ETL Training
## Agenda

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

Seng Chan You (Chan), MD, PhD
Translational Research Assistant Professor
Department of Preventive Medicine, Yonsei University, College of Medicine

Selva Muthu Kumaran Sathappan
Data Analyst
Saw Swee Hock School of Public Health, National University of Singapore

Mui Van Zandt
Senior Director
OMOP Data Networks, IQVIA

Jing Li
Senior Data Scientist
OMOP Studies, IQVIA
Introduction to ETL
ETL

• Extract, Transform, Load

• In order to get from our native/raw data into the OMOP CDM we need to design and develop an ETL process.

• Goal in ETLing is to standardize the format and terminology.

• This tutorial will teach you:
  - Best practices around designing an ETL and CDM maintenance
  - How to program an ETL
ETL Process

Data experts & CDM experts together design the ETL

Medical experts create the code mappings

All are involved in quality control

A technical person implements the ETL

**Tools**
- White Rabbit
- Rabbit In a Hat
- Usagi
- Internal Quality Checks
- Achilles
- Data Quality Dashboard
- Jenkins
- Code Repository
ETL Process

[Image -8x444 to 131x543]
[Image 134x68 to 790x451]

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

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

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

Agile Methodology
What is Agile Scrum

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

https://www.cprime.com/resources/what-is-agile-what-is-scrum/
Benefits of Agile Scrum

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

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

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

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

[https://www.cprime.com/resources/what-is-agile-what-is-scrum/](https://www.cprime.com/resources/what-is-agile-what-is-scrum/)
Agile Scrum framework

- Sprint Planning Meeting
- Sprint
- Sprints every 24 hrs.
- Finished Work
- Daily Scrum
- Sprint Retrospective
- Backlog Grooming
- Product Owner
- Begin Next Iteration Every 2 weeks
- Team
- Scrum Master
- Every 24 hrs.
**Roles in Agile Scrum**

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

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

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

Analysis – Creation of ETL Specs/Stories

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

Sprint 1
- Drug Exposure

Sprint 2
- Condition Occurrence
- Procedure Occurrence

Sprint 3
- Observation
- Payer plan period
- Cost

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

Sprint 5
- Finalize ETL Specs

Development – Implementation/Validation of ETL Specs

Sprint 0
- Initial Data On-boarding

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

Sprint 2
- Drug Exposure

Sprint 3
- Condition Occurrence
- Procedure Occurrence

Sprint 4
- Observation
- Payer plan period
- Cost

Sprint 5
- Drug Era
- Condition Era
- Observation Period
- Visit Occurrence
What is Agile?
• Project management & software development
• 2 week sprints
• Promotes continuous adaptation
Cultural and behavioural changes

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

**Agile**
- Sprint releases
- Small & frequent MVP deployments
- Predominately co-located teams
- Team are empowered to make decisions
- Scope changes made iteratively
- Plan continuously and iteratively
- Adapting change based on need and understanding
- Cross functional teams
Conversion timeline in sprint – Example

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

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

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

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

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

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

**Sprint 6**
- **Team** to perform Business Validation

**Sprint 7**
- **Analyst** to obtain sign-off
- **Developer** to run Production version
Source Data Analysis
OHDSI Tools for Analysis

OHDSI Tools

White Rabbit – Source data analysis

Usagi – Custom mapping

Rabbit in a Hat – ETL Spec writing
Source data analysis

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

1. White Rabbit Download
   https://github.com/OHDSI/WhiteRabbit

2. Find the “Latest Release” and download the WhiteRabbit zip file

3. Unzip the download

4. Double-click on `bin/whiteRabbit.bat` on Windows to start White Rabbit
White Rabbit – Location and Scan
White Rabbit – Scan
Reading the Scan

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

Series of tabs in an XLSX file

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

- Defines the tables you scanned

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table</td>
<td>Field</td>
<td>Type</td>
<td>Max length</td>
<td>N rows</td>
<td>N rows checked</td>
<td>Fraction empty</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>row_id</td>
<td>int</td>
<td>5</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>subject_id</td>
<td>int</td>
<td>5</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>hadm_id</td>
<td>int</td>
<td>6</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>admittime</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>dischtime</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>deathtime</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>admission_type</td>
<td>varchar</td>
<td>9</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>admission_location</td>
<td>varchar</td>
<td>25</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>discharge_location</td>
<td>varchar</td>
<td>25</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>insurance</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>language</td>
<td>varchar</td>
<td>4</td>
<td>-1</td>
<td>20000</td>
<td>0.42775</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>religion</td>
<td>varchar</td>
<td>22</td>
<td>-1</td>
<td>20000</td>
<td>0.00725</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>marital_status</td>
<td>varchar</td>
<td>17</td>
<td>-1</td>
<td>20000</td>
<td>0.16965</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>ethnicity</td>
<td>varchar</td>
<td>42</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>edregtime</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0.47555</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>edouttime</td>
<td>varchar</td>
<td>10</td>
<td>-1</td>
<td>20000</td>
<td>0.47555</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>diagnosis</td>
<td>varchar</td>
<td>182</td>
<td>-1</td>
<td>20000</td>
<td>0.0004</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>hospital_expire_flag</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
<tr>
<td>admissions.csv</td>
<td>has_chartevents_data</td>
<td>int</td>
<td>1</td>
<td>-1</td>
<td>20000</td>
<td>0</td>
</tr>
</tbody>
</table>
A summary column for each field, there will be as many tabs as tables selected to analyze.

<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>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>row_id</td>
<td>Frequency</td>
<td>subject_id</td>
<td>Frequency</td>
<td>hadm_id</td>
<td>Frequency</td>
<td>admittime</td>
<td>Frequency</td>
<td>discharge</td>
<td>Frequency</td>
<td>deathtime</td>
<td>Frequency</td>
<td>admission_type</td>
<td>Frequency</td>
</tr>
<tr>
<td>2</td>
<td>43575</td>
<td>111861</td>
<td>19</td>
<td>104705</td>
<td>18/1/199</td>
<td>6</td>
<td>8/27/1979</td>
<td>6</td>
<td>18001</td>
<td>EMERGENCY</td>
<td>14326</td>
<td>EMERGENCY ROOM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>42244</td>
<td>1109</td>
<td>17</td>
<td>104703</td>
<td>17/23/132</td>
<td>6</td>
<td>17/23/133</td>
<td>6</td>
<td>2</td>
<td>NEWBORN</td>
<td>2659</td>
<td>PHYS REFERRAL/K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>43576</td>
<td>113033</td>
<td>12</td>
<td>190007</td>
<td>11/17/187</td>
<td>6</td>
<td>41/7/1905</td>
<td>5</td>
<td>7/7/1913</td>
<td>ELECTIVE</td>
<td>2584</td>
<td>CLINIC REFERRAL/P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>43577</td>
<td>15080</td>
<td>11</td>
<td>19001</td>
<td>13/27/136</td>
<td>5</td>
<td>10/22/203</td>
<td>5</td>
<td>8/15/215</td>
<td>URGENT</td>
<td>431</td>
<td>TRANSFER FROM H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>43578</td>
<td>141976</td>
<td>10</td>
<td>190072</td>
<td>18/5/2169</td>
<td>5</td>
<td>5/12/2169</td>
<td>5</td>
<td>3/25/2140</td>
<td>2</td>
<td>TRANSFER FROM S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4371</td>
<td>19620</td>
<td>9</td>
<td>190171</td>
<td>11/23/200</td>
<td>5</td>
<td>1/23/2136</td>
<td>5</td>
<td>7/9/2151</td>
<td>2</td>
<td>** INFO NOT AVAIL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>17234</td>
<td>125941</td>
<td>9</td>
<td>190070</td>
<td>11/12/2124</td>
<td>5</td>
<td>6/20/2113</td>
<td>5</td>
<td>12/4/2126</td>
<td>2</td>
<td>HMO REFERRAL/SIC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>42240</td>
<td>13952</td>
<td>8</td>
<td>190077</td>
<td>19/1/2158</td>
<td>5</td>
<td>12/2/2161</td>
<td>5</td>
<td>12/4/2161</td>
<td>2</td>
<td>TRANSFER FROM O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>43572</td>
<td>123657</td>
<td>8</td>
<td>104741</td>
<td>14/14/2115</td>
<td>4</td>
<td>10/13/2106</td>
<td>5</td>
<td>2/9/2148</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>42243</td>
<td>123707</td>
<td>8</td>
<td>190075</td>
<td>13/25/2170</td>
<td>4</td>
<td>2/4/2157</td>
<td>5</td>
<td>11/27/2155</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>56890</td>
<td>119029</td>
<td>8</td>
<td>116721</td>
<td>11/5/2160</td>
<td>4</td>
<td>6/27/193</td>
<td>4</td>
<td>3/3/2122</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>17283</td>
<td>176476</td>
<td>8</td>
<td>116725</td>
<td>11/16/2179</td>
<td>4</td>
<td>1/2/2158</td>
<td>4</td>
<td>4/26/2195</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>43579</td>
<td>15727</td>
<td>8</td>
<td>187089</td>
<td>11/12/2184</td>
<td>4</td>
<td>2/26/2143</td>
<td>4</td>
<td>8/24/2195</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4869</td>
<td>127800</td>
<td>8</td>
<td>187095</td>
<td>14/21/200</td>
<td>4</td>
<td>11/23/2123</td>
<td>4</td>
<td>12/1/2199</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>30256</td>
<td>125225</td>
<td>8</td>
<td>103408</td>
<td>19/7/2182</td>
<td>4</td>
<td>7/16/2133</td>
<td>4</td>
<td>6/15/2173</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>56896</td>
<td>126643</td>
<td>8</td>
<td>187064</td>
<td>11/8/2127</td>
<td>4</td>
<td>15/2183</td>
<td>4</td>
<td>5/1/2104</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>3639</td>
<td>13929</td>
<td>7</td>
<td>128701</td>
<td>11/12/2180</td>
<td>4</td>
<td>6/2/2130</td>
<td>4</td>
<td>11/4/2106</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2305</td>
<td>196666</td>
<td>7</td>
<td>187090</td>
<td>11/12/2195</td>
<td>4</td>
<td>6/2/2149</td>
<td>4</td>
<td>2/12/2136</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>54232</td>
<td>13100</td>
<td>7</td>
<td>190067</td>
<td>16/2/2103</td>
<td>4</td>
<td>10/21/2107</td>
<td>4</td>
<td>10/8/2139</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Source Data Analysis – Exercise
Exercise – Scan Mimic data

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

Background

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

Exercises

- How many patients are there in Patients table?
- How many patients do not have date of death (dod) information?
- What is the most common condition (code) among patients?
- How many admission types are there in Admission table? What are they?
- How many patients have no insurance, just “Self Pay”?
- What is the most common drug (drug) patients use?
Using **White Rabbit** to scan Mimic Data and answer the following questions.

**Exercise Answers**

---

**Background**

---

**Exercises**

- How many patients are there in Patients table?
  - 91
- How many patients do not have date of death (dod) information?
  - 65
- What is the most common condition (code) among patients?
  - 4019
- How many admission types are there in Admission table? What are they?
  - 4; EMERGENCY, NEWBORN, ELECTIVE, URGENT
- How many patients have no insurance, just “Self Pay”?
  - 192
- What is the most common drug (drug) patients use?
  - D5W
Break – 1 hour
Vocabulary Mapping
Integration of CDM and Vocabulary

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

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

CONDITION_OCCURRENCE
person_id: 123
condition_concept_id: 312327
condition_start_date: 14Feb2013
condition_source_value: ‘427.31’
condition_source_concept_id: 44821957
Mapping to Standard Concept #1

**Step 1. Lookup the Source Concept**

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

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>44821957</td>
<td>Atrial fibrillation</td>
<td>Condition</td>
<td>ICD9CM</td>
<td>5-dig billing code</td>
<td></td>
<td>427.31</td>
</tr>
</tbody>
</table>

**Step 2. Translate to Standard**

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

<table>
<thead>
<tr>
<th>CONCEPT_ID_1</th>
<th>CONCEPT_ID_2</th>
<th>RELATIONSHIP_ID</th>
<th>VALID_START_DATE</th>
<th>VALID_END_DATE</th>
<th>INVALID_REASON</th>
</tr>
</thead>
<tbody>
<tr>
<td>44821957</td>
<td>313217</td>
<td>Maps to</td>
<td>1970-01-01</td>
<td>2099-12-31</td>
<td></td>
</tr>
</tbody>
</table>

**Step 3. Check out the standard Concept**

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

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>313217</td>
<td>Atrial fibrillation</td>
<td>Condition</td>
<td>SNOMED</td>
<td>Clinical Finding</td>
<td>$</td>
<td>49436004</td>
</tr>
</tbody>
</table>

Determines place in CDM
Mapping to Standard Concept #2

Step 1. Lookup the Source Concept

SELECT * FROM concept WHERE concept_code = '67544050474';

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>45867731</td>
<td>clopidogrel 75 MG Oral Tablet [Plavix]</td>
<td>Drug</td>
<td>NDC</td>
<td>11-digit NDC</td>
<td></td>
<td>67544050474</td>
</tr>
</tbody>
</table>

Step 2. Translate to Standard

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

<table>
<thead>
<tr>
<th>CONCEPT_ID_1</th>
<th>CONCEPT_ID_2</th>
<th>RELATIONSHIP_ID</th>
<th>VALID_START_DATE</th>
<th>VALID_END_DATE</th>
<th>INVALID_REASON</th>
</tr>
</thead>
<tbody>
<tr>
<td>45867731</td>
<td>1322185</td>
<td>Maps to</td>
<td>2015-01-29</td>
<td>2099-12-31</td>
<td></td>
</tr>
</tbody>
</table>

Step 3. Check out the standard Concept

SELECT * FROM concept WHERE concept_id = 1322185;

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1322185</td>
<td>clopidogrel 75 MG Oral Tablet [Plavix]</td>
<td>Drug</td>
<td>RxNorm</td>
<td>Branded Drug</td>
<td>S</td>
<td>213169</td>
</tr>
</tbody>
</table>
Write the SQL query to find the standard concept for this source code: R26.2

**Hint:**

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

Step 1. Lookup the Source Concept
SELECT * FROM concept WHERE concept_code = 'R26.2';

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>45602016</td>
<td>Difficulty in walking, not elsewhere classified</td>
<td>Condition</td>
<td>ICD10</td>
<td>ICD10 code</td>
<td></td>
<td>R26.2</td>
</tr>
</tbody>
</table>

Step 2. Translate to Standard
SELECT * FROM concept_relationship WHERE concept_id_1 = 45602016 AND relationship_id = 'Maps to';

<table>
<thead>
<tr>
<th>CONCEPT_ID_1</th>
<th>CONCEPT_ID_2</th>
<th>RELATIONSHIP_ID</th>
<th>VALID_START_DATE</th>
<th>VALID_END_DATE</th>
<th>INVALID_REASON</th>
</tr>
</thead>
<tbody>
<tr>
<td>45602016</td>
<td>36714126</td>
<td>Maps to</td>
<td>2018-11-28</td>
<td>2099-12-31</td>
<td></td>
</tr>
</tbody>
</table>

Step 3. Check out the standard Concept
SELECT * FROM concept WHERE concept_id = 36714126;

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CONCEPT_NAME</th>
<th>DOMAIN_ID</th>
<th>VOCABULARY_ID</th>
<th>CONCEPT_CLASS_ID</th>
<th>STANDARD_CONCEPT</th>
<th>CONCEPT_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>36714126</td>
<td>Difficulty walking</td>
<td>Condition</td>
<td>SNOMED</td>
<td>Clinical Finding</td>
<td>S</td>
<td>719232003</td>
</tr>
</tbody>
</table>
One source field can go to multiple CDM domains

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

<table>
<thead>
<tr>
<th>diagnosis_code (ICD10)</th>
<th>diagnosis_description</th>
<th>concept_id</th>
<th>concept_name (standard)</th>
<th>domain_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>I48.2</td>
<td>Chronic atrial fibrillation</td>
<td>4141360</td>
<td>Chronic atrial fibrillation</td>
<td>Condition</td>
</tr>
<tr>
<td>Z31.5</td>
<td>Genetic counseling</td>
<td>4196362</td>
<td>Genetic counseling</td>
<td>Procedure</td>
</tr>
<tr>
<td>Z82.3</td>
<td>Family history of stroke</td>
<td>4169009</td>
<td>Family history of stroke</td>
<td>Observation</td>
</tr>
<tr>
<td>R71</td>
<td>Abnormality of red blood cells</td>
<td>4098353</td>
<td>Red blood cell test</td>
<td>Measurement</td>
</tr>
</tbody>
</table>
Find out the destination table (domain) for following diagnosis data:

<table>
<thead>
<tr>
<th>diagnosis_code (ICD10)</th>
<th>diagnosis_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R10.0</td>
<td>Acute abdomen</td>
</tr>
<tr>
<td>Z01.1</td>
<td>Examination of ears and hearing</td>
</tr>
<tr>
<td>Z85.6</td>
<td>Personal history of leukaemia</td>
</tr>
<tr>
<td>R77.0</td>
<td>Abnormality of albumin</td>
</tr>
<tr>
<td>diagnosis_code (ICD10)</td>
<td>diagnosis_description</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>R10.0</td>
<td>Acute abdomen</td>
</tr>
<tr>
<td>Z01.1</td>
<td>Examination of ears and hearing</td>
</tr>
<tr>
<td>Z85.6</td>
<td>Personal history of leukaemia</td>
</tr>
<tr>
<td>R77.0</td>
<td>Abnormality of albumin</td>
</tr>
</tbody>
</table>
Vocabulary Mapping
– Exercise
Vocabulary Mapping Exercise

- On the Box, go to folder ‘C:\Users\iqvia-ohdsi\Desktop\Student’
- Open file ‘ETL Exercises - Student Sheet’
- Do exercise in **Day1 Vocabulary Mapping** tab
Custom Mapping of Unmapped Codes Using Usagi
Custom source code mapping

**Unmapped Codes**
- No existing source code mapping
- No source codes, only text
- Medical coding system doesn’t exist in OHDSI

**How much mapping is needed?**

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

**What is done?**
- Analyst manually map source codes
- Review with internal stakeholders

**Vocabulary Team**
- Group of medical and technical experts

**What is done?**
- Send the source codes
- Give us back the mapping
- Review with our internal stakeholders
Purpose of Usagi

What are unmapped codes?

- Source codes are not found in OHDSI CONCEPT table
- Source codes are found in OHDSI CONCEPT table but standard concepts are not available in CONCEPT_RELATIONSHIP table
- Source fields do not have code but only contain text description

What to do?

Use Usagi for custom mapping

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

- Requires medical expertise
- Non-English descriptions
- Time consuming
  - No capacity to custom map thousands of codes
  - Instead focus on most frequent (95%)
- Requires updating
  - A need to revisit custom mapping
  - New codes added
  - Old standard concepts become invalid

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

2. Get a copy of the Vocabulary from ATHENA – http://athena.ohdsi.org
3. Have Usagi build an index on the Vocabulary
4. Load your source codes and let Usagi process them
5. Review and update suggest mappings with medical experts
6. Export codes into the SOURCE_TO_CONCEPT_MAP
### Usagi Demo

The Usagi Demo interface displays a list of concepts related to medical conditions, diagnoses, and other health-related terms. The interface includes columns for Source code, Source term, Frequency, IPC DES, Watch score, Concept ID, Concept name, Domain, Concept class, Vocabulary, Concept code, Standard concept, and Parents/Children. Each concept is categorized under specific conditions and disorders, such as "No illness," "Condition," and "Clinical finding." The interface also provides filters and search options for querying and managing the concepts.
Break – 15min
ETL Specification Writing
What is an ETL Specification

- Document created by analysts
- Roadmap for the development team
- Used during QA process

- Cooperate with Data Owner
- Tells exactly which fields to map into the OMOP model
  - Applies rules to the data
  - Specifies what records to deduplicate or filter out completely
- Cross reference ETL Spec to ensure rules were applied
Creating ETL Specification

1. **Analyze Data**
   - Review the source data table by table, field by field
   - Study the data dictionary
   - Study any other supporting documents

2. **Work with Data Owners**
   - Confirm your understanding of the data
   - Ask questions on things that are not clear

3. **Continued Project Review**
   - Review with team
   - Review with data owners

<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person_Id</td>
<td></td>
<td>System generated id based on unique source identifier</td>
</tr>
<tr>
<td>Gender_concept_id</td>
<td>Bene_sex_ident_cd</td>
<td>If 1 then ‘8507’&lt;br&gt; If 2 then ‘8532’&lt;br&gt; All else/unknown = 0</td>
</tr>
<tr>
<td>Year_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘YYYY’.&lt;br&gt; Exclude patients with NULL or invalid year of birth</td>
</tr>
<tr>
<td>Month_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘MM’.</td>
</tr>
<tr>
<td>Day_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘DD’.</td>
</tr>
</tbody>
</table>
Tables in ETL Specification

- Standardized clinical data
  - Person
    - Observation_period
    - Visit_occurrence
      - Visit_detail
    - Condition_occurrence
    - Drug_exposure
    - Procedure_occurrence
    - Device_exposure
    - Measurement
    - Note
      - Note_NLP
    - Survey_conduct
    - Observation
    - Specimen
    - Fact_relationship

- Standardized health system data
  - Location
  - Care_site
  - Provider

- Standardized derived elements
  - Condition_era
  - Drug_era
  - Procedure_era
  - Device_era
  - Measurement
  - Note
    - Note_NLP
  - Survey_conduct
  - Observation
  - Specimen
  - Fact_relationship

- Standardized health economics
  - Cost
  - Payer_plan_period

- Standardized vocabularies
  - Concept
  - Vocabulary
  - Source_to_concept_map
  - Concept_synonym
  - Concept_ancestor
  - Source_to_concept_map
  - Drug_strength
  - Concept_relationship
  - Concept_ancestor
  - Domain
  - Concept_class
  - Vocabulary
  - Concept_synonym
  - Concept_ancestor
  - Domain
  - Concept_class
  - Source_to_concept_map
  - Standardized metadata
  - CDM_source
  - Metadata
  - Person
  - Survey_conduct
  - Note_NLP
  - Visit_detail
  - Visit_occurrence
  - Condition_occurrence
  - Drug_exposure
  - Procedure_occurrence
  - Device_exposure
  - Measurement
  - Note
    - Note_NLP
  - Survey_conduct
  - Observation
  - Specimen
  - Fact_relationship

- Standardized health economics
  - Cost
  - Payer_plan_period

- Standardized metadata
  - CDM_source
  - Metadata
  - Person
  - Survey_conduct
  - Note_NLP
  - Visit_detail
  - Visit_occurrence
  - Condition_occurrence
  - Drug_exposure
  - Procedure_occurrence
  - Device_exposure
  - Measurement
  - Note
    - Note_NLP
  - Survey_conduct
  - Observation
  - Specimen
  - Fact_relationship
ETL Spec Table Writing Sequence (Recommended)

Dimension tables
- Person
- Provider
- Care_Site
- Location

Visit tables
- Visit_Occurrence
- Visit_Detail

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

Health Economic tables
- Payer_Plan_Period
- Cost
ETL Spec Content – Common Data Elements to All Event Tables

Clinical event tables
• Condition_Occurrence
• Procedure_Occurrence
• Drug_Exposure
• Device_Exposure
• Measurement
• Observation
• Specimen

Common primary key and foreign key columns in clinical event tables

<table>
<thead>
<tr>
<th>Field name</th>
<th>Purpose and example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;entity&gt;_id</td>
<td>Primary key for the entity</td>
</tr>
<tr>
<td>Person_id</td>
<td>Foreign key to the Person table</td>
</tr>
<tr>
<td>Provider_id</td>
<td>Foreign key to the Provider table</td>
</tr>
<tr>
<td>Visit_occurrence_id</td>
<td>Foreign key to the Visit_occurrence table</td>
</tr>
</tbody>
</table>

Common vocabulary related columns in clinical event tables

<table>
<thead>
<tr>
<th>Field name</th>
<th>Purpose and example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;entity&gt;_concept_id</td>
<td><strong>Standard</strong> OMOP concept_id for source value condition_concept_id 4068155 (SNOMED &quot;Atrial arrhythmia&quot;)</td>
</tr>
<tr>
<td>&lt;entity&gt;_source_concept_id</td>
<td>OMOP concept_id for source value condition_source_concept_id 45596206 (ICD10 &quot;Atrial fibrillation and flutter&quot;)</td>
</tr>
<tr>
<td>&lt;entity&gt;_source_value</td>
<td>Verbatim information from the source data, <strong>not to be used</strong> by any standard analytics condition_source_value I48 (ICD10 &quot;Atrial fibrillation and flutter&quot;)</td>
</tr>
<tr>
<td>&lt;entity&gt;_type_concept_id</td>
<td>OMOP concept_id for the <strong>origin of the information</strong> condition_type_concept_id 32817 (&quot;EHR&quot;) Domain = ‘Type Concept’, Concept = ‘Standard’ in ATHENA</td>
</tr>
</tbody>
</table>
ETL Spec – Written in a Template

<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Applied Rule</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person_id</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender_concept_id</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year_of_birth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month_of_birth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day_of_birth</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Destination Field = OMOP field being referenced
- Source Field = field from source data that will be mapped into the Destination Field
- Applied Rule = any rules that are being applied to the data as it is mapped in
- Comment = additional notes that are relevant
<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Applied Rule</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person_Id</td>
<td></td>
<td>System generated id based on desynpuf_id</td>
<td></td>
</tr>
<tr>
<td>Gender_concept_id</td>
<td>Bene_sex_ident_cd</td>
<td>If 1 then ‘8507’</td>
<td>8507 is Male</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If 2 then ‘8532’</td>
<td>8532 is Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All else/unknown = 0</td>
<td></td>
</tr>
<tr>
<td>Year_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘YYYY’.</td>
<td>Exclude patients with NULL or invalid year of birth</td>
</tr>
<tr>
<td>Month_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘MM’.</td>
<td></td>
</tr>
<tr>
<td>Day_of_birth</td>
<td>Bene_birth_dt</td>
<td>Format is YYYY-MM-DD. Map in ‘DD’.</td>
<td></td>
</tr>
</tbody>
</table>
ETL Spec – Source and Target Tables Relationship

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

<table>
<thead>
<tr>
<th>Destination Field</th>
<th>Source Field</th>
<th>Applied Rule</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition_occurrence_id</td>
<td>A unique, system generated identifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person_id</td>
<td>Cdm.person_id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_concept_id</td>
<td>icd10_dgns_cd_1 OR icd10_dgns_cd_2 OR icd10_dgns_cd_3</td>
<td>Create one condition occurrence record for each ICD10 diagnosis code on source record</td>
<td></td>
</tr>
<tr>
<td>Condition_start_date</td>
<td>Clm_from_dt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_start_datetime</td>
<td>NULL</td>
<td>Information is not available in the source data</td>
<td></td>
</tr>
</tbody>
</table>
Writing ETL Spec with Rabbit in a Hat
Rabbit in a Hat

• Is also part of the White Rabbit Download https://github.com/OHDSI/WhiteRabbit

• Allows users to map source fields in OMOP fields

• Can read and display a White Rabbit scan document

• Provides a graphical interface to allow a user to connect source data to tables

• Generates ETL Spec document, does not generate code
Rabbit in a Hat – Start

- Double click on Rabbit in a Hat from its stored location to start the application
- Select File, Open Scan Report. Use the Scan Report we recently generated with White Rabbit
• Select source table to learn more about the fields.
Rabbit in a Hat – Learn CDM tables

- Select OMOP table to learn more about the fields.
ETL Specification Writing – Exercise & Homework
Exercise & Homework

• Map the Mimic data using Rabbit in a Hat

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