(OHDSI 2016 Symposium - Poster Presentation)

Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDM v5: ETL-CMS software and processed data available and feature-complete

Christophe G. Lambert, PhD[1], Amritansh[2], Praveen Kumar[2]
[1]Center for Global Health, Division of Translational Informatics, Dept. of Internal Medicine.
[2]Dept. of Computer Science. University of New Mexico, Albuquerque, NM.
{cglambert, amritansh, pkumar81}@unm.edu

Abstract

We announce the availability of a public OMOP Common Data Model v5 (CDM v5) dataset containing 2.33 million synthetic patients from the Centers for Medicare & Medicaid Services (CMS) Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). We anticipate that this resource will be useful for researchers in developing OHDSI tools, as well as serve as a testbed for the analysis of observational health records. Despite the synthetic nature of the data, we show, for instance, that it is representative enough of the real world to successfully apply Aphrodite for phenotype modeling. The source code for the extract-transform-load (ETL) tool is available at the OHDSI/ETL-CMS github site, and the processed data is also being made available to the OHDSI community in .csv file format. This marks the first availability of a massive open CDM v5-adhering synthetic dataset. We describe our challenges, learnings and open issues with working with the ETL process, and present results using various OHDSI tools with the data.

Introduction

The Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF), made available by Centers for Medicare & Medicaid Services (CMS)[1], provides a realistic set of administrative claims data that can be helpful in developing software/applications that can be applied to actual claims data. The DE-SynPUF data can be used to test various OHDSI tools and workflows, but as the DE-SynPUF is synthetically generated data, it has somewhat limited inferential utility for real observational studies. Nevertheless, because real Medicare/Medicaid data closely follows the DE-SynPUF format, the data can serve as a useful testbed for methods and pipelines before acquisition from sources such as http://www.resdac.org.

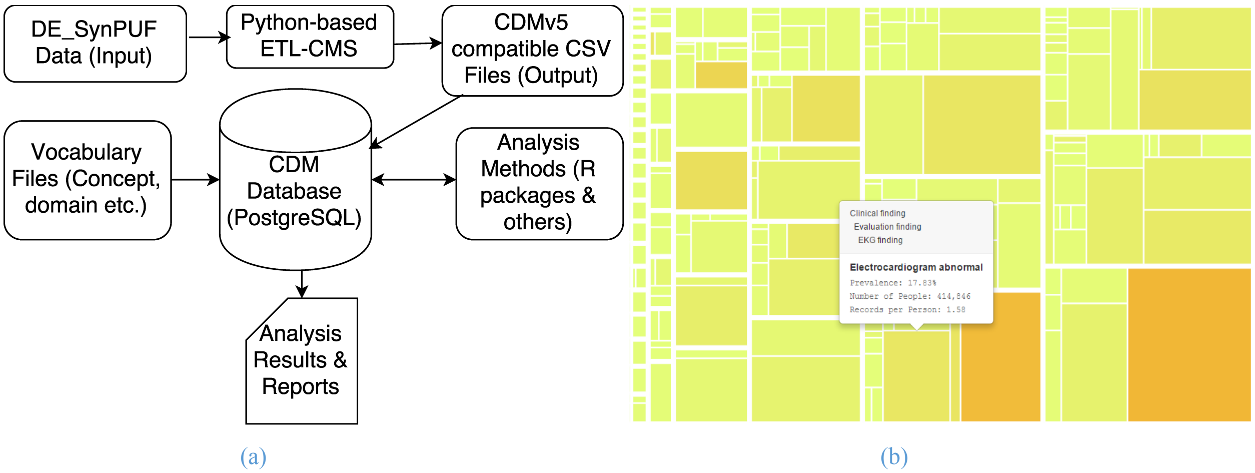
DE_SynPUF Data (Input)
Python-based ETL-CMS
CDMv5 compatible CSV Files (Output)
Vocabulary Files (Concept, domain etc.)
CDM Database (PostgreSQL)
Analysis Methods (R packages & others)
Analysis Results & Reports
Clinical finding
Evaluation finding
EKG finding
Electrocardiogram abnormal
Prevalence: 17.83%
Number of People: 414,846
Records per Person: 1.58
(a)
(b)

Figure 1. (a) The process flow of our pipeline for implementing the DE-SynPUF ETL process. Input files are converted to CDM v5 compatible .csv files by the ETL-CMS tool. Once copied into the database with all constraints, along with standardized vocabularies, Achilles analysis generates reports based on the synthetic dataset. (b) An Achilles analysis heatmap representation for measurement prevalence. The visualization was generated by the OHDSI AchillesWeb tool.

An early release of the Python-based ETL-CMS[2] software was developed by members of the CMS Working Group of the Observational Health Data Sciences and Informatics (OHDSI)[3] community to process the DE-SynPUF files and to create OMOP CDM v5-compatible[4] CSV files. Development was partial, stopping in August 2015. Our group resumed development in December 2015 and implemented the complete ETL, adding missing tables, and correcting numerous errors. We also created documentation for running the ETL, creating an OMOP CDM v5 database, and loading the DE-SynPUF data. Among many improvements, we overhauled the ETL to implement the visit_occurrrence, location, care_site, payer_plan_period tables, and we rectified numerous deficiencies in concept mapping, in order to be feature-complete with the CDM v5. All tables now conform to the constraints defined in the schema. After loading the data into PostgreSQL, we ran SQL queries to create condition_era and drug_era tables and then iteratively performed Achilles Analysis (including Achilles Heel)[5].

Achilles[5] generates summary statistics and detects quality problems with the patient-level observational health database, suitable for visualization in the OHDSI Atlas program. Achilles runs securely within the confines of our system on top of the CDM v5 database. Using the R environment, Achilles runs analyses without disclosing any patient identifiable information and stores its results in PostgreSQL. Output is exported into the JSON format to be used with AchillesWeb[5] for generating visualizations. The Achilles Heel reports were a tremendous help in identifying data quality problems associated with the output generated by ETL-CMS.

To see how consistent the synthetic data was with real data, we ran OHDSI Aphrodite analyses[6,7] to determine if we could find vocabulary concepts that are associated with phenotypes. Aphrodite uses LASSO regression in order to process a 'Silver standard'[7] of EHR data in conjunction with existing ontologies to build a phenotype model. We examined the dataset with several phenotypes (e.g. Myocardial Infarction, Suicidal Intent and Hypothyroidism) and found that some terms were clearly associated with phenotypes and others were not. Table 1 shows vocabulary concepts that were useful in classifying myocardial infarction cases by Aphrodite.

Table 1. Concept codes, names and their importance scores by Aphrodite (using R glmnet) for the 'Myocardial Infarction' phenotype.

concept_id	concept_name	glmnet score
314666	Old myocardial infarction	6.6116
4185932	Ischemic heart disease	5.9439
444406	Acute subendocardial infarction	5.4229
314059	Right bundle branch block	2.7431
46287340	Cefazolin 500 MG Injection	0.5049

Conclusion

The improvement in the existing python-based ETL-CMS software generates high fidelity data adhering to CDM v5 schema and vocabulary specifications. The ETL-CMS tool improvements and the dataset (in whole or in part) can be used by the OHDSI research community to generate EHR analytics and test tools during software development. Further, we expect that the dataset will serve as a useful learning resource for newcomers to the community.

References

1. Data Entrepreneurs' Synthetic Public Use Files, by Centers for Medicare & Medicaid Services https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-files/SynPUFs/index.html.
2. Extract-Transform-Load CMS (ETL-CMS). Retrieved 6/21/2016 from https://github.com/OHDSI/ETL-CMS.
3. The Observational Health Data Sciences and Informatics (OHDSI). Retrieved 6/21/2016 from http://www.ohdsi.org
4. Reich C, Ryan P and the OHDSI CDM and Vocabulary Development Working Group. OMOP Common Data Model V5.0.1. Retrieved 6/21/2016 from http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm.
5. Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES). Retrieved 6/21/2016 from https://github.com/OHDSI/Achilles.
6. Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE). Retrieved 6/21/2016 from https://github.com/OHDSI/Aphrodite
7. Vibhu Agarwal, Paea Lependu, Tanya Podchiyska, Rick Barber, Mary Boland, George Hripcsak, Nigam Shah. Using narratives as a source to automatically learn phenotype models. DMMI 2014. Retrieved 6/21/2016 from http://www.dmmh.org/dmmi2014_submission_4.pdf?attredirects=0&d=1.