASS_ID AS SOURCE_CONCEPT_CLASS_ID, c1.VALID_START_DATE AS
   SOURCE_VALID_START_DATE, c1.VALID_END_DATE AS SOURCE_VALID_END_DATE,stcm.INVALID_REASON AS
   SOURCE_INVALID_REASON,target_concept_id, c2.CONCEPT_NAME AS TARGET_CONCEPT_NAME,
   target_vocabulary_id, c2.domain_id AS TARGET_DOMAIN_ID, c2.concept_class_id AS TARGET_CONCEPT_CLASS_ID,
   c2.INVALID_REASON AS TARGET_INVALID_REASON, c2.standard_concept AS TARGET_STANDARD_CONCEPT
    FROM source_to_concept_map stcm
          LEFT OUTER JOIN CONCEPT c1 ON c1.concept_id = stcm.source_concept_id
          LEFT OUTER JOIN CONCEPT c2 ON c2.CONCEPT_ID = stcm.target_concept_id
    WHERE stcm.INVALID_REASON IS NULL
)
SELECT *
FROM CTE_VOCAB_MAP
/*EXAMPLE FILTERS*/
WHERE SOURCE_VOCABULARY_ID IN ('ICD9CM')
AND TARGET_VOCABULARY_ID IN ('ICD9CM')
```

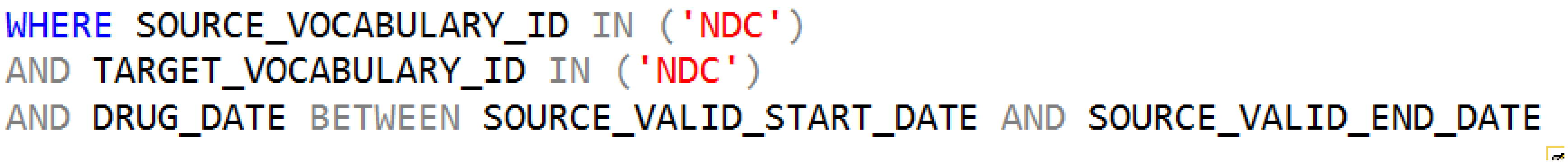
Example Filters: NDCs

- Source to Standard



```
WHERE SOURCE_VOCABULARY_ID IN ('NDC')
AND TARGET_STANDARD_CONCEPT IS NOT NULL
AND TARGET_INVALID_REASON IS NULL
AND DRUG_DATE BETWEEN SOURCE_VALID_START_DATE AND SOURCE_VALID_END_DATE
```

- Source to Source



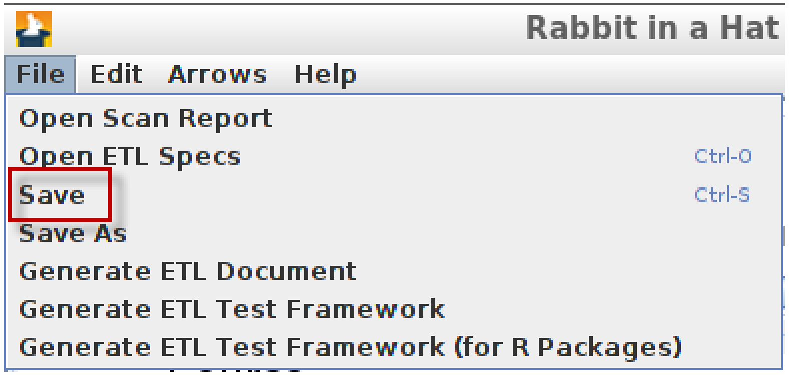
```
WHERE SOURCE_VOCABULARY_ID IN ('NDC')
AND TARGET_VOCABULARY_ID IN ('NDC')
AND DRUG_DATE BETWEEN SOURCE_VALID_START_DATE AND SOURCE_VALID_END_DATE
```

Some maps are date sensitive like NDC or DRGs

Review for incorrect mappings (e.g. source codes might map to multiple SOURCE_VOCAB_IDs)

Saving and Export to Document

- Save working document

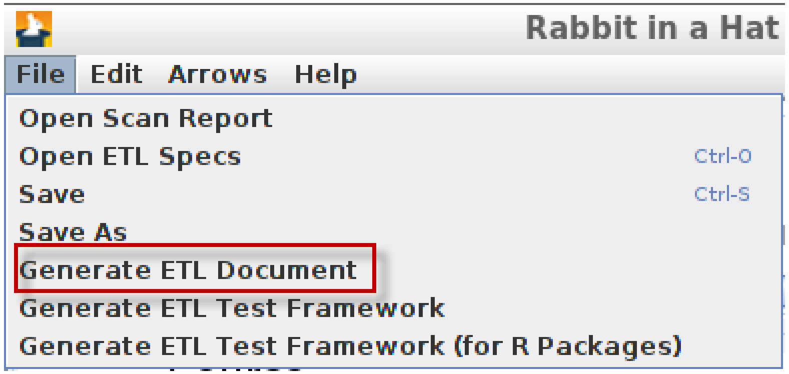


Rabbit in a Hat

File Edit Arrows Help

Open Scan Report
Open ETL Specs Ctrl-O
Save Ctrl-S
Save As
Generate ETL Document
Generate ETL Test Framework
Generate ETL Test Framework (for R Packages)

- Export to document



Rabbit in a Hat

File Edit Arrows Help

Open Scan Report
Open ETL Specs Ctrl-O
Save Ctrl-S
Save As
Generate ETL Document
Generate ETL Test Framework
Generate ETL Test Framework (for R Packages)

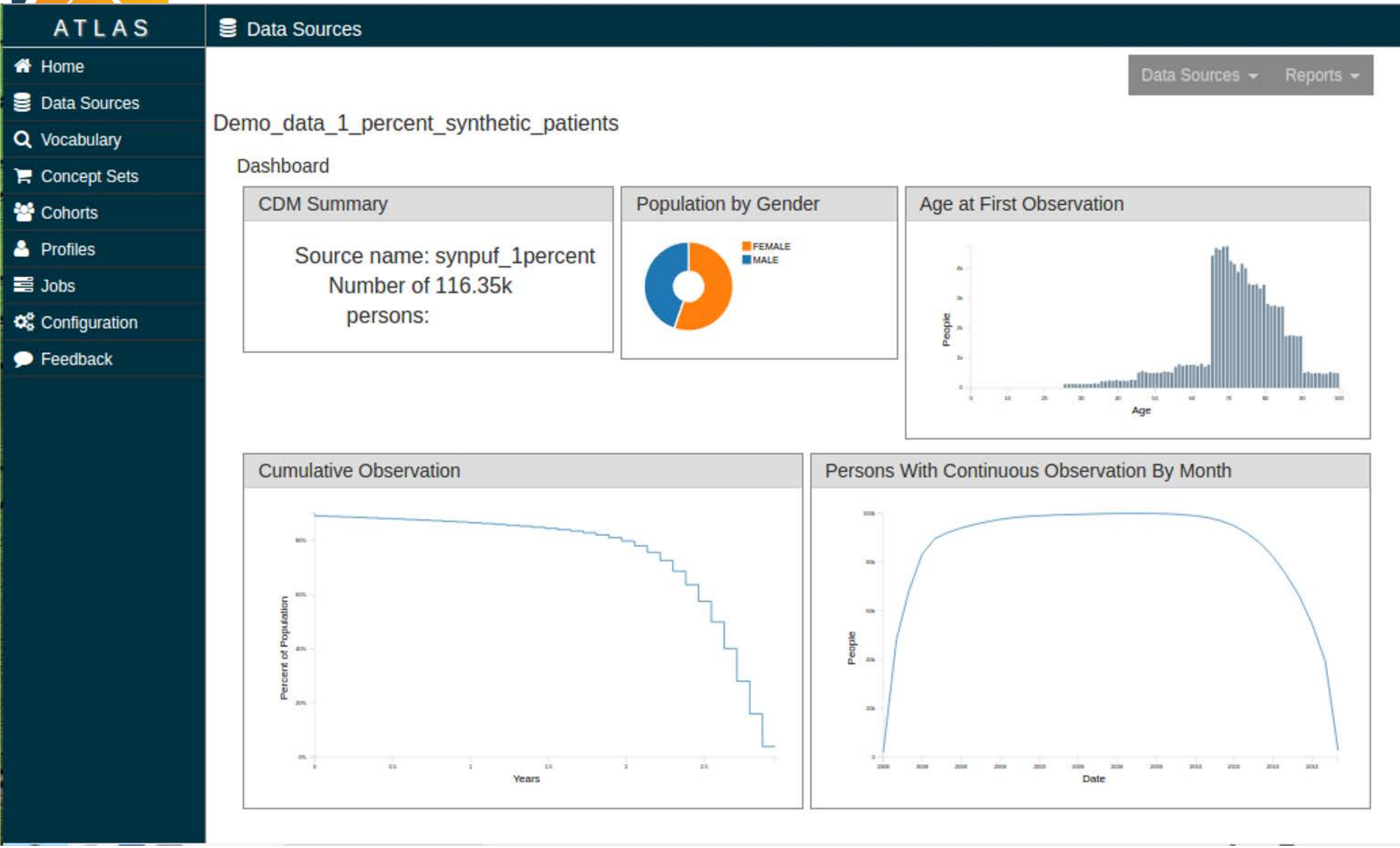
Evaluation

Testing and Quality Assurance

ACHILLES

- Interactive platform to visualize data in CDM
  - patient demographics
  - prevalence of conditions, drugs and procedures
  - distribution of values for clinical observations

- https://github.com/OHDSI/Achilles



Data Sources

Demo_data_1_percent_synthetic_patients

Dashboard

CDM Summary

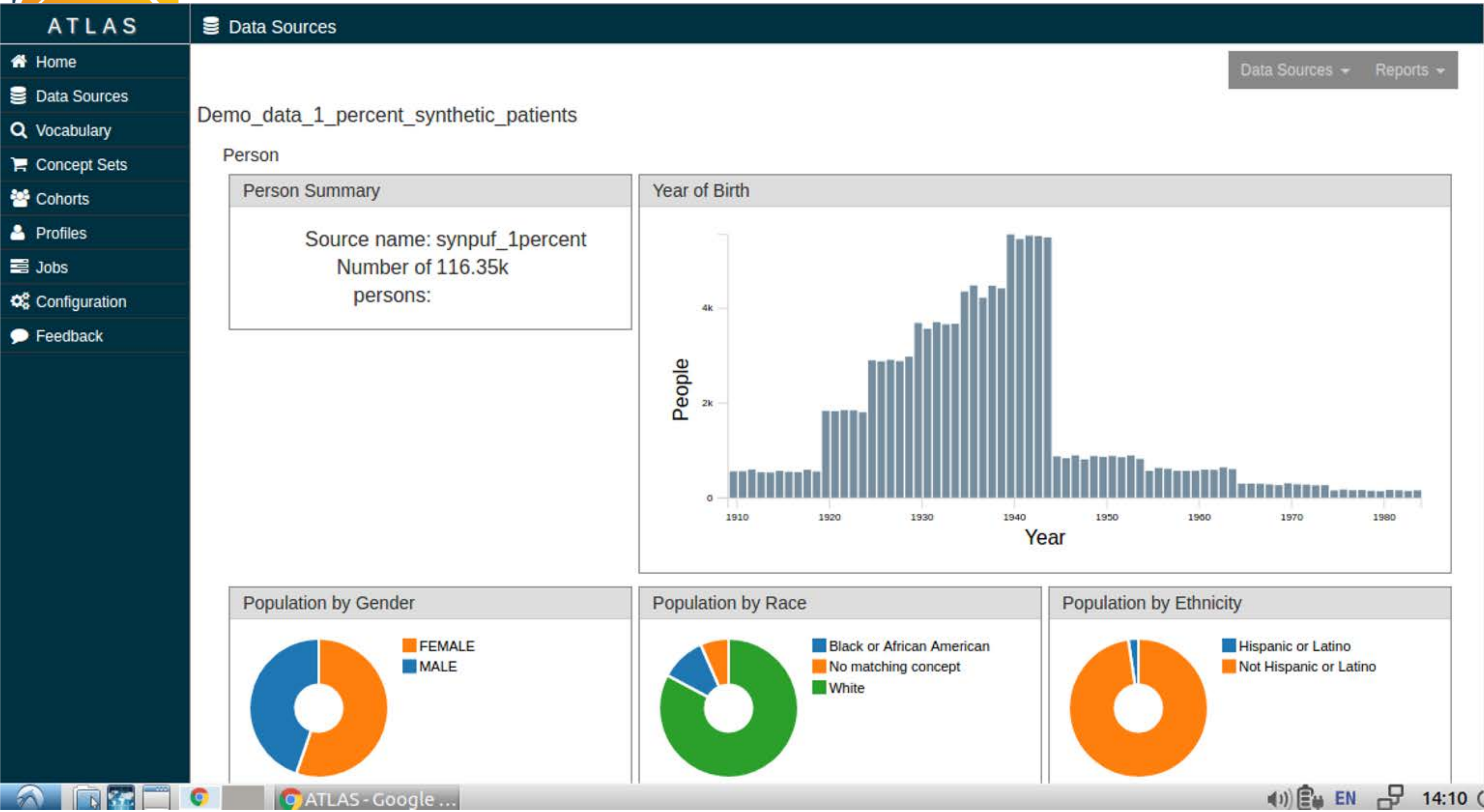
Source name: synpuf_1percent
Number of persons: 116.35k

Population by Gender

Age at First Observation

Cumulative Observation

Persons With Continuous Observation By Month

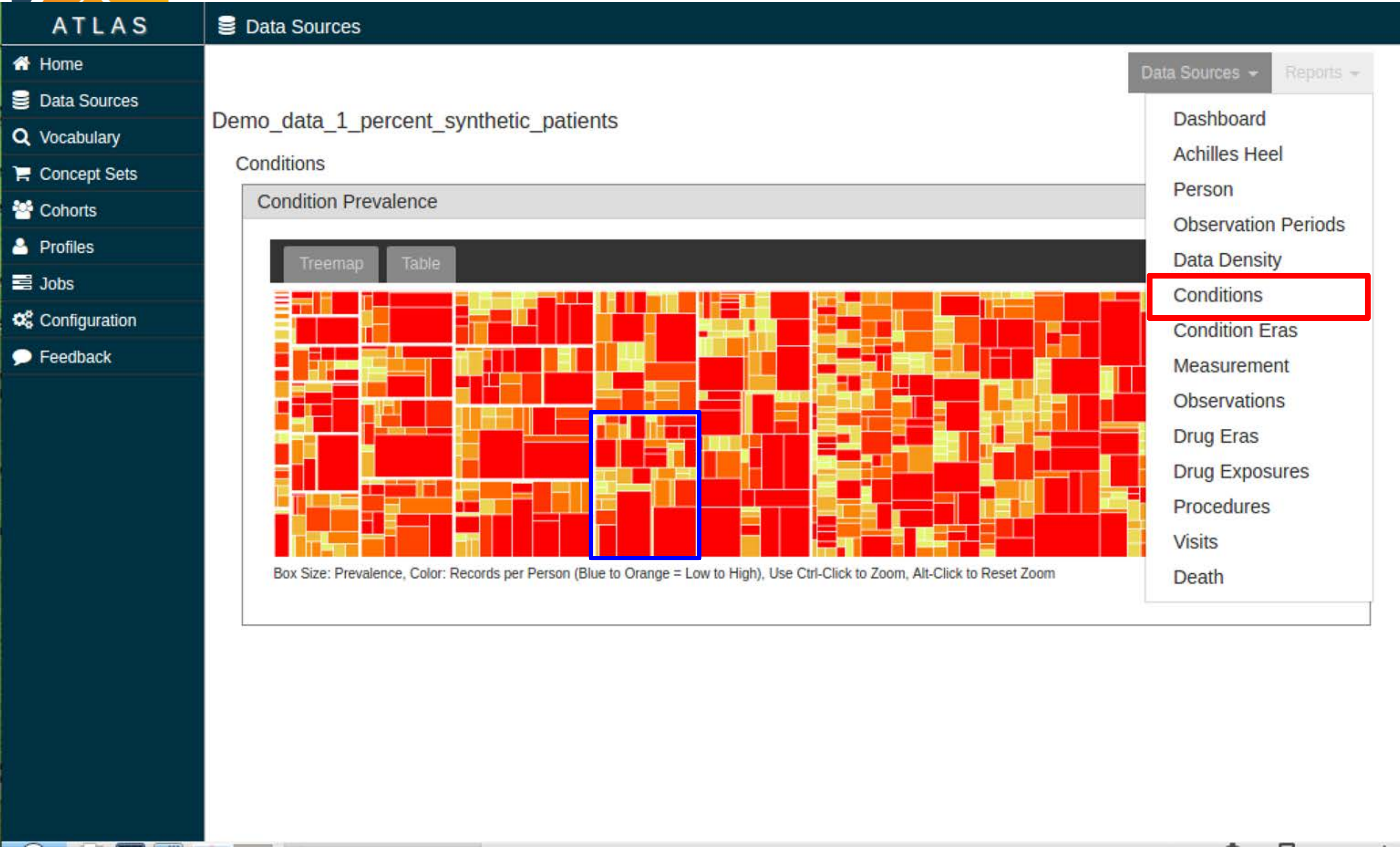


Demo_data_1_percent_synthetic_patients

Person

Person Summary

Source name: synpuf_1percent
Number of 116.35k
persons:

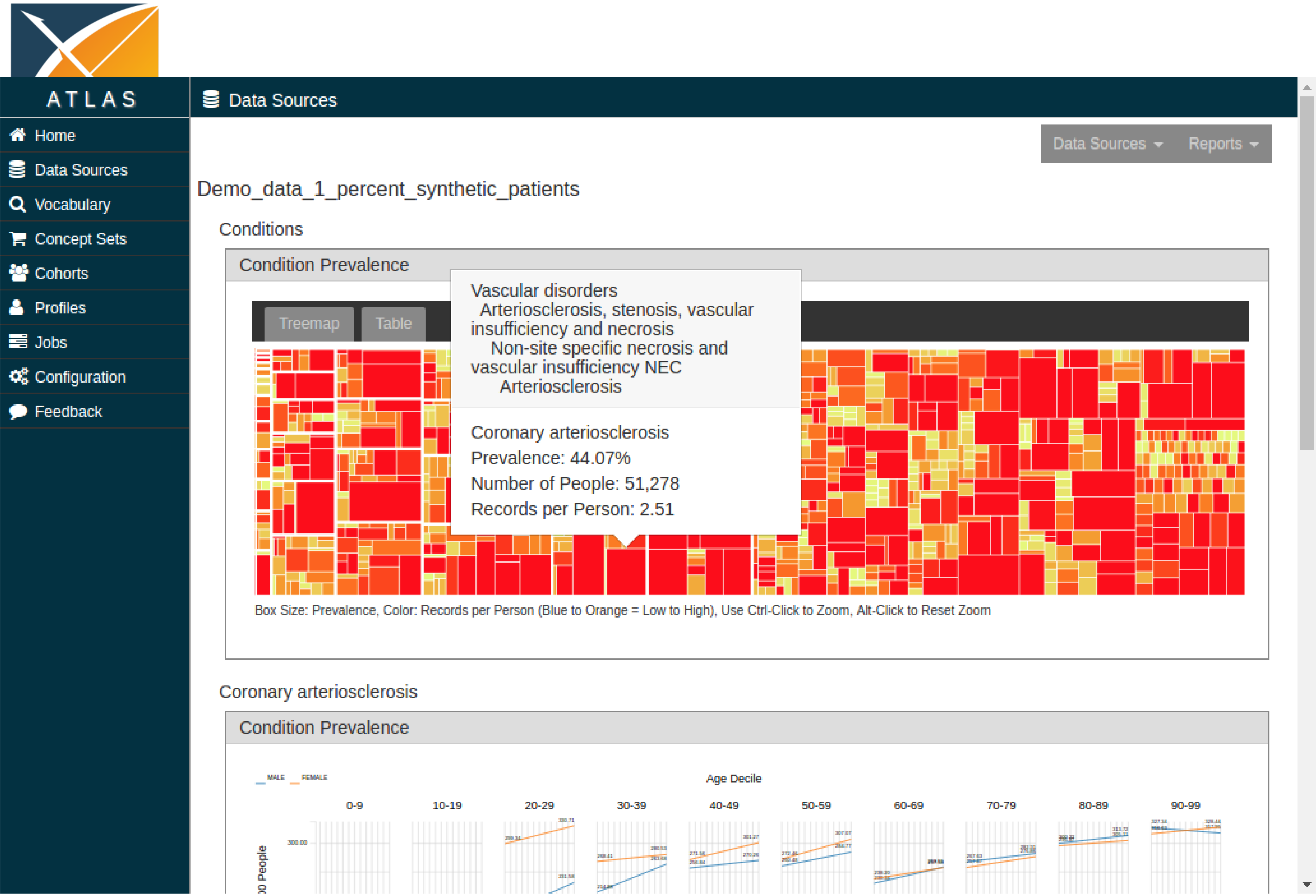


Demo_data_1_percent_synthetic_patients

Conditions

Condition Prevalence

Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom



Demo_data_1_percent_synthetic_patients

Conditions

Condition Prevalence

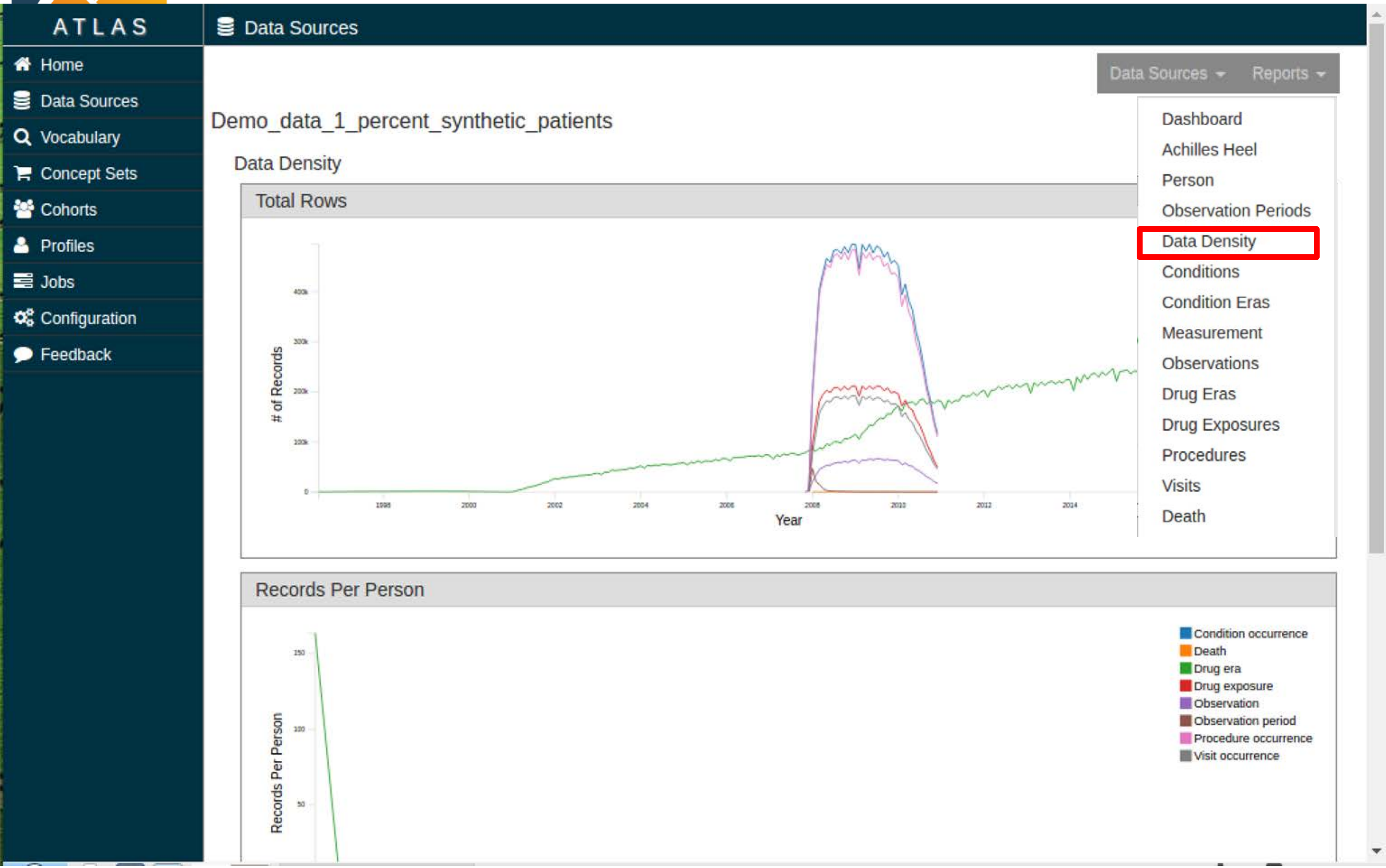
Vascular disorders
Arteriosclerosis, stenosis, vascular insufficiency and necrosis
Non-site specific necrosis and vascular insufficiency NEC
Arteriosclerosis

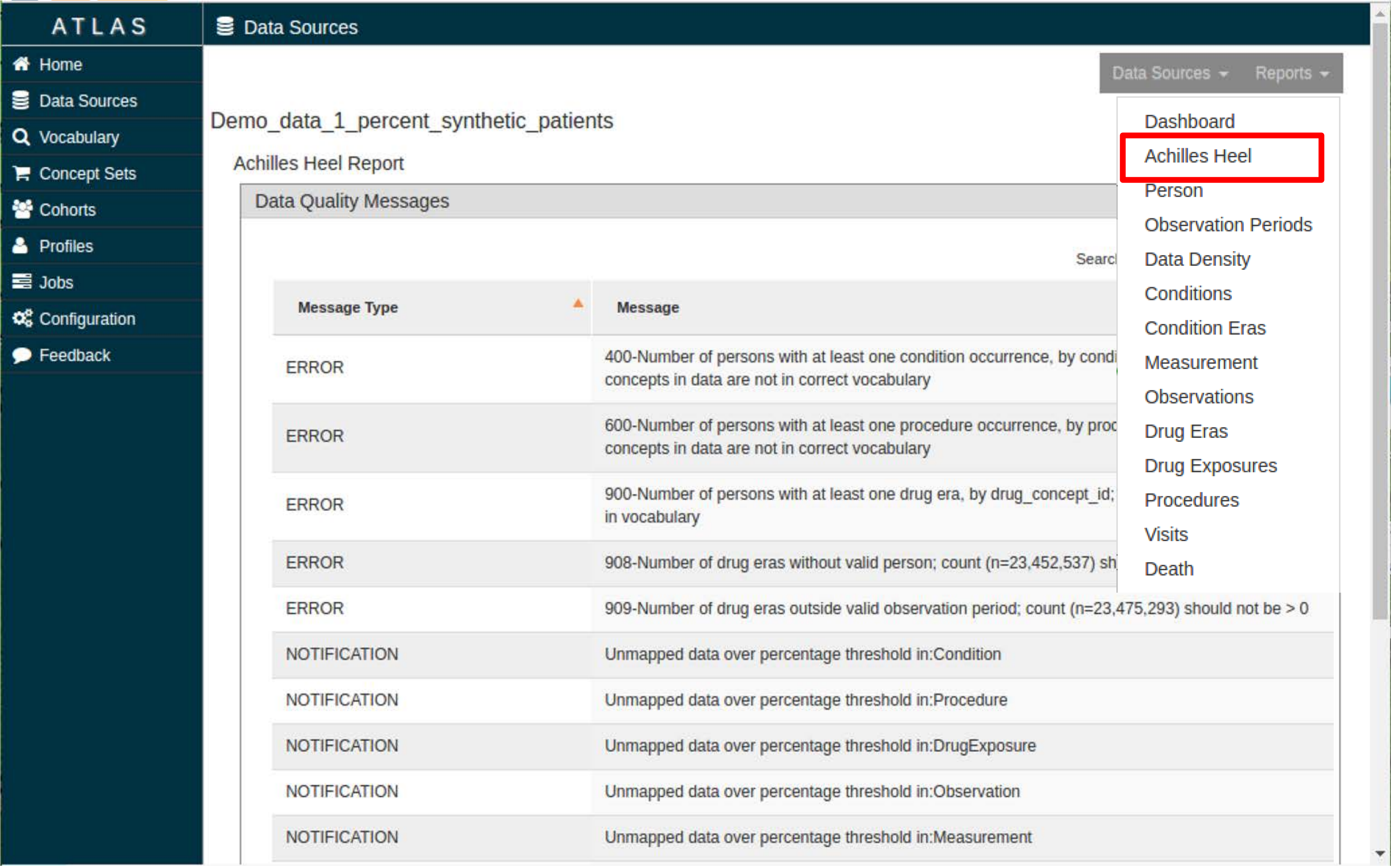
Coronary arteriosclerosis
Prevalence: 44.07%
Number of People: 51,278
Records per Person: 2.51

Box Size: Prevalence, Color: Records per Person (Blue to Orange = Low to High), Use Ctrl-Click to Zoom, Alt-Click to Reset Zoom

Coronary arteriosclerosis

Condition Prevalence


ATLAS
Home
Data Sources
Vocabulary
Concept Sets
Cohorts
Profiles
Jobs
Configuration
Feedback
Data Sources
Data Sources
Reports
Demo_data_1_percent_synthetic_patients
Data Density
Total Rows
# of Records
400k
300k
200k
100k
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
Year
Dashboard
Achilles Heel
Person
Observation Periods
Data Density
Conditions
Condition Eras
Measurement
Observations
Drug Eras
Drug Exposures
Procedures
Visits
Death
Records Per Person
Records Per Person
150
100
50
Condition occurrence
Death
Drug era
Drug exposure
Observation
Observation period
Procedure occurrence
Visit occurrence



Demo_data_1_percent_synthetic_patients

Achilles Heel Report

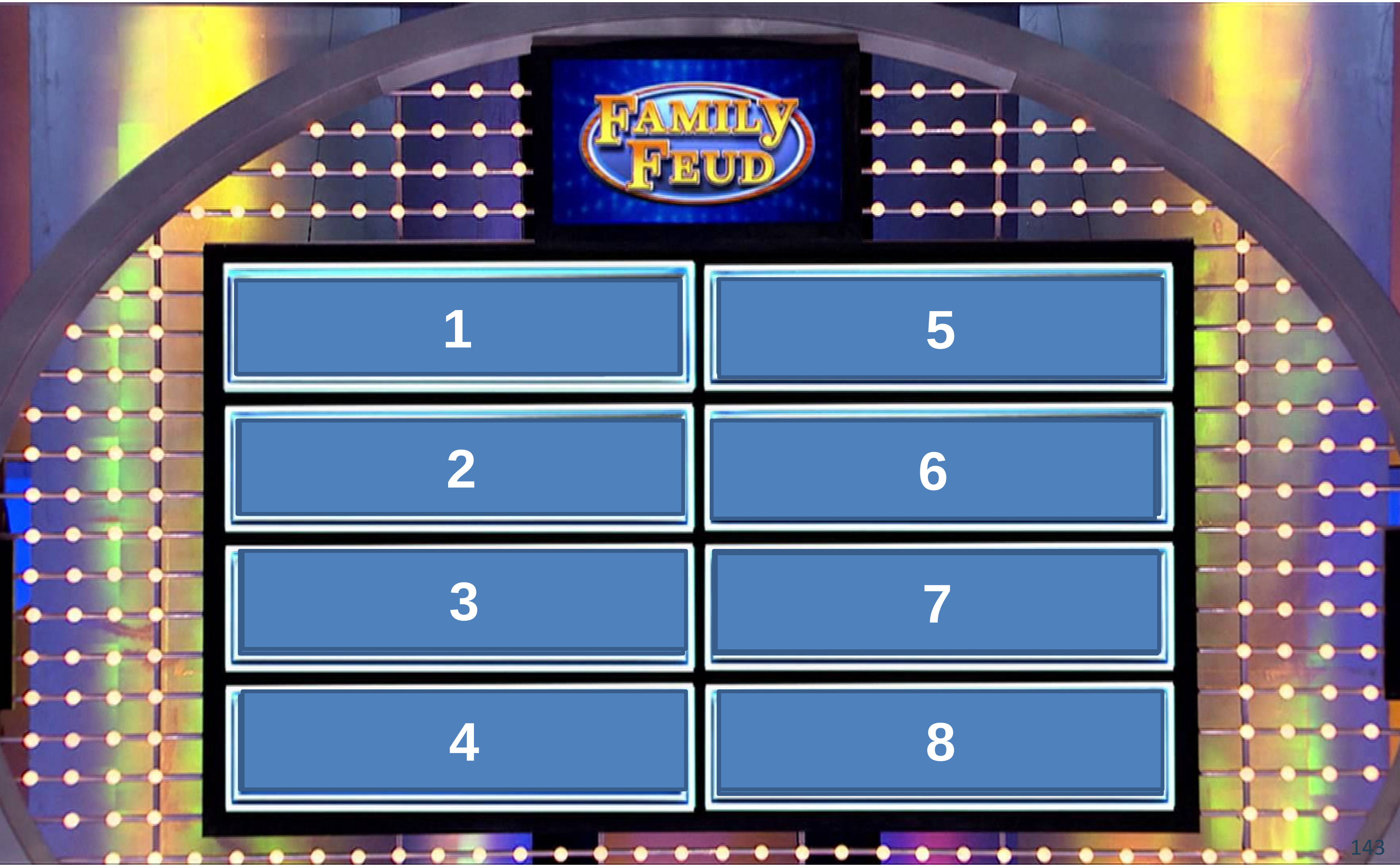
Data Quality Messages

Message Type	Message
ERROR	400-Number of persons with at least one condition occurrence, by condi concepts in data are not in correct vocabulary
ERROR	600-Number of persons with at least one procedure occurrence, by proc concepts in data are not in correct vocabulary
ERROR	900-Number of persons with at least one drug era, by drug_concept_id; in vocabulary
ERROR	908-Number of drug eras without valid person; count (n=23,452,537) sh
ERROR	909-Number of drug eras outside valid observation period; count (n=23,475,293) should not be > 0
NOTIFICATION	Unmapped data over percentage threshold in:Condition
NOTIFICATION	Unmapped data over percentage threshold in:Procedure
NOTIFICATION	Unmapped data over percentage threshold in:DrugExposure
NOTIFICATION	Unmapped data over percentage threshold in:Observation
NOTIFICATION	Unmapped data over percentage threshold in:Measurement

ETL Pitfalls

- Privacy Issues
  - Removal of ICD9/10 codes that are considered privacy issues, such as death or sexual abuse
  - Using “fake” date in Death table to indicate a death
- Patient Cleansing
  - Test patients
- Differing Business Rules
  - Institutions decide not to follow vocabulary classifications

Conclusion



1

2

3

4

5

6

7

8

Join the journey

Interested in OHDSI?

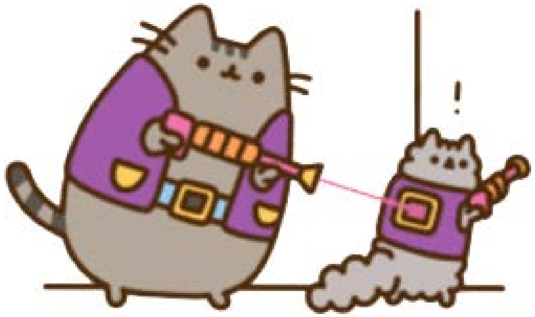
Join the Journey:

http://www.ohdsi.org/join-the-journey/

Questions:

http://forums.ohdsi.org/

APPENDIX



Lasers

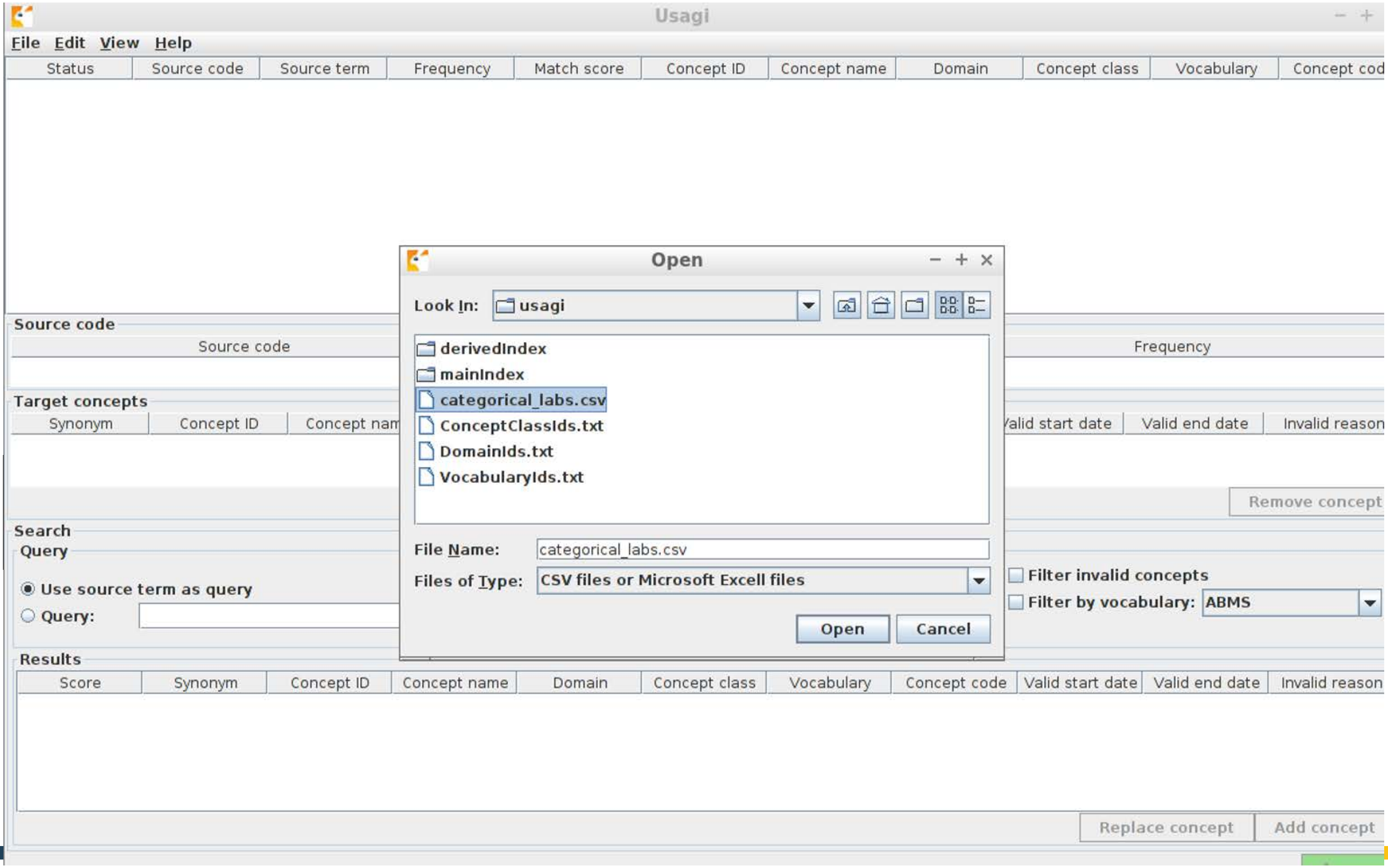
USAGI

- Tool to help in mapping codes from a source system into the standard terminologies stored in OMOP Vocabulary

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

USAGI Exercise

Sample File



Look In: usagi

- derivedIndex
- mainIndex
- categorical_labs.csv
- ConceptClassIds.txt
- DomainIds.txt
- VocabularyIds.txt

File Name: categorical_labs.csv

Files of Type: CSV files or Microsoft Excell files

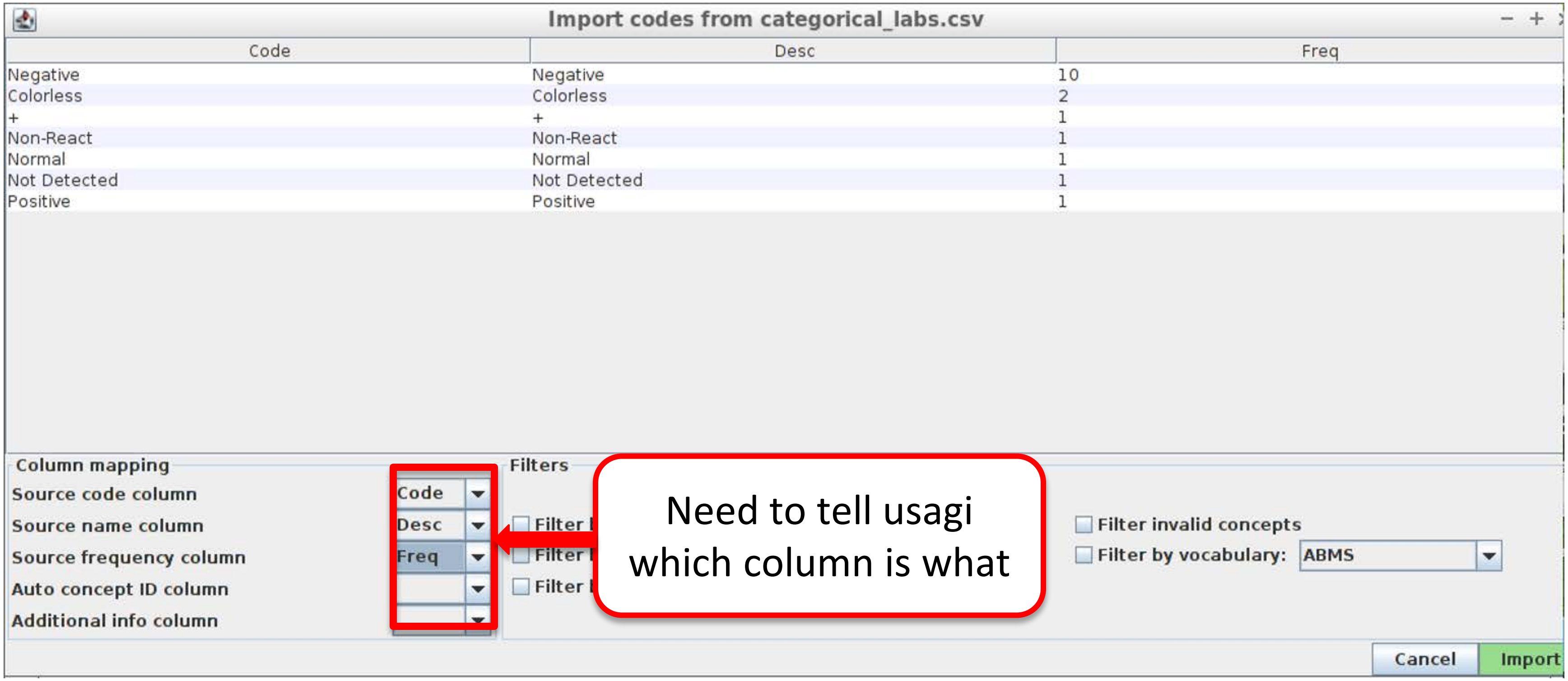
● Use source term as query

○ Query:

□ Filter invalid concepts

□ Filter by vocabulary: ABMS

USAGI Exercise



Import codes from categorical_labs.csv

Code	Desc	Freq
Negative	Negative	10
Colorless	Colorless	2
+	+	1
Non-React	Non-React	1
Normal	Normal	1
Not Detected	Not Detected	1
Positive	Positive	1

Column mapping

Source code column: Code

Source name column: Desc

Source frequency column: Freq

Auto concept ID column:

Additional info column:

Filters

□ Filter

□ Filter

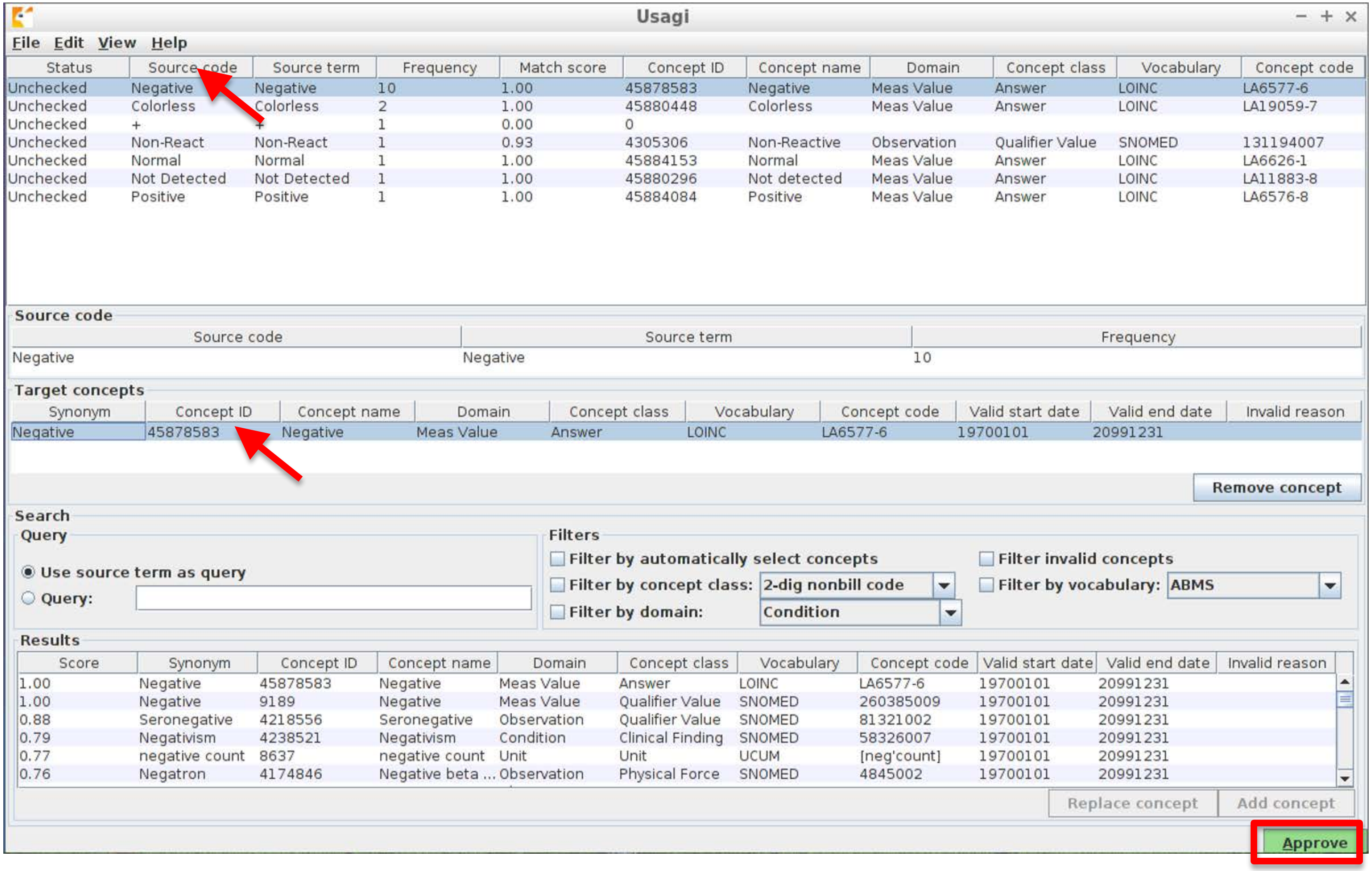
□ Filter

□ Filter invalid concepts

□ Filter by vocabulary: ABMS

Need to tell usagi which column is what

USAGI Exercise



Status	Source code	Source term	Frequency	Match score	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
Unchecked	Negative	Negative	10	1.00	45878583	Negative	Meas Value	Answer	LOINC	LA6577-6
Unchecked	Colorless	Colorless	2	1.00	45880448	Colorless	Meas Value	Answer	LOINC	LA19059-7
Unchecked	+	+	1	0.00	0					
Unchecked	Non-React	Non-React	1	0.93	4305306	Non-Reactive	Observation	Qualifier Value	SNOMED	131194007
Unchecked	Normal	Normal	1	1.00	45884153	Normal	Meas Value	Answer	LOINC	LA6626-1
Unchecked	Not Detected	Not Detected	1	1.00	45880296	Not detected	Meas Value	Answer	LOINC	LA11883-8
Unchecked	Positive	Positive	1	1.00	45884084	Positive	Meas Value	Answer	LOINC	LA6576-8

Source code

Source code	Source term	Frequency
Negative	Negative	10

Target concepts

Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
Negative	45878583	Negative	Meas Value	Answer	LOINC	LA6577-6	19700101	20991231	

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by automatically select concepts

□ Filter by concept class: 2-dig nonbill code

□ Filter by domain: Condition

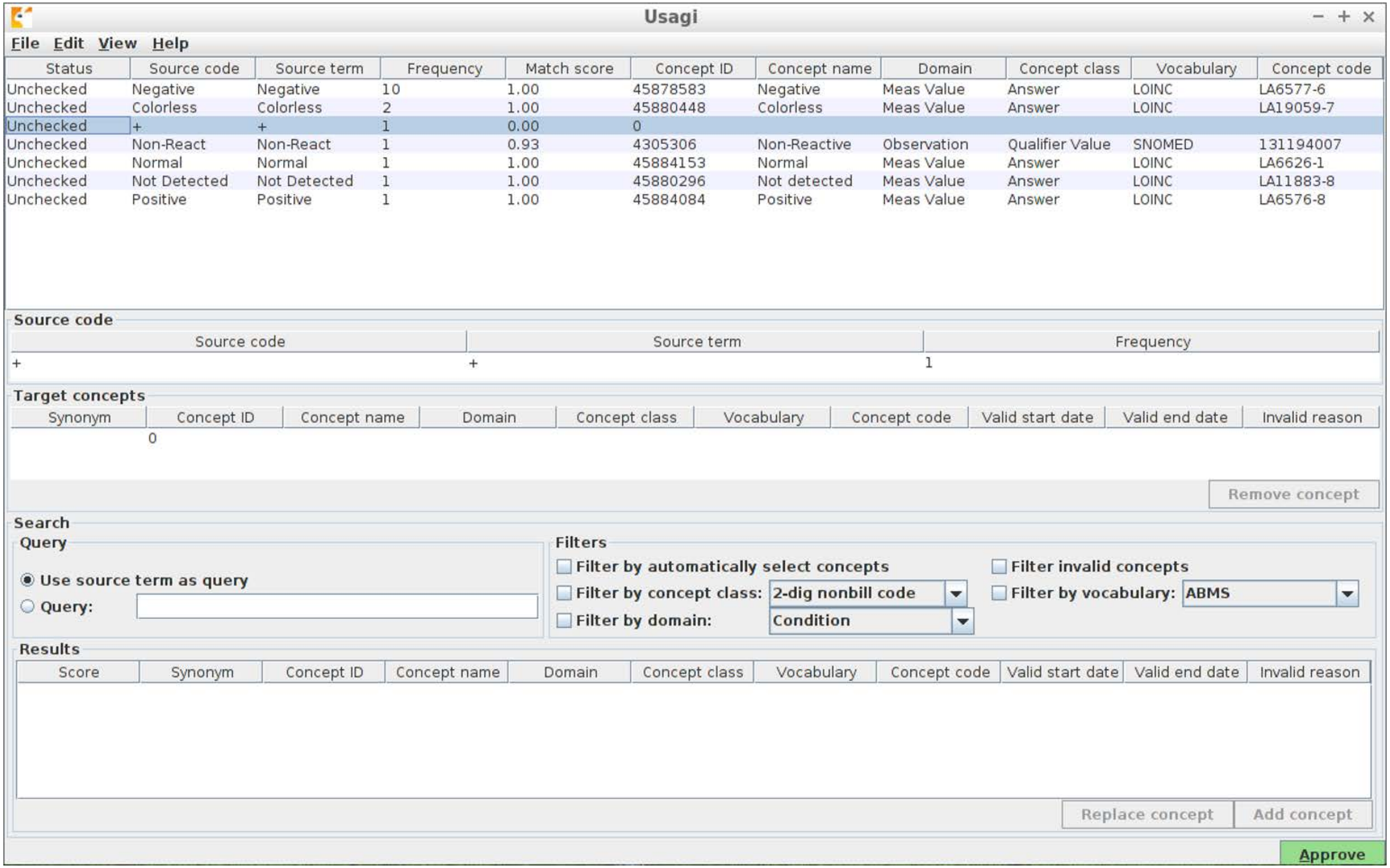
□ Filter invalid concepts

□ Filter by vocabulary: ABMS

Results

Score	Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
1.00	Negative	45878583	Negative	Meas Value	Answer	LOINC	LA6577-6	19700101	20991231	
1.00	Negative	9189	Negative	Meas Value	Qualifier Value	SNOMED	260385009	19700101	20991231	
0.88	Seronegative	4218556	Seronegative	Observation	Qualifier Value	SNOMED	81321002	19700101	20991231	
0.79	Negativism	4238521	Negativism	Condition	Clinical Finding	SNOMED	58326007	19700101	20991231	
0.77	negative count	8637	negative count	Unit	Unit	UCUM	[neg'count]	19700101	20991231	
0.76	Negatron	4174846	Negative beta ...	Observation	Physical Force	SNOMED	4845002	19700101	20991231	

USAGI Exercise



Status	Source code	Source term	Frequency	Match score	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
Unchecked	Negative	Negative	10	1.00	45878583	Negative	Meas Value	Answer	LOINC	LA6577-6
Unchecked	Colorless	Colorless	2	1.00	45880448	Colorless	Meas Value	Answer	LOINC	LA19059-7
Unchecked	+	+	1	0.00	0					
Unchecked	Non-React	Non-React	1	0.93	4305306	Non-Reactive	Observation	Qualifier Value	SNOMED	131194007
Unchecked	Normal	Normal	1	1.00	45884153	Normal	Meas Value	Answer	LOINC	LA6626-1
Unchecked	Not Detected	Not Detected	1	1.00	45880296	Not detected	Meas Value	Answer	LOINC	LA11883-8
Unchecked	Positive	Positive	1	1.00	45884084	Positive	Meas Value	Answer	LOINC	LA6576-8

Source code

Source code	Source term	Frequency
+	+	1

Target concepts

Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
	0								

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by automatically select concepts

□ Filter by concept class: 2-dig nonbill code

□ Filter by domain: Condition

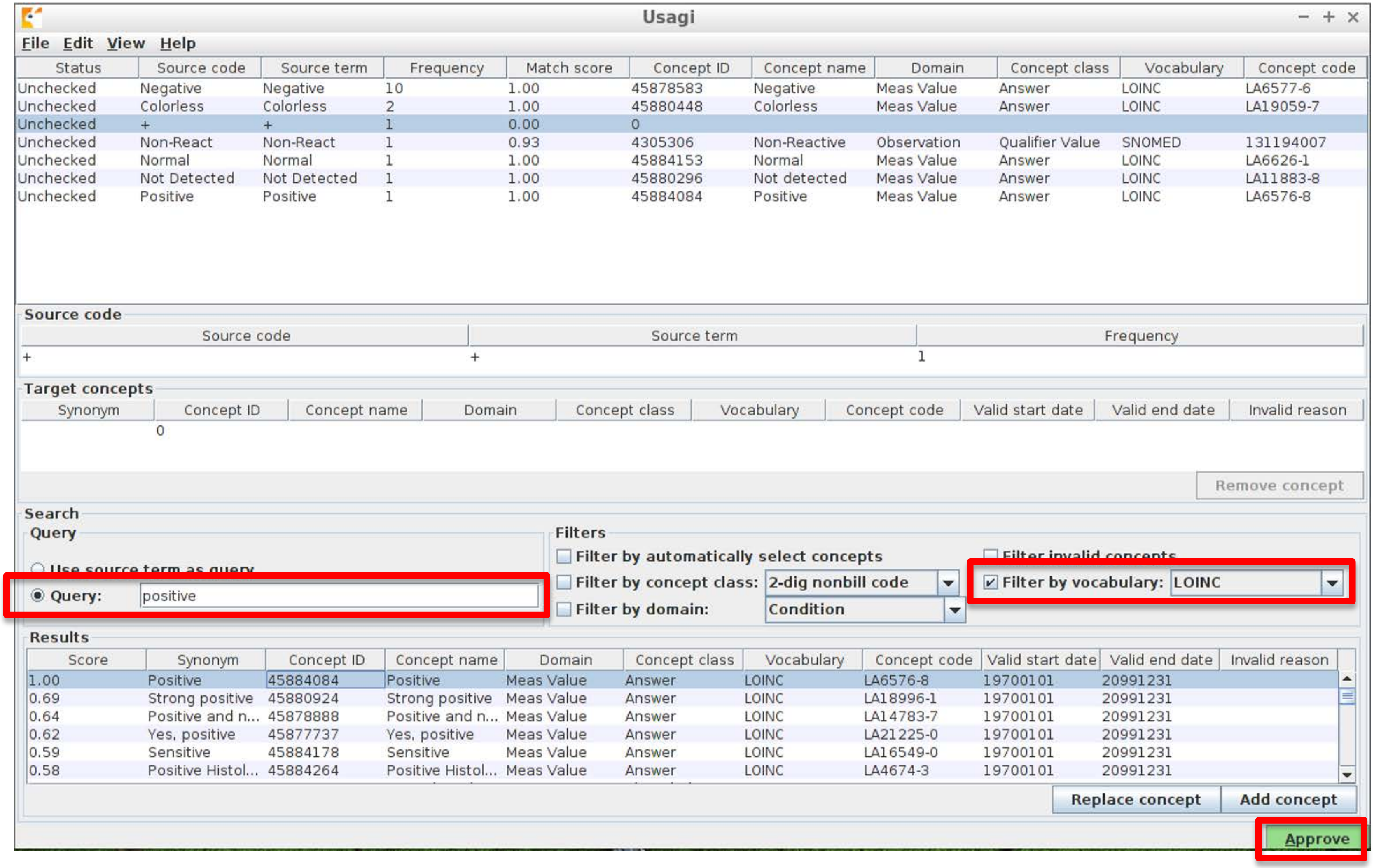
□ Filter invalid concepts

□ Filter by vocabulary: ABMS

Results

Score	Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason

USAGI Exercise



Status	Source code	Source term	Frequency	Match score	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code
Unchecked	Negative	Negative	10	1.00	45878583	Negative	Meas Value	Answer	LOINC	LA6577-6
Unchecked	Colorless	Colorless	2	1.00	45880448	Colorless	Meas Value	Answer	LOINC	LA19059-7
Unchecked	+	+	1	0.00	0					
Unchecked	Non-React	Non-React	1	0.93	4305306	Non-Reactive	Observation	Qualifier Value	SNOMED	131194007
Unchecked	Normal	Normal	1	1.00	45884153	Normal	Meas Value	Answer	LOINC	LA6626-1
Unchecked	Not Detected	Not Detected	1	1.00	45880296	Not detected	Meas Value	Answer	LOINC	LA11883-8
Unchecked	Positive	Positive	1	1.00	45884084	Positive	Meas Value	Answer	LOINC	LA6576-8

Source code

Source code	Source term	Frequency
+	+	1

Target concepts

Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
	0								

Search

Query

○ Use source term as query

● Query: positive

Filters

□ Filter by automatically select concepts

□ Filter invalid concepts

□ Filter by concept class: 2-dig nonbill code

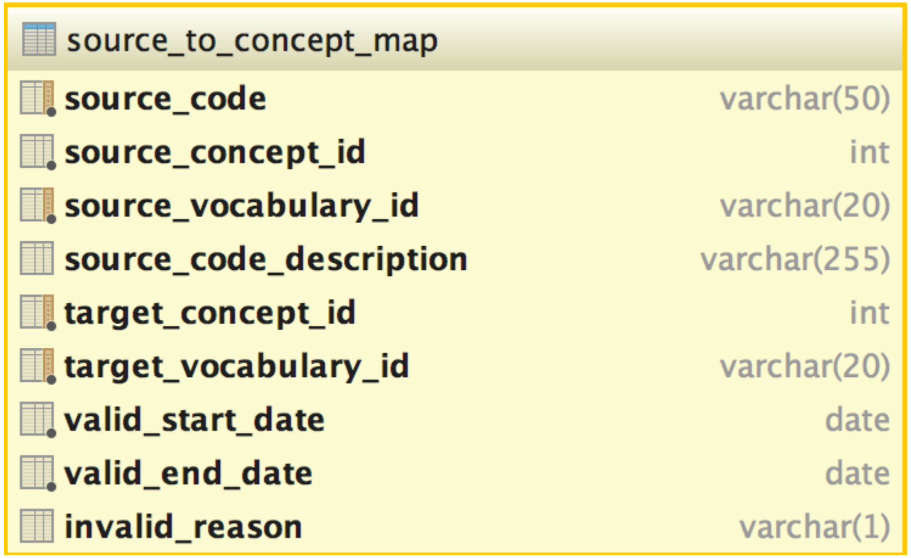
☑ Filter by vocabulary: LOINC

□ Filter by domain: Condition

Results

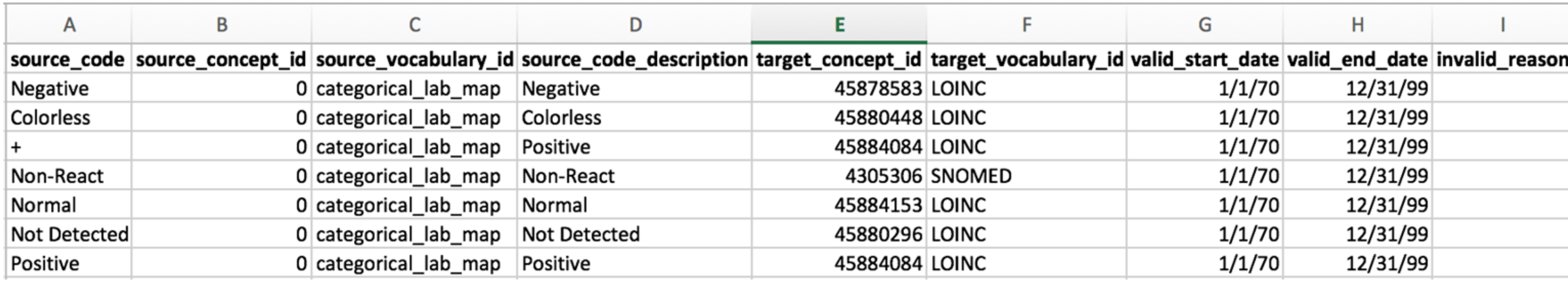
Score	Synonym	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Valid start date	Valid end date	Invalid reason
1.00	Positive	45884084	Positive	Meas Value	Answer	LOINC	LA6576-8	19700101	20991231	
0.69	Strong positive	45880924	Strong positive	Meas Value	Answer	LOINC	LA18996-1	19700101	20991231	
0.64	Positive and n...	45878888	Positive and n...	Meas Value	Answer	LOINC	LA14783-7	19700101	20991231	
0.62	Yes, positive	45877737	Yes, positive	Meas Value	Answer	LOINC	LA21225-0	19700101	20991231	
0.59	Sensitive	45884178	Sensitive	Meas Value	Answer	LOINC	LA16549-0	19700101	20991231	
0.58	Positive Histol...	45884264	Positive Histol...	Meas Value	Answer	LOINC	LA4674-3	19700101	20991231	

USAGI Exercise



source_to_concept_map

source_code	varchar(50)
source_concept_id	int
source_vocabulary_id	varchar(20)
source_code_description	varchar(255)
target_concept_id	int
target_vocabulary_id	varchar(20)
valid_start_date	date
valid_end_date	date
invalid_reason	varchar(1)



A	B	C	D	E	F	G	H	I
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
Negative	0	categorical_lab_map	Negative	45878583	LOINC	1/1/70	12/31/99	
Colorless	0	categorical_lab_map	Colorless	45880448	LOINC	1/1/70	12/31/99	
+	0	categorical_lab_map	Positive	45884084	LOINC	1/1/70	12/31/99	
Non-React	0	categorical_lab_map	Non-React	4305306	SNOMED	1/1/70	12/31/99	
Normal	0	categorical_lab_map	Normal	45884153	LOINC	1/1/70	12/31/99	
Not Detected	0	categorical_lab_map	Not Detected	45880296	LOINC	1/1/70	12/31/99	
Positive	0	categorical_lab_map	Positive	45884084	LOINC	1/1/70	12/31/99	