Transforming Maine's All Payer Claims Database to an internationally used common data model to support collaborative research

Tega Dibie, Adam Black, Tom Merrill, Dr. Kate Ahrens

Introduction

Health data sharing is hindered by the disparate nature of data sources and privacy concerns, therefore there is a critical need to standardize health data sources into a common data model that can be used for collaborative research.

Observational Medical Outcomes Partnership (OMOP)

- OMOP Common Data Model is a standardized data model designed to accommodate various health-related data sources, such as administrative claims, electronic health records, and survey data.
- The OMOP common data model has been adopted internationally by health care researchers. Its open-source code allows for a transparent and reproducible process for generating evidence across sites.

All Payers Claims Database (APCD)

The APCD is a repository containing public and commercial claims data for nearly all residents of Maine since 2003, representing over 700 million individual claims.

Research Objectives

- Complete an extract, transform, and load (ETL) process to convert the APCD into the OMOP common data model, a process which includes mapping eligibility information, diagnoses, procedures, and prescriptions filled to standard codes in the common data model.
- Perform a proof of concept study to validate the performance of the common data model.

Data Sources

- The primary source of data for this endeavour is Maine's APCD as released by Maine Health Data Organization (MHDO).
- Medical claims, Medical eligibility, Pharmacy claims, National Provider identifier (NPI) table, residential zip codes.

OMOP Vocabulary

- The OMOP vocabulary allows for the mapping of various medical codes, such as CPT4, ICD-9, ICD-10, RxNorm, SNOMED, HCPCS, NDC etc., from the source tables to standard concepts (unique across all databases regardless of coding system in source).
- The CDM vocabulary allows for the introduction of hierarchical relationships among different related conditions/concepts and more.

Methods

Extract, Transform, and Load (ETL) Development

ETL code was developed and implemented using SQL and R (open source programming language) and released on github as an R package. All data transformation logic was implemented in SQL and ETL orchestration (i.e. triggering execution of SQL and logging progress) was implemented in R. SQL code was written to run on SQL server but could be modified for use in other platforms.

Execution Framework

All transformation took place in a single SQL server database. Database schemas were used to logically separate the stages of the ETL.

Step 1: raw/source tables - Source data as released by MHDO.
Step 2: staging tables - First level cleaning and restructuring of source data.
Step 3: transform tables - Intermediate tables with source and standard mappings.
Step 4: Target Table - The final Common Data Model tables including the OMOP vocabulary tables (see https://github.com/OHDSI/CommonDataModel/wiki/)

Figure 1: ETL Process Map

Mapping source codes to a standard representation

Health care data is represented in a wide variety of coding systems across the world. In order to develop interoperable epidemiological studies, that can run on many different datasets, the representation of healthcare data must be standardized. The ETL process we performed mapped the diagnosis, procedure, and drug codes in the APCD to their standard representation in the OMOP vocabulary.

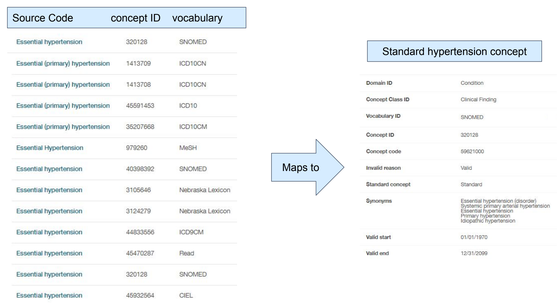


Figure 2: Example of mapping hypertension source codes to a standard representation

Results

ETL code and documentation

- We've publicly made available the code and documentation related to the transformation of the APCD to the OMOP Common Data Model.
- ETL code can be accessed via https://github.com/ablack3/APCDtoOMOP
- ETL specification documentation can be accessed via https://github.com/ablack3/APCDtoOMOP/blob/master/docs/Maine%20APCD%20to%20OMOP%20ETL%20specification.pdf

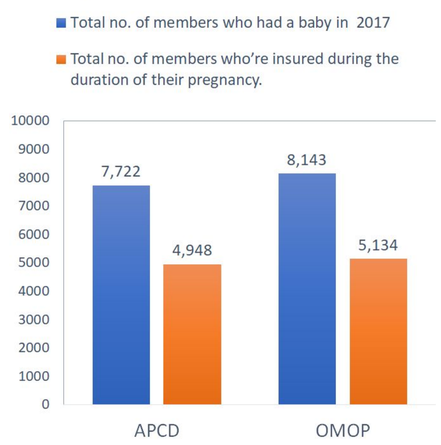
Data Quality Test results

- To validate the APCD to OMOP conversion, we conducted an identical analysis on the raw APCD data and on the OMOP Common Data Model, and compared the obtained results.
- For our analysis, we found the total number of members who delivered a baby in 2017 and were fully insured during the duration of their pregnancy.

Cohort comparison

	Total no. of Members who had a baby in 2017	Total no. of Members who're insured during the duration of their pregnancy.
Raw APCD	7,722	4,948
OMOP CDM	8,143	5,134

Figure 3: APCD VS OMOP



- As shown in Figure 3, the results obtained from the OMOP Common Data Model and APCD are within + or - 5% of each other.

Discussion

Why this is important?

- OMOP Common Data Model allows for the comparison of results from disparate data sources.
- Scalability: The CDM is optimized to handle computational analysis involving extremely large datasets (> 1 billion observations)
- The availability of specialized web tools such as Atlas enables database exploration, cohort definition, and population level analysis.
- Cross-collaboration: The CDM enables cross institutional collaboration with a network of international researchers currently using the OMOP CDM.
- Various methods/packages that allow for Patient Level estimations and predictions are readily available.

Challenges to implementation

Developing and running the ETL is difficult for a number of reasons. First, multiple competencies are required including deep understanding of the source data, OMOP CDM, SQL programming, data engineering, and database administration. In addition it requires dedicated IT resources and involvement.

Cross Institutional Collaboration

This research work was undertaken by the University of Southern Maine and Maine Medical Center. We opted for setting up identical computing environments in parallel with each institution. The intention was to develop an ETL that would work in both institutions. Identical source data files were obtained from the MaineHealth Data Organization for this project. Each institution took responsibility for their copy of the data and infrastructure. Code was shared using a private git repository on Azure DevOps hosted by MMCRI.

Conclusion

The extract, transform, and load process successfully converted Maine's All Payer Claims Database to the OMOP common data model. Future work should examine the comparability of the multiple data sources , as well as extend the use of the OMOP common data model in cross-institutional research.

Acknowledgements

This work was funded by the Maine Economic Improvement Fund, and completed under a Memorandum of Understanding with the Maine Health Data Organization and the University of Southern Maine and Maine Medical Center to support workforce training in health care data analyses and research.