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

<table>
<thead>
<tr>
<th>Anthony Sena</th>
<th>Jungmi Han</th>
</tr>
</thead>
<tbody>
<tr>
<td>Janssen R&amp;D Epidemiology Analytics</td>
<td>Columbia University</td>
</tr>
</tbody>
</table>

![Anthony Sena](image1.png)  
![Jungmi Han](image2.png)
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.
<table>
<thead>
<tr>
<th>Time</th>
<th>Type</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM-8:15AM</td>
<td></td>
<td>Introductions</td>
</tr>
<tr>
<td>8:15AM-9:15AM</td>
<td>Foundational</td>
<td>What is OMOP/OHDSI? OMOP Common Data Model (CDM) – Why and How</td>
</tr>
<tr>
<td>9:15AM-10:00AM</td>
<td></td>
<td>How to retrieve data from OMOP CDM</td>
</tr>
<tr>
<td>10:00AM-10:15AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>10:15AM-10:45AM</td>
<td>Implementation</td>
<td>Setup and Performing of an Extract Transform and Load process into the CDM</td>
</tr>
<tr>
<td>10:45AM-11:30AM</td>
<td></td>
<td>Using WhiteRabbit and Rabbit-In-A-Hat to Build an ETL</td>
</tr>
<tr>
<td>11:30AM-11:45AM</td>
<td>Evaluation</td>
<td>Testing and Quality Assurance</td>
</tr>
<tr>
<td>11:45AM-12:00PM</td>
<td></td>
<td>Wrap up</td>
</tr>
</tbody>
</table>
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 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
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

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCEPT_ID</td>
<td>313217</td>
</tr>
<tr>
<td>CONCEPT_NAME</td>
<td>Atrial fibrillation</td>
</tr>
<tr>
<td>DOMAIN_ID</td>
<td>Condition</td>
</tr>
<tr>
<td>VOCABULARY_ID</td>
<td>SNOMED</td>
</tr>
<tr>
<td>CONCEPT_CLASS_ID</td>
<td>Clinical Finding</td>
</tr>
<tr>
<td>STANDARD_CONCEPT</td>
<td>S</td>
</tr>
<tr>
<td>CONCEPT_CODE</td>
<td>49436004</td>
</tr>
<tr>
<td>VALID_START_DATE</td>
<td>01-Jan-70</td>
</tr>
<tr>
<td>VALID_END_DATE</td>
<td>31-Dec-99</td>
</tr>
</tbody>
</table>

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

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

- 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

- Contains one record per each unique location
- Location is highly variable across sources, of limited use thus far
OBSERVATION_PERIOD

- 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

• Can have death without cause

• Can only have 1 death per person
VISIT_OCCURRENCE

• 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

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

- Procedures have the least standardized vocabularies that causes some redundancy
CONDITION_OCCURRENCE

• Vocabulary: SNOMED -> classification

• Data sources:
  – Billing diagnosis (inpatient, outpatient)
  – Problem list

• Individual records <> distinct episodes
DRUG_EXPOSURE

- 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

- 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

• 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

- 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

• To capture of biomarker / tissue bank
NOTE

• To capture unstructured free text

• Coming soon in CDM 5.x: NLP and LOINC Clinical Document Ontology (CDO) annotations
Health Economics

- 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
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_DEFINITION** 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

- 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

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 handle combination products?

How do we infer discontinuation?

How do we handle gap?

NDC: 00179198801
Lisinopril 5 MG Oral Tablet

NDC: 00310013010
ZESTRIL 5 MG TABLET

NDC: 0038013210
Lisinopril 20 MG Oral Tablet [Zestril]

NDC: 0038013134
Lisinopril 10 MG Oral Tablet [Zestril]

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

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

VirtualBox

PostgreSQL

PostgreSQL

pgAdmin

cdm

webapi

Broadsea

WebTools

Atlas
Penelope
Calypso

WebAPI

Tomcat

Methods Library

OHDSI R packages

Studio

synpuf_100k

WhiteRabbit

RabbitInAHat

OHDSI R packages

Studio
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 (64-bit)
- **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**.

![VirtualBox Memory Configuration](image)

- **Do not add a virtual hard disk**
- **Create a virtual hard disk now**
- **Use an existing virtual hard disk file**

![VirtualBox Hard Disk Configuration](image)
OHDSI in a Box – Start Up
OhDSI in a Box – International Keyboards

[Image of a computer interface with various menus and settings for input methods and preferences.]

- Keyboard Input Methods (IBus Daemon) is not running. Do you wish to start it?
  - No
  - Yes

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

- IBus Preferences
  - General
    - Input Method
      - English - English (US)
      - German - German
  - Advanced
    - Use system keyboard layout
    - Share the same input method among all applications
OHDSI in a Box – Adjust Resolution
OHDSI in a Box – Clipboard
OHDSI in a Box – Timeout
OHDSI in a Box – Ready
CDM Database – pgAdmin III New Server

![New Server Registration](image)

<table>
<thead>
<tr>
<th>Properties</th>
<th>SSL</th>
<th>SSH Tunnel</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Host</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Port</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance DB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Username</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Password</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Store password</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Name: ohdsi
- Host: localhost
- Port: 5432
- Maintenance DB: postgres
- Username: ohdsi
- Password: ohdsi
- Colour: [ ]
- Group: Servers
CDM Database – Connect
CDM Database – Open SQL Sheet

Execute arbitrary SQL queries.

--- 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

```sql
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

Data Used for Demonstration

• Five types of data:

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

SynPUF High Level Diagram

Beneficiary Summary

- Inpatient Claims
- Outpatient Claims
- Carrier Claims
- Prescription Drug Events (PDE)
Mapping SynPUF to CDM

SynPUF

Beneficiary Summary
- Inpatient Claims
- Outpatient Claims
- Carrier Claims
- Prescription Drug Events (PDE)

CDM

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
- Payer_plan_period
- Cost

Standardized derived elements
- Cohort
- Cohort_attribute
- Condition_era
- Drug_era
- Dose_era

Standardized health economics
Some Example Questions

New Users of Warfarin who are >=65?

New Users of Warfarin with prior Atrial Fibrillation?
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?

Drug exposures capture records about the utilization of a drug when ingested or otherwise introduced into the body.
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
How do I define Warfarin?

• Writing SQL Statement

• OHDSI Tool ATLAS
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
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
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
New Users of Warfarin

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
Step 1: Get the codes you need

```sql
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
```
Step 2: Find Drug Exposures

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
Step 3: Find 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
New Users of Warfarin

Ex 1

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
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
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
Step 1: Start with the previous query

```sql
/* (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,
    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
```
Step 2: Add the Person Table to calculate age

```sql
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
```
New Users of Warfarin <= 65 years of age

```sql
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

/* (Exercise 2) Warfarin New Users 65 or Older at Index */
```
How do I define new users of Warfarin with prior Atrial Fibrillation?
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
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

)nu.*,

SELECT 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

78
Step 2: Define 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
### Step 3: Prior Atrial Fibrillation

Keeps condition within the same observable time, exclude if you want all time prior.
How do I define new users of Warfarin with prior Atrial Fibrillation?
New Users of Warfarin with prior 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 *, 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 coCONDITION_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.

Appendix

<table>
<thead>
<tr>
<th>Available CDM Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Name</strong></td>
</tr>
<tr>
<td>OHDSI CDM V5 Database</td>
</tr>
</tbody>
</table>

**Distinct People**

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?

Extract

Extract Load

source_1

source_2

source_3

Transform

Load
cdm
ETL: Real world scenario

Truven MarketScan Commercial Claims and Encounters (CCAE)

<table>
<thead>
<tr>
<th>enrolid</th>
<th>admdate</th>
<th>pdx</th>
<th>dx1</th>
<th>dx2</th>
<th>dx3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1570337021</td>
<td>5/31/2000</td>
<td>41071</td>
<td>41071</td>
<td>4241</td>
<td>V5881</td>
</tr>
</tbody>
</table>

Optum Extended SES

<table>
<thead>
<tr>
<th>patid</th>
<th>fst_dt</th>
<th>diag1</th>
<th>diag2</th>
<th>diag3</th>
<th>diag4</th>
</tr>
</thead>
<tbody>
<tr>
<td>259000476532</td>
<td>5/30/2000</td>
<td>41071</td>
<td>27800</td>
<td>4019</td>
<td>2724</td>
</tr>
</tbody>
</table>

Premier

<table>
<thead>
<tr>
<th>pat_key</th>
<th>period</th>
<th>icd_code</th>
<th>icd_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>-17197140</td>
<td>1/1/2000</td>
<td>410.71</td>
<td>P</td>
</tr>
<tr>
<td>-17197140</td>
<td>1/1/2000</td>
<td>414.01</td>
<td>S</td>
</tr>
<tr>
<td>-17197140</td>
<td>1/1/2000</td>
<td>427.31</td>
<td>S</td>
</tr>
<tr>
<td>-17197140</td>
<td>1/1/2000</td>
<td>496</td>
<td>S</td>
</tr>
</tbody>
</table>

Japan Medical Data Center

<table>
<thead>
<tr>
<th>member_id</th>
<th>admission_date</th>
<th>icd10_level4_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0041437</td>
<td>4/11/2013</td>
<td>I214</td>
</tr>
<tr>
<td>M0041437</td>
<td>4/11/2013</td>
<td>A539</td>
</tr>
<tr>
<td>M0041437</td>
<td>4/11/2013</td>
<td>B182</td>
</tr>
<tr>
<td>M0041437</td>
<td>4/11/2013</td>
<td>E14-</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>enrolid</th>
<th>admdate</th>
<th>pdx</th>
<th>dx1</th>
<th>dx2</th>
<th>dx3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1570337021</td>
<td>5/31/2000</td>
<td>41071</td>
<td>41071</td>
<td>4241</td>
<td>V5881</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>157033702</td>
<td>5/31/2000</td>
<td>41071</td>
<td>Inpatient claims - primary position</td>
</tr>
<tr>
<td>157033702</td>
<td>5/31/2000</td>
<td>41071</td>
<td>Inpatient claims - 1st position</td>
</tr>
<tr>
<td>157033702</td>
<td>5/31/2000</td>
<td>4241</td>
<td>Inpatient claims - 2nd position</td>
</tr>
<tr>
<td>157033702</td>
<td>5/31/2000</td>
<td>V5881</td>
<td>Inpatient claims - 3rd position</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
<th>CONDITION_SOURCE_CONCEPT_ID</th>
<th>CONDITION_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>157033702</td>
<td>5/31/2000</td>
<td>41071</td>
<td>Inpatient claims - primary position</td>
<td>44825429</td>
<td>444406</td>
</tr>
</tbody>
</table>
OMOP CDM = Standardized structure: same tables, same fields, same datatypes, same conventions across disparate sources

• Consistent structure optimized for large-scale analysis
• Structure preserves all source content and provenance
OMOP CDM = Standardized content: common vocabularies across disparate sources

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

<table>
<thead>
<tr>
<th>Truven CCAE: CONDITION_OCCURRENCE</th>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
<th>CONDITION_SOURCE_CONCEPT_ID</th>
<th>CONDITION_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>157033702</td>
<td>5/31/2000</td>
<td>41071</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inpatient claims - primary position</td>
<td></td>
<td>44825429</td>
<td>444406</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optum Extended SES: CONDITION_OCCURRENCE</th>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
<th>CONDITION_SOURCE_CONCEPT_ID</th>
<th>CONDITION_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>259000474406532</td>
<td>5/30/2000</td>
<td>41071</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inpatient claims - 1st position</td>
<td></td>
<td>44825429</td>
<td>444406</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Premier: CONDITION_OCCURRENCE</th>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
<th>CONDITION_SOURCE_CONCEPT_ID</th>
<th>CONDITION_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-171971409</td>
<td>1/1/2000</td>
<td>410.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hospital record - primary</td>
<td></td>
<td>44825429</td>
<td>444406</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>JMDC: CONDITION_OCCURRENCE</th>
<th>PERSON_ID</th>
<th>CONDITION_START_DATE</th>
<th>CONDITION_SOURCE_VALUE</th>
<th>CONDITION_TYPE_CONCEPT_ID</th>
<th>CONDITION_SOURCE_CONCEPT_ID</th>
<th>CONDITION_CONCEPT_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4149337</td>
<td>4/11/2013</td>
<td>1214</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inpatient claims</td>
<td></td>
<td>45572081</td>
<td>444406</td>
</tr>
</tbody>
</table>
ETL Process: Roles

• Members of the team
  – CDM expert
  – Local data expert
  – Data engineer
  – Person with medical knowledge
  – Business stakeholder
ETL Process: Agile

Setup

Sprint 0

Sprint 1

Sprint 2

Sprint 3

Sprint 4

Release

During Sprint

ETL Specs → Develop
Evaluation → Execute

Post Sprint

Business Validation/Sign-off → Issues/Corrections
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
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence

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

Sprint 4
- Finalize ETL Specs

Development – Implementation/Validation of ETL specs

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

Atlas Data Sources (Achilles)
Best Practices Documented

Implementation

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

- [https://github.com/OHDSI/WhiteRabbit](https://github.com/OHDSI/WhiteRabbit)
- Click on “releases”
- “Latest Release” and download the WhiteRabbit zip file
Getting WhiteRabbit

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

[Image of software interface showing the working folder path set to H:\2016\WhiteRabbit_v0.5.14\SynPuf DE 1.0 Sample 1]
Specify the Location of Data
Scanning your Data
Scanning your Data

![White Rabbit interface for scanning data with options for tables, cell counts, distinct values, and rows per table.]
Scanning your Data
Run the Scan Report on Your Data!

- Link on desktop

- Execute

- 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”.
• 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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Table</td>
<td>Field</td>
<td>Type</td>
<td>Max length</td>
<td>N rows</td>
<td>N rows ch</td>
<td>Fraction empty</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Beneficiary_Summary.csv</td>
<td>DESYNPUF</td>
<td>varchar</td>
<td>16</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_BIR</td>
<td>int</td>
<td>8</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_DEA</td>
<td>int</td>
<td>8</td>
<td>-1</td>
<td>99999</td>
<td>0.98449</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_SEX</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_RAC</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_ESR</td>
<td>varchar</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Beneficiary_Summary.csv</td>
<td>SP_STATE</td>
<td>int</td>
<td>2</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE.COL</td>
<td>int</td>
<td>3</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE_HI.</td>
<td>int</td>
<td>2</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE.SM.</td>
<td>int</td>
<td>2</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Beneficiary_Summary.csv</td>
<td>BENE.HM.</td>
<td>int</td>
<td>2</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Beneficiary_Summary.csv</td>
<td>PLAN.CVF</td>
<td>int</td>
<td>2</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.ALZHD.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.CHF.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.CHRN.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.CNCR.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.COPD.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.DEPRE.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.DIABE.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Beneficiary_Summary.csv</td>
<td>SP.ISCHM.</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>99999</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Table Tabs

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

---

<table>
<thead>
<tr>
<th>#</th>
<th>Variable names</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DESYNPUF_ID</td>
<td>DESYNPUF: Beneficiary Code</td>
</tr>
<tr>
<td>2</td>
<td>BENE_BIRTH_DT</td>
<td>DESYNPUF: Date of birth</td>
</tr>
<tr>
<td>3</td>
<td>BENE_DEATH_DT</td>
<td>DESYNPUF: Date of death</td>
</tr>
<tr>
<td>4</td>
<td>BENE_SEX_IDENT_CD</td>
<td>DESYNPUF: Sex</td>
</tr>
<tr>
<td>5</td>
<td>BENE_RACE_CD</td>
<td>DESYNPUF: Beneficiary Race Code</td>
</tr>
</tbody>
</table>

### Variable Name: BENE_BIRTH_DT
- Type: Num
- Format: YYYYMMDD

---

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DESYNPUF_ID</td>
<td>Frequency</td>
<td>BENE_BIRTH_DT</td>
<td>Frequency</td>
<td>BENE_DEATH_DT</td>
<td>Frequency</td>
<td>BENE_SEX_IDENT_CD</td>
</tr>
<tr>
<td>1</td>
<td>List truncated...</td>
<td>19421001</td>
<td>466</td>
<td>98448</td>
<td>2</td>
<td>55211</td>
<td>44788</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>19410401</td>
<td>434</td>
<td>20080901</td>
<td>151</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>19390401</td>
<td>433</td>
<td>20080101</td>
<td>142</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>19431101</td>
<td>433</td>
<td>20081001</td>
<td>137</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>19400301</td>
<td>428</td>
<td>20080301</td>
<td>135</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>19410501</td>
<td>426</td>
<td>20081101</td>
<td>134</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>19390301</td>
<td>414</td>
<td>20080401</td>
<td>131</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>19400501</td>
<td>414</td>
<td>20080701</td>
<td>129</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>19410801</td>
<td>414</td>
<td>20080501</td>
<td>125</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>19391201</td>
<td>413</td>
<td>20080801</td>
<td>120</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>19431001</td>
<td>413</td>
<td>20081201</td>
<td>119</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>19411001</td>
<td>412</td>
<td>20080201</td>
<td>118</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>19411100</td>
<td>412</td>
<td>20080601</td>
<td>110</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

---

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

786.05 - Shortness of breath
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

Source
- beneficiary_summary.csv
- carrier_claims.csv
- inpatient_claims.csv
- outpatient_claims.csv
- prescription_drug.csv

CDMV5.0.1
- location
- person
- observation_period
- care_site
- visit_occurrence
- provider
- condition_occurrence
Map Raw Tables to CDM Tables

Source:
- beneficiary_summary.csv
- carrier_claims.csv
- inpatient_claims.csv
- outpatient_claims.csv
- prescription_drug.csv

CDMV5.0.1:
- location
- person
- drug_exposure
- observation_period
- care_site
- visit_occurrence
- provider
Prescription Drug Events contain records pertaining to Part D synthetic events. Medicare Part D are prescription drug benefits.
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

cannot forget carrier claims, inpatient claims, and outpatient claims as well. they contain procedure codes like hcpcs that may convert into drugs as well.
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.
• For today’s example we’ll start with the PERSON table
<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Logic</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>person_id</td>
<td></td>
<td></td>
<td>Autonumber</td>
</tr>
<tr>
<td>gender_concept_id</td>
<td>bene_sex_ident_cd</td>
<td>Source Value - Standard Concept Id 1 - 8507 2 - 8532 If gender is not 1 or 2, please discard person.</td>
<td>1-Male 2-Female</td>
</tr>
<tr>
<td>year_of_birth</td>
<td>bene_birth_dt</td>
<td>Take first 4 digits (starting from left)</td>
<td>BENE_BIRTH_DT = YYYYMMDD</td>
</tr>
<tr>
<td>month_of_birth</td>
<td>bene_birth_dt</td>
<td>Take 5th and 6th digit starting from the left</td>
<td>BENE_BIRTH_DT = YYYYMMDD</td>
</tr>
<tr>
<td>day_of_birth</td>
<td>bene_birth_dt</td>
<td>Take last two digits starting from the left</td>
<td>BENE_BIRTH_DT = YYYYMMDD</td>
</tr>
<tr>
<td>time_of_birth</td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>race_concept_id</td>
<td>bene_race_cd</td>
<td>Source Value - Concept ID 1 - 8527 2 - 8516 3 - 0 5 - 0 Else set to 0.</td>
<td>1-White 2-Black 3-Others 5-Hispanic</td>
</tr>
<tr>
<td>ethnicity_concept_id</td>
<td>bene_race_cd</td>
<td>Source Value - Concept ID 1 - 38003564 2 - 38003564 3 - 0 5 - 38003563 Else set to 0.</td>
<td>1-White 2-Black 3-Others 5-Hispanic</td>
</tr>
<tr>
<td>location_id</td>
<td>sp_state_code</td>
<td>Use the BENE_COUNTY_CD and SP_STATE_CODE to lookup in the LOCATION table the LOCATION_ID.</td>
<td></td>
</tr>
<tr>
<td>provider_id</td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>care_site_id</td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>person_source_value</td>
<td>desynpuf_id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender_source_value</td>
<td>bene_sex_ident_cd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender_source_concept_id</td>
<td></td>
<td>Set to 0.</td>
<td></td>
</tr>
<tr>
<td>race_source_value</td>
<td>bene_race_cd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>race_source_concept_id</td>
<td></td>
<td>Set to 0.</td>
<td></td>
</tr>
<tr>
<td>ethnicity_source_value</td>
<td>bene_race_cd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnicity_source_concept_id</td>
<td></td>
<td>Set to 0.</td>
<td></td>
</tr>
</tbody>
</table>
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
• 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”

46275076 - “rituximab Injection”
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_VOCABULARY_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')
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.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 AS 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**

```sql
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**

```sql
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

• Export to document
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
Demo_data_1_percent_synthetic_patients

Dashboard

CDM Summary
Source name: synpuf_1percent
Number of 116.35k persons:

Population by Gender

Age at First Observation

Cumulative Observation

Persons With Continuous Observation By Month
Source name: synpuf_1percent
Number of 116.35k persons:
Demo_data_1_percent_synthetic_patients

Conditions

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

Age Decile

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Message Type</td>
<td>Message</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>400-Number of persons with at least one condition occurrence, by condition concepts in data are not in correct vocabulary</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>600-Number of persons with at least one procedure occurrence, by procedure concepts in data are not in correct vocabulary</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>900-Number of persons with at least one drug era, by drug_concept_id; in vocabulary</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>908-Number of drug eras without valid person; count (n=23,452,537) should not be &gt; 0</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>909-Number of drug eras outside valid observation period; count (n=23,475,293) should not be &gt; 0</td>
<td></td>
</tr>
<tr>
<td>NOTIFICATION</td>
<td>Unmapped data over percentage threshold in:Condition</td>
<td></td>
</tr>
<tr>
<td>NOTIFICATION</td>
<td>Unmapped data over percentage threshold in:Procedure</td>
<td></td>
</tr>
<tr>
<td>NOTIFICATION</td>
<td>Unmapped data over percentage threshold in:DrugExposure</td>
<td></td>
</tr>
<tr>
<td>NOTIFICATION</td>
<td>Unmapped data over percentage threshold in:Observation</td>
<td></td>
</tr>
<tr>
<td>NOTIFICATION</td>
<td>Unmapped data over percentage threshold in:Measurement</td>
<td></td>
</tr>
</tbody>
</table>
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

CDM standardizes the structure

CDM is a patient centric model

ETL Process is an Iterative One

White Rabbit your data to understand it

Plan your ETL in Rabbit in a Hat

OMOP Vocabulary used to standardize the terminology

OHDSI is an Open Source Collaborative Community

ACHILLES Helps you evaluate your data
Join the journey

Interested in OHDSI?

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

Questions:  
http://forums.ohdsi.org/
APPENDIX
USAGI

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

USAGI Exercise

Sample File
USAGI Exercise

Need to tell usagi which column is what
### USAGI Exercise

#### USAGI Table

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

#### Source code table

<table>
<thead>
<tr>
<th>Source code</th>
<th>Source term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

#### Target concepts

<table>
<thead>
<tr>
<th>Synonym</th>
<th>Concept ID</th>
<th>Concept name</th>
<th>Domain</th>
<th>Concept class</th>
<th>Vocabulary</th>
<th>Concept code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>45878583</td>
<td>Negative</td>
<td>Meas Value</td>
<td>Answer</td>
<td>LOINC</td>
<td>LA6577-6</td>
</tr>
</tbody>
</table>

#### Search

- **Query**: Use source term as query
- **Filters**
  - Filter by automatically select concepts
  - Filter by concept class: 2-dig nonbill code
  - Filter by concept: Z-digit nonbill code
  - Filter by vocabulary: ABMS
  - Filter by domain: Condition

#### Results

<table>
<thead>
<tr>
<th>Score</th>
<th>Synonym</th>
<th>Concept ID</th>
<th>Concept name</th>
<th>Domain</th>
<th>Concept class</th>
<th>Vocabulary</th>
<th>Concept code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Negative</td>
<td>45878583</td>
<td>Negative</td>
<td>Meas Value</td>
<td>Answer</td>
<td>LOINC</td>
<td>LA6577-6</td>
</tr>
<tr>
<td>1.00</td>
<td>Negative</td>
<td>9189</td>
<td>Negative</td>
<td>Meas Value</td>
<td>Qualifier Value</td>
<td>LOINC</td>
<td>LA19059-7</td>
</tr>
<tr>
<td>0.88</td>
<td>Seronegative</td>
<td>4218556</td>
<td>Seronegative</td>
<td>Observation</td>
<td>Qualifier Value</td>
<td>LOINC</td>
<td>SL19059-7</td>
</tr>
<tr>
<td>0.79</td>
<td>Negativism</td>
<td>4238921</td>
<td>Negativism</td>
<td>Condition</td>
<td>Clinical Finding</td>
<td>LOINC</td>
<td>LA6577-6</td>
</tr>
<tr>
<td>0.76</td>
<td>Negatron</td>
<td>4174846</td>
<td>Negative beta...</td>
<td>Observation</td>
<td>Physical Force</td>
<td>LOINC</td>
<td>LA6577-6</td>
</tr>
</tbody>
</table>

**Approve**
USAGI Exercise
### USAGI Exercise

![USAGI Exercise Image]
# USAGI Exercise

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>source_code</td>
<td>source_concept_id</td>
<td>source_vocabulary_id</td>
<td>source_code_description</td>
<td>target_concept_id</td>
<td>target_vocabulary_id</td>
<td>valid_start_date</td>
<td>valid_end_date</td>
<td>invalid_reason</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Negative</td>
<td>45878583</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>Colorless</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Colorless</td>
<td>45880448</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Positive</td>
<td>45884084</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>Non-React</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Non-React</td>
<td>4305306</td>
<td>SNOMED</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Normal</td>
<td>45884153</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>Not Detected</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Not Detected</td>
<td>45880296</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0</td>
<td>categorical_lab_map</td>
<td>Positive</td>
<td>45884084</td>
<td>LOINC</td>
<td>1/1/70</td>
<td>12/31/99</td>
<td></td>
</tr>
</tbody>
</table>