# Converting the data in the U.S. CMS Virtual Research Data Center to the OHDSI Common Data Model version 5

**Fabrício S. P. Kury, MD[1], Vojtech Huser[1]**
**[1] 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. 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 its own CDM and provides multiple tools to facilitate data analysis[1]. 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 at al. produced ETL code for CMS Synthetic data (SynPuf) files using Python[2], but Python is not permitted in the VRDC. Evans at al. produced another ETL for the same input data using Apache Spark[3], but did not make the code 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.

## Objective and Methods

In 2013 the CMS announced a novel way for researchers to access its data, namely, the Virtual Research Data Center (VRDC)[4]. It consists of a virtual Windows remote desktop accessible only inside a protected, internet-less Virtual Private Network. Due to security reasons, VRDC users cannot install any additional application and must adhere 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 its restrictions, 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. However, each VRDC Data Use Agreement provides only 500 GB of dedicated storage[5]. Additional storage can be purchased, but, even still, performing ETL on all VRDC data files is unfeasible in most cases due to the large data volume[6]. Therefore, our goals were (1) to produce a SAS program that can run in the VRDC and ETL 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 SQL queries that create each CDM v5 table as a SQL View.

*Table 1: ETL for 1,000,000 beneficiaries for years 1999 until 2012.*

| VRDC table | VRDC table size* | CDM v5 table | CDM v5 table size* | Comput. time |
|---|---|---|---|---|
| MBSF_AB_x | 993.1 MB | PERSON | 90.1 MB | 50.7 seconds |
| PDE Files** | 755.0 MB | DRUG_EXPOSURE** | 337.8 MB | 17.1 seconds |
| MBSF_AB_x | 48.5 MB | DEATH | 10.9 MB | 4.9 seconds |
| MBSF_D_x | 595.9 MB | OBSERVATION_PERIOD | ***1,775.4 MB*** | ***8 min 54 secs*** |

* Only the rows referring to the same 1,000,000 beneficiaries. ** Drug data is limited to years 2006 to 2012.

We found that the CDM v5 tables contain largely less data than the VRDC tables, thereby reducing the burden of the 500 GB limit. For tables PERSON, DRUG_EXPOSURE and DEATH, each beneficiary occupied altogether, in average, 460 bytes of storage space. The table OBSERVATION_PERIOD was excluded from this calculation because, although the data it holds is correct, it is still too far from its ideal form – it presently contains a separate row for each month of Part D enrollment of each beneficiary, which explains its very large size. The ETL computation time was principally dependent on the structure of the SQL View rather than the number of patients included or resulting table size.

## Discussion and Limitations

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 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 available under his/her DUA. The SQL Views-based approach provides three important advantages over creating separate tables in CDM format. First, the tables in CDM format can be accessed as if they concretely existed, while in fact they do not, so no individual storage space is spent and therefore full-file analyses can be possible despite the 500 GB storage limitation of the VRDC. Second, if you already know which beneficiaries you want to investigate, the CDM v5 tables can be accessed with a WHERE clause specifying the desired *person_ids* (*BENE_IDs*) – when doing so, the SQL optimization engine will propagate this restriction throughout the entire ETL process since its beginning, making the whole process much faster to execute. Third, if concretely existing CDM v5 tables are ever needed (e.g.: in case any data needs to be modified after the ETL), they can be easily created by copying the rows from the SQL Views.

In some circumstances the ETL process was not clear. For example, we were unable to find CDM v5 concept IDs to separately represent enrollment of a beneficiary in Medicare Parts A, B, D and/or a Health Management Organization (Part C). Another example, the concept of "ethnicity" does not exist in the CMS data and our approach was to derive it from the race code. Our experience shows the expectable finding that the CMS VRDC tables and the CDM v5 tables are likely not wholly translatable from one to the other or vice versa. For example, the "drug exposure table" in the CMS VRDC (i.e. the Part D Event files) contain "Benefit Phase", "Generic Name" and "Brand Name" for each filled drug prescription, while the CDM v5 table contains "Lot Number", "Visit Occurrence ID" and "Stop Reason".

Finally, our source code, available online, presently constitutes a small fraction of what would be needed to make the ETL usable. For example, dummy (zero) concept IDs were used in most cases while the original CMS VRDC code was copied to the source_id/value field. Further work is needed to extend and polish this ETL, and it remains to be seen whether SQL Views hold sufficient power for this task, or whether they will need to be replaced with actual tables.

## Conclusion

We have initiated development of a SAS program that performs ETL of the CMS claims data inside the VRDC. Our approach to the ETL provides means to overcome some of the challenges inherent to the VRDC, including the handling of its large datasets. We welcome commentary, corrections, contributions and reuse of our code made available online at https://github.com/fabkury/cms_vrdc_etl.

## References

1. Analytic Tools, available at http://www.ohdsi.org/analytic-tools/. Accessed September 14, 2015.
2. Danese et al. Python-based ETL of SynPUF data to CDMv5-compatible CSV files. Available at https://github.com/OHDSI/ETL-CMS/tree/master/python_etl. Accessed September 14, 2015.
3. Evans at al. 1000 Person sample of CMS SynPUF simulated data in CDMV5 format. Available at http://www.ltscomputingllc.com/downloads. Accessed September 14, 2015.
4. CMS Announces New Data Sharing Tool. Available at https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2013-Press-releases-items/2013-11-12.html. Accessed September 14, 2015.
5. Introduction to the Virtual Research Data Center (VRDC). Available at http://www.resdac.org/sites/resdac.org/files/Introduction%20to%20the%20Virtual%20Research%20Data%20Center%20%28VRDC%29%20-%20Slides.pdf. Accessed September 14, 2015.
6. File Sizes for 100% Research Identifiable Claims Files for years 2009-2011. Available at http://www.resdac.org/resconnect/articles/195. Accessed September 14, 2015.