Common Data Model
v4 to v5 Conversion Script

Anthony Sena
11 August 2015
Agenda

• OMOP CDM & conversion from CDM V4 to V5
• Where can I find the conversion script?
• Using the conversion script
• Questions/Comments
How OHDSI Works

Source data warehouse, with identifiable patient-level data

ETL

Standardized, de-identified patient-level database (OMOP CDM v5)

Standardized large-scale analytics

Analysis results

OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

Summary statistics results repository

OHDSI.org

OHDSI Data Partners
Objectives in OMOP Common Data Model development

• One model to accommodate both administrative claims and electronic health records
  – Claims from private and public payers, and captured at point-of-care
  – EHRs from both inpatient and outpatient settings
  – Also used to support registries and longitudinal surveys

• One model to support collaborative research across data sources both within and outside of US

• One model that can be manageable for data owners and useful for data users (efficient to put data IN and get data OUT)

• Enable standardization of structure, content, and analytics focused on specific use cases
Evolution of the OMOP Common data model

OMOP CDM now Version 5, following multiple iterations of implementation, testing, modifications, and expansion based on the experiences of the OMOP community who bring on a growing landscape of research use cases.

http://omop.org/CDM
New tables in CDMv5

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

Standardized health system data
- Location
- Care_site
- Provider
- Payer_plan_period
- Visit_cost
- Procedure_cost
- Drug_cost
- Device_cost

Standardized derived elements
- Cohort
- Drug_era
- Dose_era
- Condition_era

Standardized health economics

Standardized meta-data
- CDM_source
  - Concept
  - Vocabulary
  - Concept_relationship
  - Relationship
  - Concept_synonym
  - Concept_ancestor
  - Source_to_concept_map
  - Drug_strength
  - Cohort_definition

Standardized vocabulary
Codes in one coding system can belong to different domains. For example:

ICD-9 V80.1 (Screening for glaucoma)

is a diagnose code, but indicates that a procedure was performed. In the new vocabulary it is mapped to

• SNOMED 171215009 (Glaucoma screening)

This concept belongs to the Procedure domain, and should go into the Procedure_occurrence table
Vocabulary-driven ETL

• Only Standard concepts are allowed in the CDM concept_id fields (see standard_concept field in Concept)

• Each Standard concept has one (and only one) domain (see domain_id field in Concept), which defines where the concept should go
Many-to-many mappings

Source data:
- Diagnoses
- Procedures
- Drugs
- Lab values

CDM v5:
- Condition_occurrence
- Procedure_occurrence
- Drug_exposure
- Measurement
- Observation
- Device_exposure

Typical:
- Diagnoses
  - 90% conditions
  - 8% procedures
  - 2% other
- Procedures
  - 70% procedures
  - 20% measurements
  - 10% devices
- Drugs
  - 90% drugs
  - 10% devices
- Lab values
  - 100% measurements
Progression of OMOP CDM ETLs

Past
- Patient-level data in source system/schema
- OMOP CDM v4

Future
- Patient-level data in source system/schema
- OMOP CDM v5

Temp bridge
- Patient-level data in source system/schema
- OMOP CDM v4
- V4 ➔ V5 ETL
- OMOP CDM v5
Conversion script mappings

Data carries forward from V4 to V5
Where can I find the script?

- GitHub -> CommonDataModel -> [Version 4 To Version 5 Conversion](#)

- What you’ll find on GitHub:
  - An “OHDSI-SQL” (aka templateSQL) script that you can modify and run through SqlRender
  - Subfolders that contain rendered SQL files for the different RDBMS (SQL Server, PostgreSQL, Oracle, etc)

- Script has 2 sections:
  - Conversion from V4 to V5
  - Queries that will provide some basic stats to perform some quality checks of the conversion
How do I use it?

• Prerequisites
  – Create a target **CDM v5 database** on the same database server as your CDM v4 database
  – Download the V5 vocabulary from **Athena** and load it into the CDM v5 database
  – Ensure that you have enough disk space to accommodate the creation of the new. We estimate you will need 6x the storage of your current V4 database.
  – If you made any changes in your V4 database, be sure to carry those over to V5 and to check the conversion script for potential impact. For example, if you converted a field to BIGINT in V4, you will need to ensure this is carried forward.
How do I use it?

• Running the script
  – The script requires read rights from the V4 database and will need owner privileges to the V5 database
  – Perform a find & replace on the script for the search terms:
    • [SOURCE_CDMV4].[SCHEMA] - Your V4 database name + schema
    • [SOURCE_CDMV4] - Your V4 database name
    • [TARGET_CDMV5].[SCHEMA] - Your V5 database name + schema
    • [TARGET_CDMV5] - Your V5 database name
  – Run the script & review the output to see how your data was moved from V4 to V5.
Checking the results

• 3 Queries provide results:
  1. Row count comparisons between V4 and V5 tables
  2. Mapping of V4 tables to target V5 tables using the vocabulary
  3. Total of the V4 tables mapped to their V5 domains versus the V5 row counts to tie out the mapping.
Reviewing the results

Row counts changes due to vocabulary driven conversion process

<table>
<thead>
<tr>
<th>Database Name</th>
<th>TableName</th>
<th>Rows</th>
<th>Database Name</th>
<th>TableName</th>
<th>Rows</th>
<th>Row Count Chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>[CDMV4]</td>
<td>CARE_SITE</td>
<td>1</td>
<td>[CDMV5]</td>
<td>care_site</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>DEATH</td>
<td>3</td>
<td>[CDMV5]</td>
<td>death</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>NULL</td>
<td>NULL</td>
<td>[CDMV5]</td>
<td>device_exposure</td>
<td>8347</td>
<td>8347</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>LOCATION</td>
<td>54</td>
<td>[CDMV5]</td>
<td>location</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>NULL</td>
<td>NULL</td>
<td>[CDMV5]</td>
<td>measurement</td>
<td>108865</td>
<td>108865</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>OBSERVATION</td>
<td>3811</td>
<td>[CDMV5]</td>
<td>observation</td>
<td>38243</td>
<td>34432</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>OBSERVATION_PERIOD</td>
<td>6968</td>
<td>[CDMV5]</td>
<td>observation_period</td>
<td>6968</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>PAYER_PLAN_PERIOD</td>
<td>22459</td>
<td>[CDMV5]</td>
<td>payer_plan_period</td>
<td>22459</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>PERSON</td>
<td>6000</td>
<td>[CDMV5]</td>
<td>person</td>
<td>6000</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>PROCEDURE_COST</td>
<td>777007</td>
<td>[CDMV5]</td>
<td>procedure_cost</td>
<td>777007</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>PROVIDER</td>
<td>7170225</td>
<td>[CDMV5]</td>
<td>provider</td>
<td>7170225</td>
<td>0</td>
</tr>
<tr>
<td>[CDMV4]</td>
<td>VISIT_OCCURRENCE</td>
<td>376185</td>
<td>[CDMV5]</td>
<td>visit_occurrence</td>
<td>376185</td>
<td>0</td>
</tr>
</tbody>
</table>
Reviewing the results

V4 Tables

<table>
<thead>
<tr>
<th>TableName</th>
<th>Domain</th>
<th>RowCount</th>
<th>COUNT(*) from V4</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition_Occurrence</td>
<td>condition</td>
<td>464849</td>
<td>529057</td>
<td>28</td>
</tr>
<tr>
<td>Condition_Occurrence</td>
<td>measurement</td>
<td>8416</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_Occurrence</td>
<td>observation</td>
<td>31522</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_Occurrence</td>
<td>procedure</td>
<td>24298</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>529085</strong></td>
<td><strong>529057</strong></td>
<td><strong>28</strong></td>
</tr>
<tr>
<td>Drug_Exposure</td>
<td>drug</td>
<td>729950</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>729950</strong></td>
<td><strong>729950</strong></td>
<td><strong>0</strong></td>
</tr>
<tr>
<td>Observation</td>
<td>measurement</td>
<td>3811</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>3811</strong></td>
<td><strong>3811</strong></td>
<td><strong>0</strong></td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>device</td>
<td>8347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>drug</td>
<td>20975</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>measurement</td>
<td>96638</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>observation</td>
<td>6721</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>procedure</td>
<td>634209</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>766890</strong></td>
<td><strong>766837</strong></td>
<td><strong>53</strong></td>
</tr>
</tbody>
</table>

Mapping to V5 Domain

Row Count in V5 target tables for rows in the V4 condition_occurrence table

Original row count in V4 for comparison
Reviewing the results

Rows from V4 that map to target tables in V5 with row counts:

\[(634209 + 24298 = 605507)\]
Special Thanks

• Patrick Ryan & Chris Knoll – script authors
• Lee Evans – Testing Environment
Questions or comments

Questions or comments?

Thanks!