Converting the data in the U.S. CMS Virtual Research Data Center to the OHDSI Common Data Model v5

Fabrício S. P. Kury, MD[1], Vojtech Huser, MD, PhD[1]
[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

Abstract

The data made available by the U.S. Centers for Medicare & Medicaid Services (CMS) through the Virtual Research Data Center (VRDC) represent a considerable portion of the total U.S. population and spending on healthcare (currently 79 million patients). The volume of the data, and the restricted VRDC environment, bring particular considerations to the ETL to the OHDSI Common Data Model v5. In this poster we initiate the effort towards enabling OHDSI investigations in the VRDC and discuss its particularities.

Background and Introduction

Big Data research in healthcare is increasingly adopting a Common Data Model (CDM) to allow execution of analyses across several datasets. The Observational Health Data Sciences and Informatics (OHDSI) collaborative maintains a CDM and provides multiple tools to facilitate data analysis. In our research on drug usage, we wanted to take advantage of existing analyses authored by OHDSI researchers on data available to us from the U.S. Centers for Medicare & Medicaid Services (CMS) via the Virtual Research Data Center (VRDC). For that purpose, we sought to transform the CMS VRDC data into OHDSI CDM version 5. While there are two previous works that convert CMS data to the OHDSI CDM, they are not usable inside the VRDC. Danese *et al.* (*https://github.com/OHDSI/ETL-CMS/*) produced ETL code for CMS Synthetic data (SynPuf) files using Python, but Python is not permitted in the VRDC. Evans *et al.* (*http://www.ltscomputingllc.com/downloads/*) produced another ETL for the same input data using Apache Spark, but the code is not publicly available. Therefore, while the work by Danese *et al.* was sometimes helpful as an example, we set out to write our own ETL code that can be executed in the SAS environment inside the VRDC.

Objectives and Methods

In 2013 the CMS announced a novel way for researchers to access its data, namely, the Virtual Research Data Center (VRDC). It consists of a virtual Windows remote desktop accessible only inside a protected, internet-less Virtual Private Network. Due to security policies, VRDC users cannot install any additional applications and are limited to using the SAS software for data analysis. Data transport from/to the VRDC is restricted, and many system functionalities are not accessible, such as the command line shell. Despite these limitations, among the great advantages of the VRDC is the fact that it provides access to the full version of most data files, rather than limited samples. For interim storage, each VRDC Data Use Agreement allows for 500 GB of project-based space. Additional storage can be purchased; nonetheless, performing ETL on all VRDC data files is unfeasible in most cases due to the large data volume. Therefore, our goals were

(1) to produce a SAS program that can run in the VRDC and transform the CMS data into OHDSI CDM v5; and
(2) to evaluate the storage requirements for CDM v5 tables.

Results

We produced a SAS program, available at *https://github.com/fabkury/cms_vrdc_etl* which performs partial ETL of the **PERSON**, **DEATH**, **DRUG_EXPOSURE** and **OBSERVATION_PERIOD** tables. The program is solely composed of SAS macros that generate Structured Query Language (SQL) queries that create each CDM v5 table as a SQL View.

Table 1: ETL for 10 million beneficiaries for years 2007 until 2012.†

VRDC table	VRDC size* (MB)	CDM v5 table	CDM v5 size* (MB)	Beneficiaries	Time‡ (mm:ss)
MBSF_AB_x	4,767.2	PERSON	904.4	10,000,000	05:22
PDE files	7,845.0	DRUG_EXPOSURE	3,445.1	193,821	06:24
MBSF_AB_x	200.3	DEATH	43.6	1,887,414	00:35
MBSF_D_x	4,026.4	OBSERVATION_PERIOD	2,775.9	6,512,641	53:42

†Choice of years is because Medicare Part D enrollment started gradually along 2006. *Only the rows referring to the same beneficiaries. ‡For computing the actual materialized CDM v5 table, not just the SQL View.

Discussion and Conclusion

We have produced an open source code to perform partial ETL of CMS data into OHDSI CDM v5 in a manner suitable for use inside the restricted VRDC environment. The use of SAS macros simplified the code and permits the user to easily limit the ETL to specific years or data files as intended or available under his/her DUA. The use of SQL Views provides three important advantages. First, by using SQL Views instead of physical tables we did not utilize any space within the project-based 500 GB storage quota. Second, if it is known which beneficiaries are meant to be investigated, the Views will automatically impose that restriction from the earliest point possible in the ETL process. Third, if concretely existing CDM v5 tables are ever needed, they can be easily created by copying the rows from the Views.

In some circumstances the ETL process was not clear. For example, we were unable to find suitable CDM v5 vocabulary concepts to separately represent enrollment of a beneficiary in Medicare Parts A, B, D and/or a Health Management Organization (Part C) (*period_type_concept_id*). Our ETL project confirms prior observations that data transformation may lead to loss of useful information detail. For example, the “drug exposure table” in the CMS VRDC schema (i.e. the PDE files) contains columns such as "Benefit Phase”, “Generic Name” and “Brand Name” for each filled drug prescription, for which we could not identify appropriate CDM v5 columns. Conversely, the CDM v5 table can capture “Lot Number”, “Visit Occurrence ID” and “Stop Reason”, for which we had no source data.

As we progress towards a more usable ETL, we also hope to re-evaluate whether SQL Views hold sufficient power for the task.

Contact: Fabrício Kury – kuryf@nih.gov

Conflicts of Interest: The authors declare no conflicts of interest.