OMOP CDM ETL conversion: dataset verification and validation

Alexandra Orlova, Mikhail Archakov, PSM II, Odysseus Data Services, Cambridge MA USA

Abstract

One of the goals of Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is to create standards for the representation of source data to facilitate transparent and reproducible research. Odysseus team is working on converting many different raw datasets to OMOP CDM format and has developed a set of standard policies and approaches to test and validate resulting OMOP data.

Introduction

As OMOP CDM focuses on healthcare data, the critical issue of data quality required for this model should be considered as the highest priority. Since Odysseus Data Services is working on converting multiple various data sources to OMOP format (e.g. MarketScan, Optum, CPRD, etc.), developing standard policies, approaches and tools to perform data assessment and validation of the resulting OMOP instances is essential.

In order to make sure OMOP instance is compliant with OHDSI standards and requirements, the OMOP conversion was performed according to ETL specification, and to check data integrity, Odysseus uses several stages of verification:

- **Unit testing**. This includes checking basic field constraints, as well as some OMOP specific requirements for concept_id fields.
- **Integration testing** includes evaluation of row counts in final tables, complex business rules, custom requirements etc.
- **Data statistics** shows record counts for OMOP tables, displays mapping rates for concepts fields, shows field top values, etc.
- **Analytical validation**. This is a complex process performed by analyst following a specific methodology.

In order to improve verification efficiency of the verification process, some of the stages were automated using Odysseus-developed tools.

Unit testing

Unit testing is used to check field constraints, OMOP requirements, and overall data integrity. Due to the fact that data warehouses, such as Redshift and Hive, do not enforce field constraints like foreign keys, primary keys, etc., these constraints should also be tested.

Standard checks implemented in unit testing:

- **Basic field constraints**: required fields, foreign keys, primary keys, maximum length for string values
- **Check for duplicate records**
- **Basic integrity checks**: event_start_date values should not be greater than event_end_date, etc.
- **Concept checks**: all concept_id values in OMOP instance should be valid, standard, and fall to the specific domain (e.g. condition_concept_id should only contain concepts that belong to the Condition domain), and to specific vocabulary.

In addition to general OMOP requirements, datasets-specific checks can be implemented:

- **Checks for specific values**: used if the field can only contain a set of constant values (e.g. gender_concept_id field) or a range of values (e.g. year_of_birth).
- **Foreign keys to source tables**: used for the fields mapped directly from source dataset to OMOP.
- **By-rule testing**: used if the table is populated in ETL using multiple mapping rules.

Integration testing

Integration tests are used to verify that specific business rules and requirements were implemented according to the ETL specification. Integration tests are complex queries, tailored to specific datasets and/or business rules.

The standard formula for the integration test query is: **row_count(Tsource) – row_count(Ttarget)**,

where **Tsource** is table from the source dataset with business rules-related filters applied, and **Ttarget** is the resulting OMOP table. The test is considered passed if such query returns 0, meaning row counts are equal.

Data statistics

Data statistics are gathered after the OMOP conversion is done and collected to the conversion reports.

Statistics include detailed per-table information for the resulting OMOP instance, such as:

- Row counts, both table total and by mapping rule.
- Most frequent values in fields.
- Percentage of distinct / NULL values.
- Distribution of records by date
- Top unmapped codes
- Mapping rate for records and distinct codes

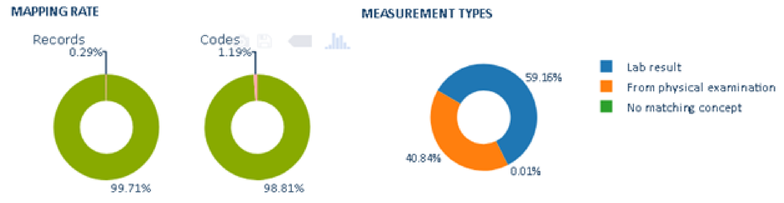

Figure 1. Report example: mapping rates (left) and record distribution by measurement type (right)

While statistical report generation is automated, the analyst is still required to review and verify all the reports.

Analytical validation

Validation is a part of the standard ETL process. It serves to assess ETL conversion results and to make certain that there are no unexpected data losses and source data is converted properly.

Overall, there are 4 types of tables included into ETL conversion:

- Source tables
- Filtered source tables
- Stage CDM tables
- Final CDM tables

Source tables are tables needed to be converted. This is data 'as is'.

Filtered source tables are Source tables after applying cleaning rules. Cleaning rules include common OMOP CDM filters (e.g. exclude patients without year of birth) and specific for particular data asset filters (e.g. exclude particular data feeds or years due to data unreliability).

Stage CDM tables are intermediate OMOP CDM tables with duplicate records (We consider as duplicates in CDM tables records that have identical info except for identity field).

Final CDM tables are Stage CDM tables after removing duplicates. This is data 'to be'.

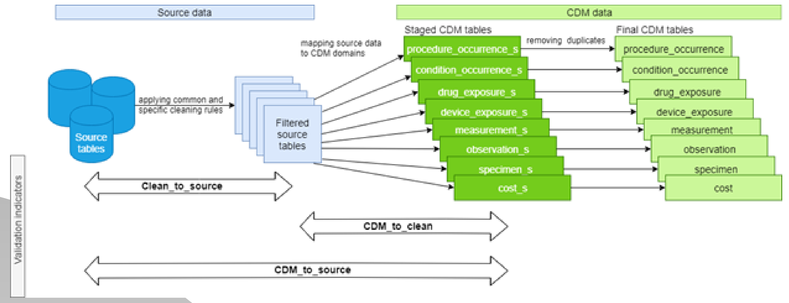

Figure 2. Validation

Indicators

Validation contains three main indicators:

1. **Clean_to_source**: comparison of Source tables and Filtered source tables gives an understanding of the amount of records that do not meet including criteria.
2. **CDM_to_clean**: comparison of Filtered source tables and Stage CDM tables shows data losses during mapping source data to different OMOP CDM domains.
3. **CDM_to_source:** comparison of Source tables and Stage CDM tables reflects the total percentage of included into OMOP CDM re-

Each indicator consists of two: ratio of total amount of records and ratio of total amount of patients. For example, Clean_to_source indicator has ratio of total amount of records from Filtered source tables to the total amount of records from Source tables and ratio of total amount of patients from Filtered source tables to total amount of patients from Source tables. See all indicators and their ratios in the table 1 below

Table 1. Validation indicators

Ratio	Indicators		
	Clean_to_source	CDM_to_clean	CDM_to_source
Total records	total amount of records from Filtered source tables / total amount of records from Source tables	total amount of records from Stage CDM tables / total amount of records from Filtered source tables	total amount of records from Stage CDM tables / total amount of records from Source tables
Patients	total amount of patients from Filtered source tables / total amount of records from Source tables	total amount of patients from Stage CDM tables / total amount of records from Filtered source tables	total amount of patients from Stage CDM tables / total amount of records from Source tables

Methodology

We developed validation methodology that is divided into five parts:

1. High level counts
2. Basic metrics
3. Concept group validation
4. Code list validation
5. Cost validation

Each part contains all three indicators described above. The difference between parts is the object of comparison. Validation results are stored in a separate table named 'Validation'.

High level counts

The main purpose of this section is a comparison of source and OMOP CDM total numbers. This part allows viewing total data losses and distribution of data losses by different types of tables (Source tables, Filtered source tables, CDM Stage tables). See example in table 2 below.

Basic metrics

This section is intended to make sure basic patient metrics were converted properly. We identify seven metrics (Height, Weight, BMI, Blood pressure, Cholesterol, LDL, Tobacco consumption) and created a list of concepts from OMOP CDM vocabularies that denote these metrics. We use this list to find metrics in all three types of tables. See example in table 2 below.

Concept group validation

The main aim of this section is checking data consistency. We check that connection between particular diagnosis and treatment is not losing during conversion. We identify three pairs of diagnosis and treatment (e.g. Diabetes and Metformin / Sitagliptin). This method allows comparing 'diagnosis – treatment' rate in source data and in OMOP CDM tables and if there is a significant difference, it means conversion might have wrong logic or data losses. For this part of validation we calculate only one ratio (total amount of patients). See example in table 2 below.

Code list validation

This section helps to ensure that there are no huge losses in particular groups of medical events. We made several groups (e.g. ACS; Diabetes; Hypertension; Disorders of lipoid metabolism; Other chronic nonalcoholic liver disease; Cardiovascular disease; Chronic kidney disease; Overweight, obesity and other hyperalimentation) and identify templates to find medical codes from these groups in source and OMOP CDM data. If there is a significant difference between source and OMOP CDM counts, then we should check source to OMOP CDM mapping logic. See example in table 2 below.

Cost validation

Cost validation is not applicable for each project. Purpose of this part is to compare the number of records with cost in source data and in OMOP CDM tables and, also, compare the amount of cost (minimum, maximum and average cost). We identify most common five procedures and five drugs and compare costs for them (e.g. Acetaminophen 325 mg oral tablet; Pantoprazole sodium 40 mg oral tablet; Collection of venous blood by venipuncture; Radiologic examination, chest; single view, frontal).

Table 2. Example of Validation table

test id	type id	group name	source counts	clean counts	cdm counts	clean to source	CDM to clean	CDM to source
High level counts	Total records		3,500,000,000	2,900,000,000	2,900,500,000	82.86%	100.02%	82.87%
High level counts	Total patients		250,000,000	238,000,000	238,000,000	95.20%	100.00%	95.20%
Basic metrics	Total records	BMI	3,000	2,800	2,800	93.33%	100.00%	93.33%
Basic metrics	Total patients	BMI	2,900	2,600	2,600	89.66%	100.00%	89.66%
Basic metrics	Total records	Cholesterol	10,000	8,600	8,600	86.00%	100.00%	86.00%
Basic metrics	Total patients	Cholesterol	9,000	7,200	7,200	80.00%	100.00%	80.00%
Basic metrics	Total records	LDL	20,000	19,000	19,000	95.00%	100.00%	95.00%
Basic metrics	Total patients	LDL	15,000	13,700	13,700	91.33%	100.00%	91.33%
Basic metrics	Total records	Tobacco consumption	350,000	330,000	330,000	94.29%	100.00%	94.29%
Basic metrics	Total patients	Tobacco consumption	300,000	280,000	280,000	93.33%	100.00%	93.33%
Concept group validation	Patients	Diagnosis 1	50,000,000	49,900,000	49,900,000	99.80%	100.00%	99.80%
Concept group validation	Patients	Diagnosis 1 - Drug Treatment 1	4,900,000	4,800,000	4,800,000	97.96%	100.00%	97.96%
Concept group validation	Patients	Diagnosis 2	30,000,000	27,600,000	27,600,000	92.00%	100.00%	92.00%
Concept group validation	Patients	Diagnosis 2 - Drug Treatment 2	250,000	230,000	230,000	92.00%	100.00%	92.00%
Concept group validation	Patients	Diagnosis 3	29,000,000	27,800,000	27,800,000	95.86%	100.00%	95.86%
Concept group validation	Patients	Diagnosis 3 - Drug Treatment 3	1,000,500	997,000	997,000	99.65%	100.00%	99.65%
Concept group validation	Patients	Drug Treatment 1	5,100,000	4,730,000	4,730,000	92.75%	100.00%	92.75%
Concept group validation	Patients	Drug Treatment 2	360,000	340,000	340,000	94.44%	100.00%	94.44%
Concept group validation	Patients	Drug Treatment 3	2,280,000	2,000,000	2,000,000	87.72%	100.00%	87.72%

Conclusion

Compliance with existing conventions and policies plays a crucial role in OMOP CDM conversion. Validation and testing are important and helpful parts of OMOP CDM conversion that show if there are any data losses or data inconsistency, and help to identify incorrect mapping logic.