# Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDM*v*5: 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 (**CDMv5**) 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.

## Background

The need for an open and free dataset for patient health data analytics (administrative claims and/or electronic health records) has been felt for a long time. Due to various administrative and legal hurdles, students' use of licensed or data containing protected health information (PHI) is often disallowed. Usually one has to go through an institutional review board (IRB) process for access and take human subjects training to work in such research. This creates barriers to entry for prospective researchers in the Observational Health Data Sciences and Informatics (OHDSI) community[1]. Moreover, until now, open datasets which are freely available have been small and aren't suitable to test OHDSI tools with their full range of features or scalability. To support the growth of the OHDSI community, we felt it was crucial to make available a large open dataset in the OMOP CDMv5 format[2], for testing and learning about patient health data.

The Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) was made available by the Center for Medicare and Medicaid Services(CMS) with the goal of providing a realistic set of administrative claims data in the public domain, while providing the very highest degree of protection to the Medicare beneficiaries' PHI[3]. It contains 2.33 million synthetic patients based on real Medicare claims data, further divided into 20 subsets. The claims data was collected from 2008-2010 and contain drugs, procedures, visits, conditions, providers, costs, deaths, and patient demographics. The claims' format is similar to real claims data as obtainable from http://www.resdac.org.

An early release of the Python-based ETL-CMS software was developed by the CMS Working Group of the OHDSI community to process the DE-SynPUF files and to create CDMv5-compatible CSV files[4]. 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.

## Methods

We created detailed documentation for running the ETL, creating an OMOP CDM v5 database, and loading the DE-SynPUF data[4]. Among many improvements, we overhauled the existing ETL-CMS tool to implement the visit_occurrence, 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 database, we ran SQL queries to create condition_era and drug_era tables and then iteratively performed Achilles Analysis (including Achilles Heel)[5]. Fig. 2 illustrates the overall workflow of our software. All OMOP CDMv5 database tables are now populated to the extent the SynPUF data allows: visit_occurrence, payer_plan_period, location, care_site, etc. (Fig. 1). We made sure no data in the output CDMv5 csv files violates the defined database constraints. All 20 SynPUF parts consistently reference shared information such as provider, care_site, and location. Table 1 summarizes the different source to target vocabulary mappings made by the ETL. Unmapped concepts were handled by assigning the target concept id to '0' if there was no mapping defined from source to target concept. Input data sorting is now consistent across all platforms. A log file is created for the input records with undefined ICD9/HCPCS/NDC codes. The updated version of the ETL-CMS software is available for download on the ETL-CMS github repository[4]. For convenience, we have made the complete output data files, as well as a 1/20th subset, available on the OHDSI ftp website[6].

The processed data still has several caveats:

1. The output data has limits on its inferential research value:
   - It is synthetic data derived from real data.
   - Modifications from real data are undocumented.
2. Trade-offs were made for certain transformations:
   - Drugs were not assigned to visits.
   - Observation periods were defined by earliest and latest event.
   - Payer plan periods employed complex estimates.
3. Input DE-SynPuf records with undefined ICD9/HCPCS/NDC codes were not processed:
   - Some appear to be typos.
   - Some appear to be real but non-standard codes (e.g. 04.22).
4. Drug quantity and days supply are problematic:
   - 6% of drug_exposure quantity and days_supply were 0.
   - The derived dose_era table was therefore left empty.
5. Location information uses SSA codes:
   - Converted to 2-letter state codes (not to specifications).
   - All non-states were lumped into code "54".

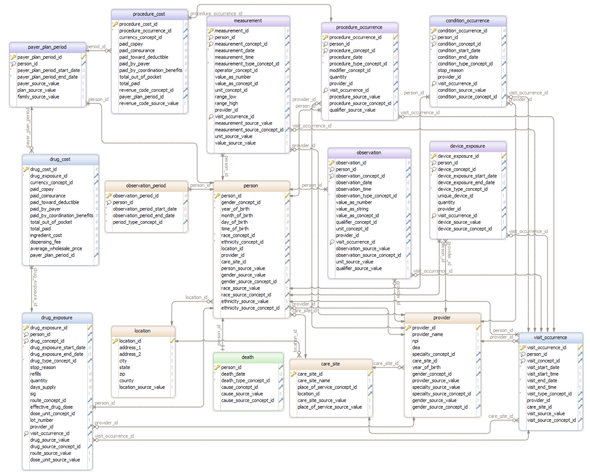


**Figure 1. Entity-relationship diagram of all tables populated by our ETL-CMS software modifications.** We strictly adhere to CDMv5 specifications .

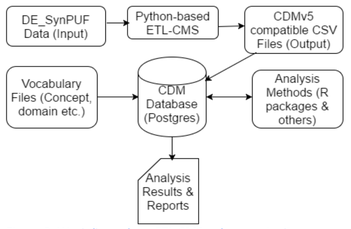


**Figure 2. Work flow of our ETL-CMS software pipeline**.

| Source Vocabulary | Target Vocabulary | No. of Concepts mapped | Example |
|---|---|---|---|
| ICD9Proc | ICD9Proc | 3182 | Excision or destruction of lesion or tissue of abdominal wall or umbilicus (ICD-9: 54.3 self-map) |
| ICD9CM | SNOMED | 13174 | Organic sleep apnea, unspecified (ICD-9: 327.20) → Breathing-related sleep disorder (SNOMED: 111489007) |
| CPT4 | RxNorm | 5 | Diphtheria, tetanus toxoids, and whole cell pertussis vaccine (DTP), for intramuscular use (CPT4: 490701) → diphtheria toxoid vaccine, inactivated (RxNorm: 798304) |
| CPT4 | CPT4 | 7904 | Anesthesia for procedures on major vessels of neck; simple ligation(CPT4: 00352 self map) |
| HCPCS | RxNorm | 575 | Injection, rho d immune globulin, intravenous, human, solvent detergent, 100 iu (HCPCS: J2792) → Rho(D) Immune Globulin (RxNorm: 35465) |
| HCPCS | SNOMED | 33 | End-stage renal disease patient with a hematocrit or hemoglobin not documented (HCPCS: G8387) → End stage renal disease (SNOMED: 46177005) |
| NDC | RxNorm | 274380 | Lithium Carbonate 300 mg Oral Tablet (NDC:43353093509) → Lithium Carbonate 300 mg Oral Tablet (RxNorm: 197890) |

**Table 1. Source to Target vocabulary mapping.** We count the number of instances and provide examples of concept code mappings.

Achilles[5] was an essential tool in developing the ETL and visualizing the resulting data. Achilles generates summary statistics and detects quality problems with the patient-level data, 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 For generating visualizations (Fig. 3, 5). The Achilles Heel reports were a tremendous help in identifying data quality problems associated with the output generated by ETL-CMS (Fig. 4).

To see how consistent the synthetic data were with real data, we ran OHDSI Aphrodite[7] analyses to determine if we could find vocabulary concepts that are associated with phenotypes. Aphrodite uses LASSO regression in order to process a 'Silver standard' of EHR data in conjunction with existing ontologies to build a phenotype model[8]. We examined the dataset with several phenotypes (e.g. Myocardial Infarction, Suicidal Intent and Hypothyroidism) to determine which additional terms were clearly associated with these phenotypes, with results in Table 2.

## Results

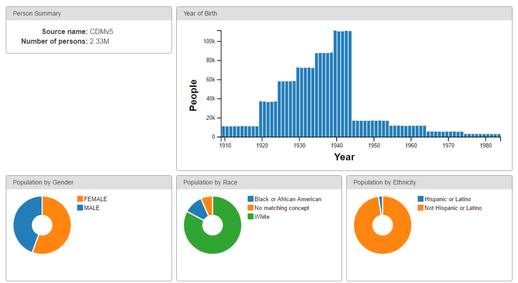


**Figure 3. Achilles WebTool dashboard.** The dashboard displays person level data such as total size of population, year of birth, gender, age and ethnicity .

| Message Type | Message |
|---|---|
| WARNING | 4-Number of persons by race; data with unmapped concepts |
| WARNING | 400-Number of persons with at least one condition occurrence, by condition_concept_id; data with unmapped concepts |
| WARNING | 402-Number of persons by condition occurrence start month, by condition_concept_id; 1,270 concepts have a 100% change in monthly count of events |
| WARNING | 420-Number of condition occurrence records by condition occurrence start month; theres a 100% change in monthly count of events |
| WARNING | 600-Number of persons with at least one procedure occurrence, by procedure_concept_id; data with unmapped concepts |
| WARNING | 602-Number of persons by procedure occurrence start month, by procedure_concept_id; 470 concepts have a 100% change in monthly count of events |
| WARNING | 620-Number of procedure occurrence records by procedure occurrence start month; theres a 100% change in monthly count of events |
| WARNING | 700-Number of persons with at least one drug exposure, by drug_concept_id; data with unmapped concepts |
| WARNING | 702-Number of persons by drug exposure start month, by drug_concept_id; 1,771 concepts have a 100% change in monthly count of events |
| WARNING | 720-Number of drug exposure records by drug exposure start month; theres a 100% change in monthly count of events |
| WARNING | 800-Number of persons with at least one observation occurrence, by observation_concept_id; data with unmapped concepts |
| WARNING | 802-Number of persons by observation occurrence start month, by observation_concept_id; 112 concepts have a 100% change in monthly count of events |
| WARNING | 820-Number of observation records by observation start month; theres a 100% change in monthly count of events |
| WARNING | 902-Number of persons by drug era start month, by drug_concept_id; 152 concepts have a 100% change in monthly count of events |
| WARNING | 1,000-Number of persons with at least one condition era, by condition_concept_id; data with unmapped concepts |

**Figure 4. Achilles Heel report.** The report helps in identifying possible errors and warnings related to data quality. Depicted are the remaining warnings associated with the current version of the ETL-CMS output files, representing non-serious issues that cannot be remedied due to limitations of the content of the source data.

| concept_id | Name of concept | GImnet 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 |

**Table 2. Aphrodites' glmnet score for Myocardial Infarction (MI) phenotype**. Concepts with high scores are not MI events but are predictive of such events.

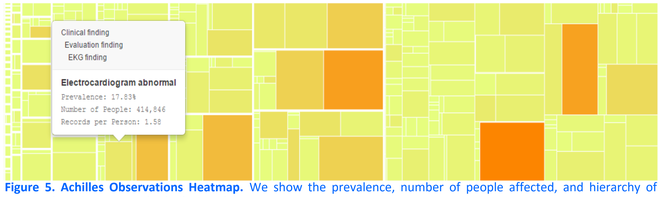


**Figure 5. Achilles Observations Heatmap.** We show the prevalence, number of people affected, and hierarchy of observations. Larger box size indicates higher prevalence and darker orange corresponds to more records per person.

## Conclusions

The improvement in the existing python-based ETL-CMS software generates high fidelity data adhering to the CDM v5 vocabulary specifications. The ETL-CMS tool improvements and the dataset (in whole or in part) can be used by the OHDSI research community to perform patient health data analytics and to test tools during software development. Further, we expect that the dataset will serve as a useful learning resource for newcomers to the community. As further OHDSI tools are developed, this free dataset can be a common resource to test and implement against. Finally, we acknowledge and thank, members of the OHDSI community for their help and advice during our ETL odyssey.

## References

1. The Observational Health Data Sciences and Informatics (OHDSI). Retrieved 9/12/2016 from http://www.ohdsi.org
2. Reich C, Ryan P and the OHDSI CDM and Vocabulary Development Working Group. OMOP Common Data Model V5.0.1. Retrieved 9/12/2016 from http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm.
3. 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.
4. Extract-Transform-Load CMS (ETL-CMS). Retrieved 9/12/2016 from https://github.com/OHDSI/ETL-CMS.
5. Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES). Retrieved 6/21/2016 from https://github.com/OHDSI/Achilles.
6. DE_SynPuf output files on OHDSI website. Retrieved on 9/12/2016 from ftp://ftp.ohdsi.org/synpuf/
7. Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE).Retrieved 6/21/2016 from https://github.com/OHDSI/Aphrodite
8. 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.