Mapping Common Data Model data tables to an HDF5 file for reproducible machine learning workflows

Janos G. Hajagos, Ph.D.
Department of Biomedical Informatics
Stony Brook University

Objective:

Predicting which inpatients are at highest risk for a 30-day readmission is an important factor for allocating care management resources. There is a rich literature of building machine learning models for predicting hospital readmissions [1]. This poster demonstrates the first steps in building a reproducible workflow for predicting inpatient readmission based on normalized data stored in the OHDSI (CDM) Common Data Model.

Key Concepts:

OHDSI Common Data Model allows healthcare data from various sources to be stored in a single schema with a standardized vocabulary. It grew out of the work to rigorously evaluate methods and data sets for detecting adverse drug events.

HDF5 is a flexible file container for storing arrays in an organized structure. The concept of groups which is similar to file paths allows the data to be stored in a hierarchy. It supports a range of data types and compression methods. It has been used for storing and analyzing the data for the LIGO experiment to detect gravitational waves, see: (https://losc.ligo.org/s/events/LVT151012/LOSC_Event_tutorial_LVT151012.html).

JSON (Javascript Object Notation) is a format for storing data in a machine independent way. Here it is used as an intermediary data format between a relational database and HDF5.

Jupyter notebooks grew out of the iPython project. It is used for creating reproducible and documented computations using a range of backends (Python, R, Julia).

CMS SYNPUF (Synthetic Public Use File) was created to encourage developers to build applications that utilize CMS's Medicare data. It does not contain any PHI (Protected Health Information). It is complex and large enough to demonstrate the feasibility for a project. It was mapped to the OHDSI CDM in 2016.

Methods:

A 100,000 encounter subset of CMS's SYNPUF (Synthetic Public Use File) and OHDSI vocabulary files were loaded into a PostgreSQL (version 9.6) relational database system. Tables were first denormalized at the visit occurrence level and the results were stored in temporary database tables. A total of 66,700 inpatient visits were extracted from the relational database as a set of JSON (Javascript Object Notation) documents. The JSON documents were mapped to a single HDF5 file. The mapped file contained multiple matrices where each row represents a separate inpatient visit and a column a feature associated with a domain. Both mapping steps are controlled by separate configuration files. The generated HDF5 file is then post processed and a 30-day readmission flag is appended to the HDF5 file as a separate dataset.

Using a Jupyter/IPython3 notebook features are selected across multiple groups and assembled into a single matrix (66,700 rows by 5,686 columns). All conditions, observations, procedures, and measurements were included. Additionally, the gender (female), length of stay in days, age in years, and past history of readmissions were included. In total there were 6,241 30-day hospital readmissions. A random forest model was trained to predict 30-day readmission following discharge. To measure performance of the classifier the AUC (Area Under the Curve) for the ROC (Receiver Operating Curve) was calculated

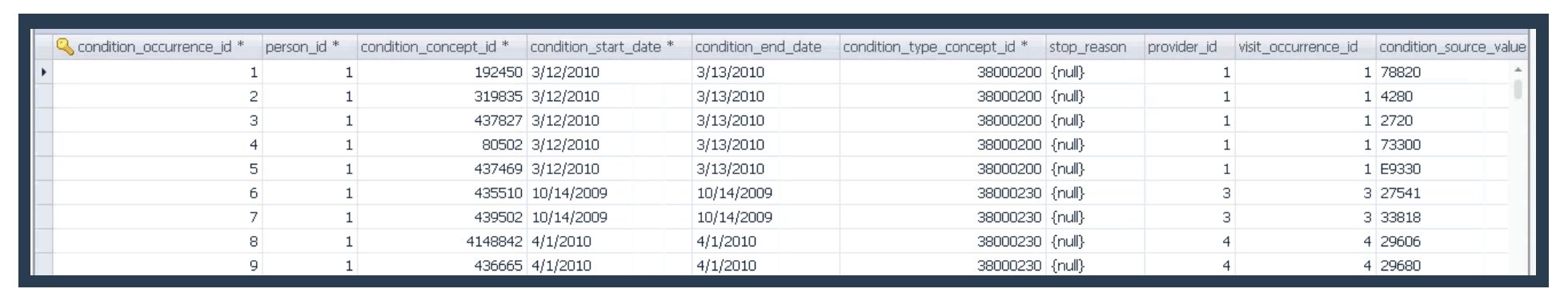
Results:

The HDF5 file generated from the SYNPUF inpatient visits set is 7.59 Mb (Megabytes), the post processed file is 9.40 Mb, and the file used in the predictive model fitting is 4.85 Mb. The complete analysis of the data can be found in the Github project: https://github.com/SBU-BMI/MappingOHDSI2HDF.

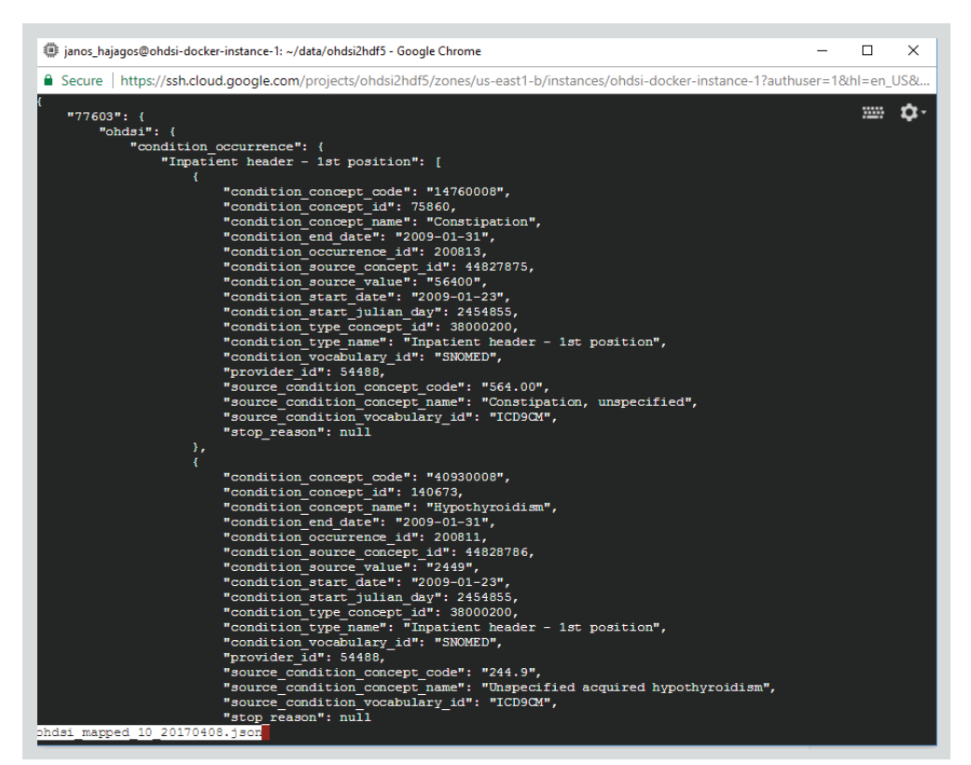
A random forest model with a population of 500 trees and two feature selection steps: remove zero variance features and select K best features (ANOVA F-score with 250 features) were utilized to predict 30-day inpatient readmission. On the testing set which was not utilized for training the total (AUC) area under the curve was 0.53. The predictive ability of the model trained on the CMS's synthetic data to predict 30-day readmission is poor.

Mapping OHDSI CDM data to HDF5

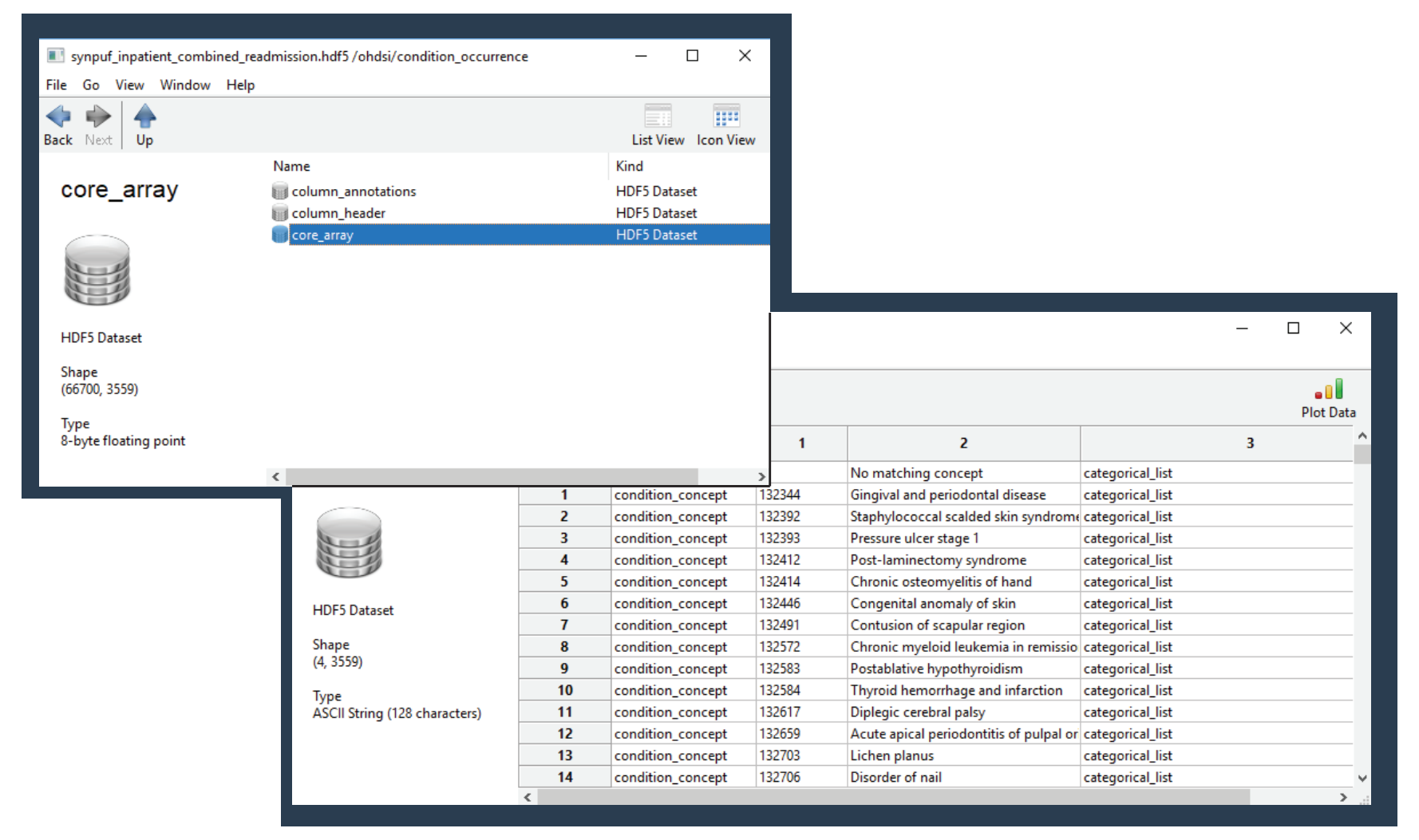
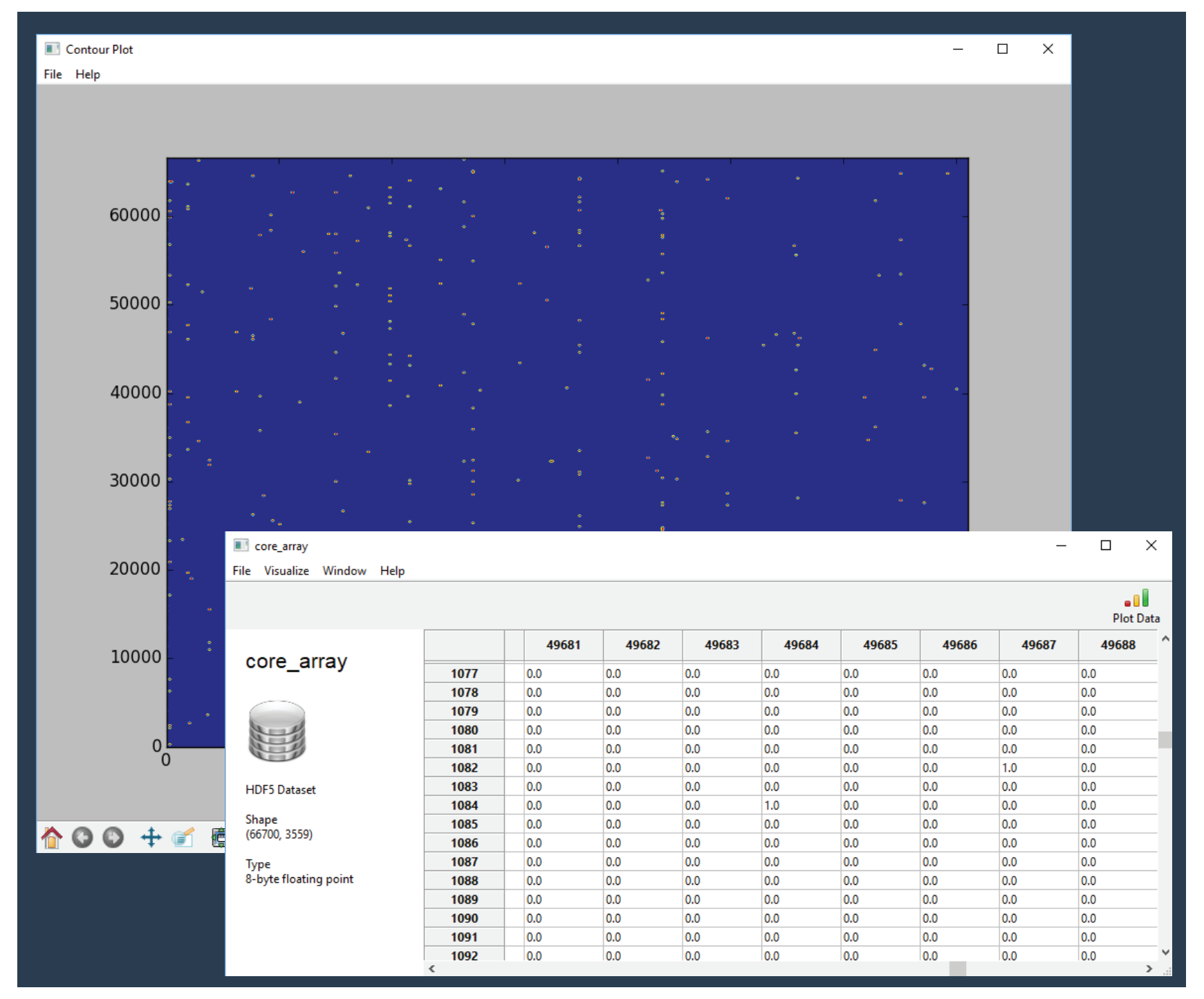
Database table



JSON Document



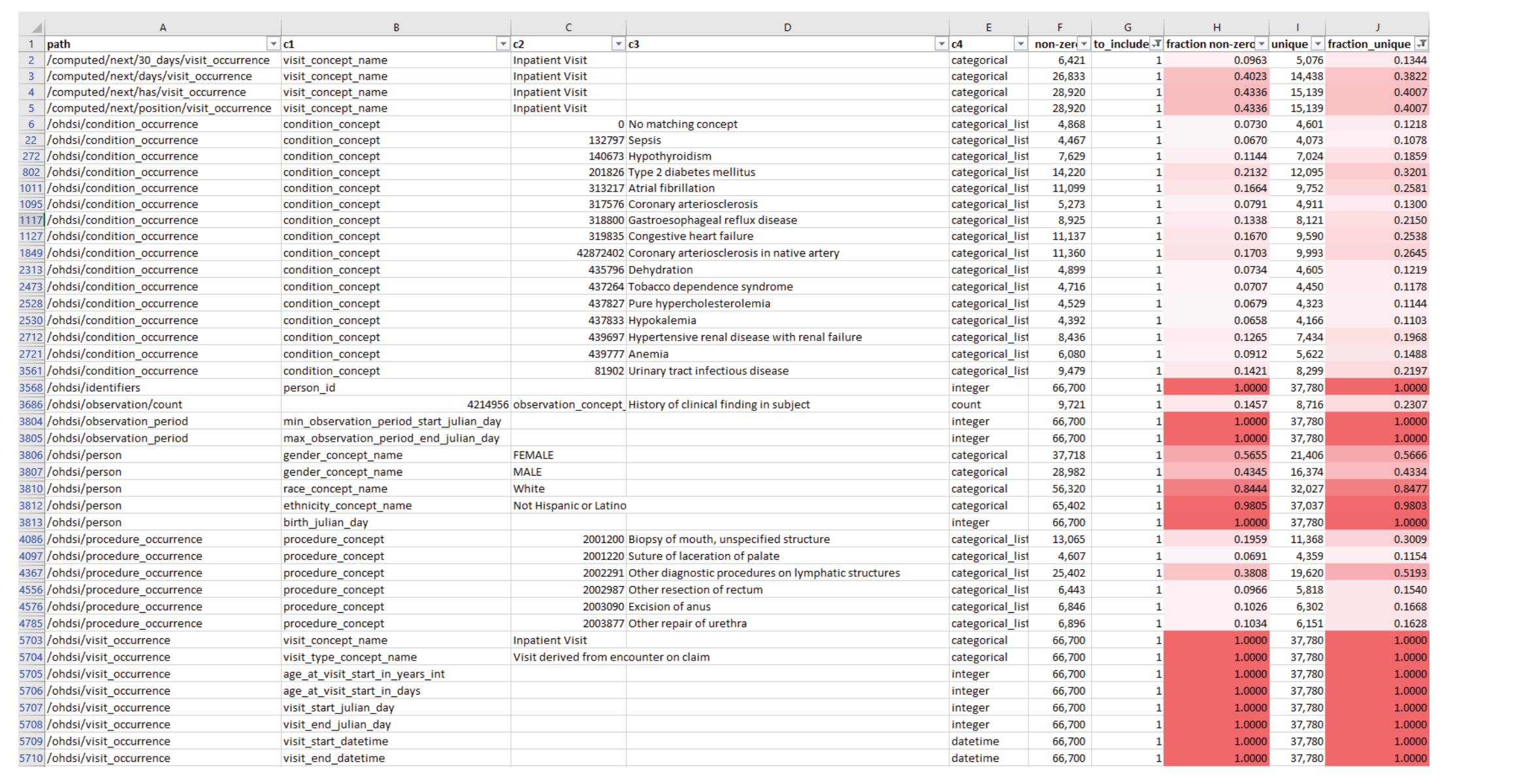
HDF5

Mapping Results

CDM Table	HDF5 Group	Number of columns	Non-zero elements
person	/ohdsi/person	9	268,370
visit_occurrence	/ohdsi/person	7	533,600
condition _occurrence	/ohdsi/condition _occurrence	3,559	410,535
procedure _occurrence	/ohdsi/procedure _occurrence	1,888	177,437
measurement	/ohdsi/ measurement	41	5,551
observation	/ohdsi/ observation	194	32,652

Variables occurring for 10% or greater of synthetic patients

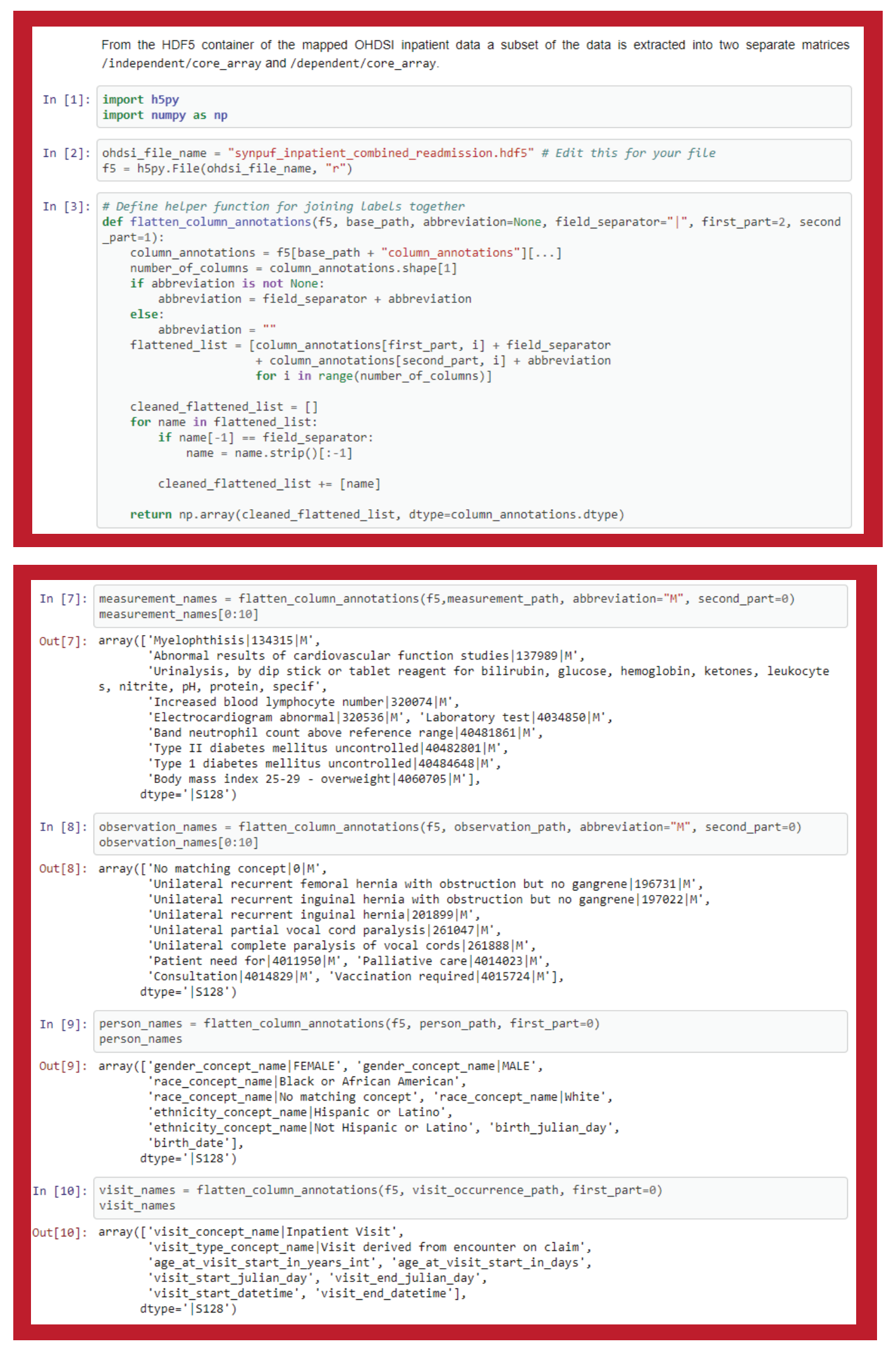
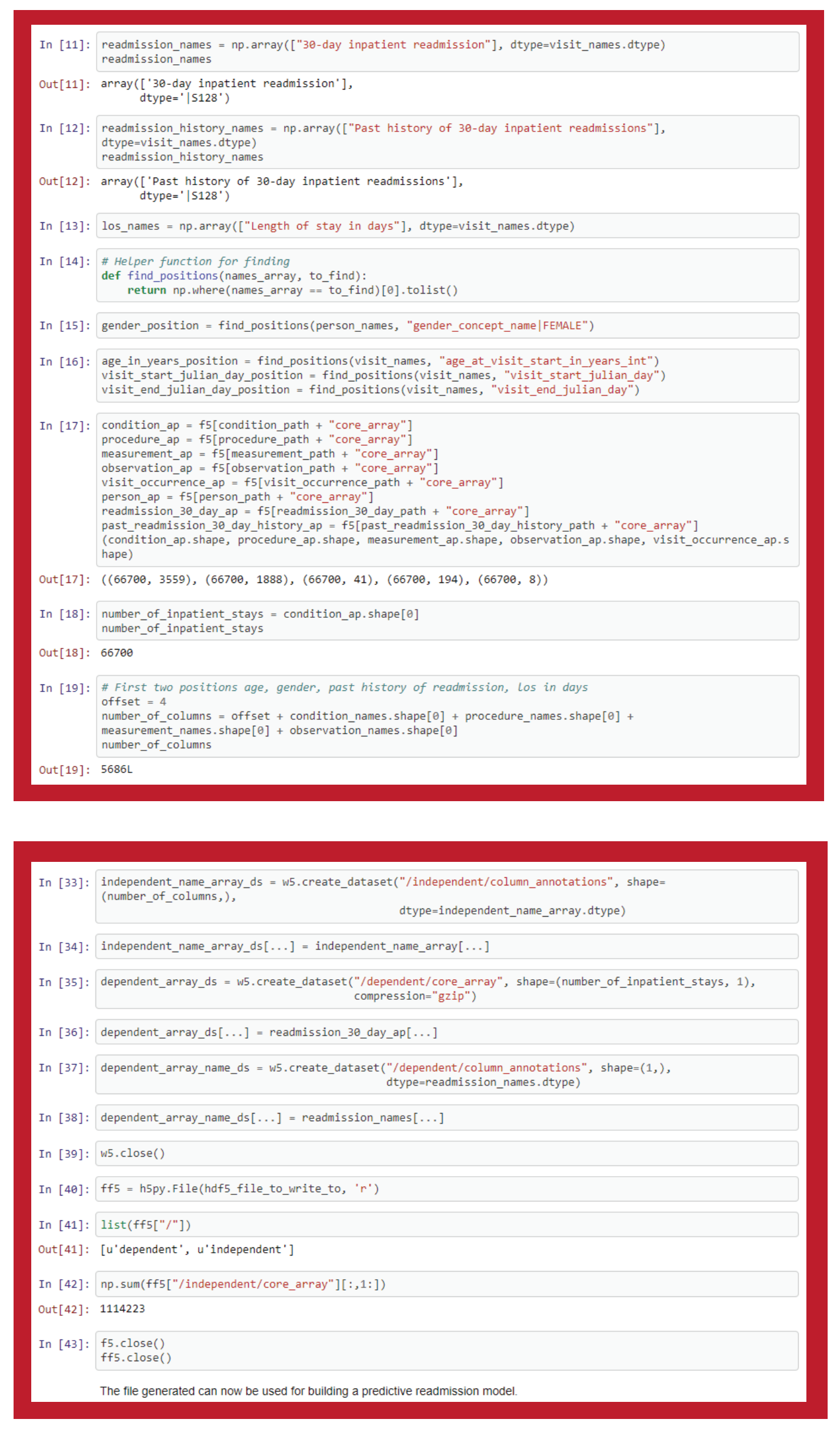


Conclusion:

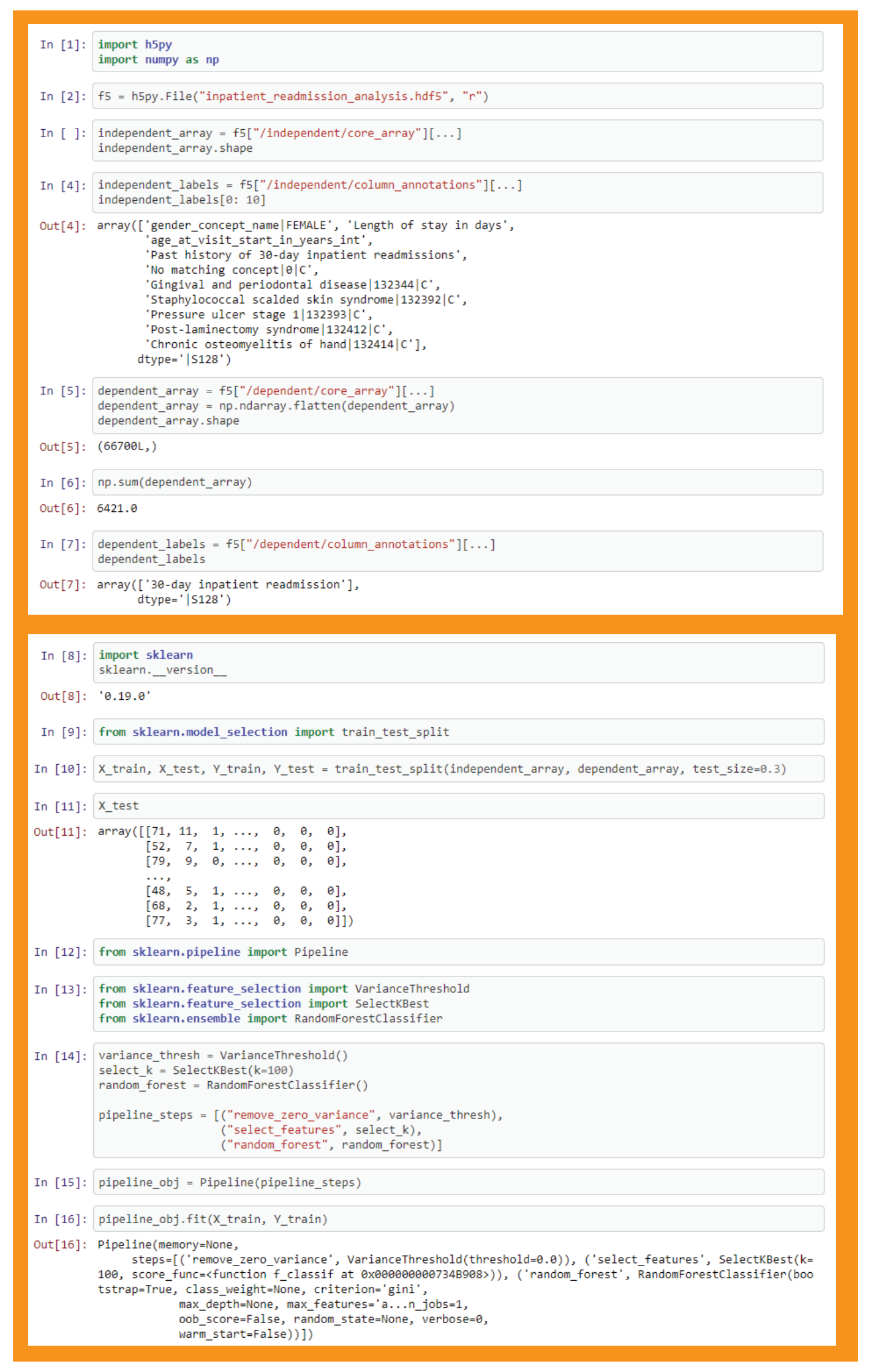
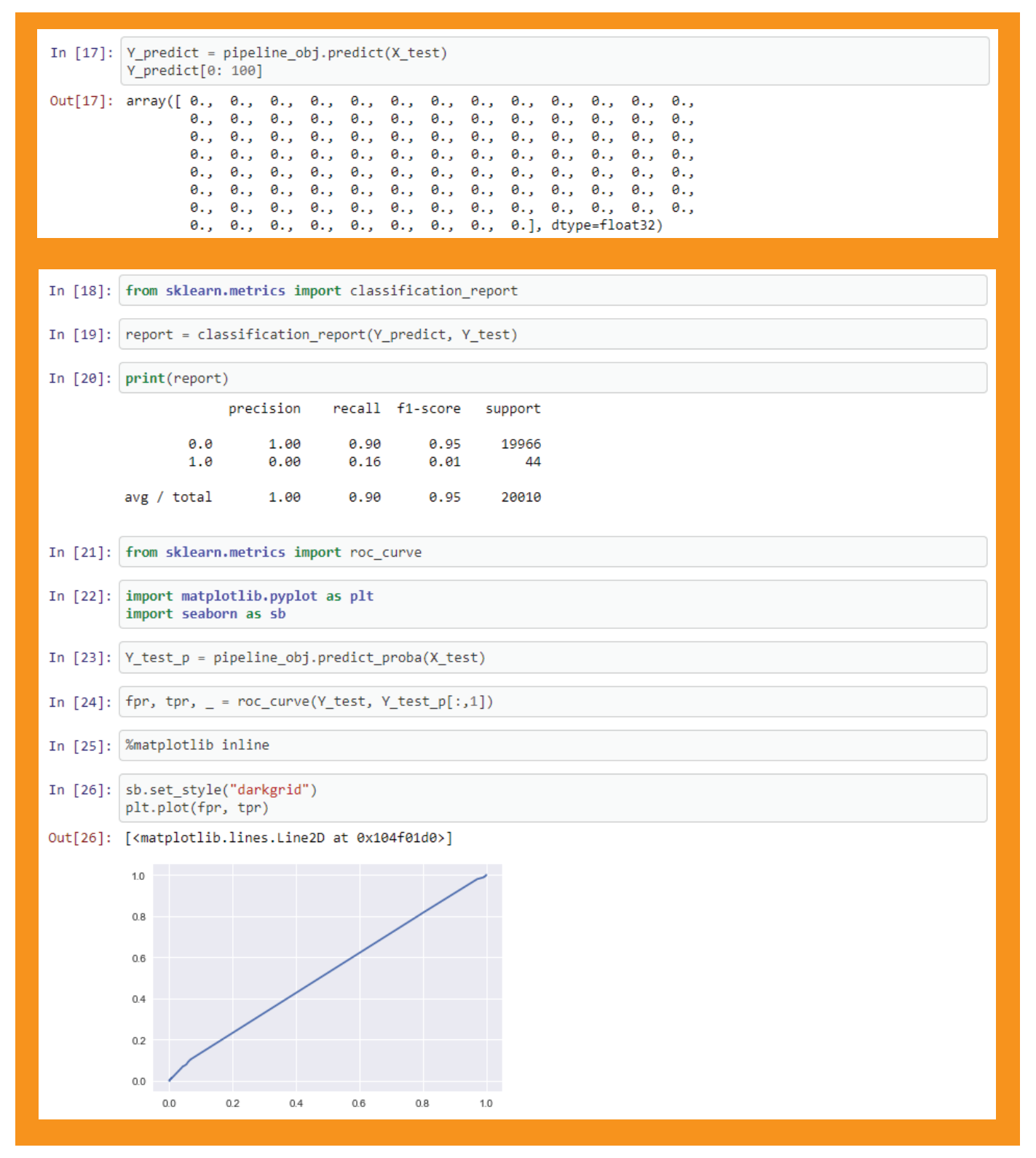
The approach described here shows that it is possible to run machine learning workflow against health care data of realistic size and complexity using data in the CDM as a source. It was not expected that the model would produce meaningful results due to the synthetic nature of the underlying data. By utilizing the CDM as the data model it removes some of the uncertainty in the data modeling process. Using the approach and tools described here it would be possible to train multiple models across different CDM datasets and compare the performance of 30-day predictive models and estimate the ability of the transference of the trained models across different datasets.

Jupyter Notebooks

Data processing

Building a predictive readmission model

Abstract:

Synthetic inpatient claims data in the OMOP Common Data Model were mapped to matrices in a HDF5 file. The content of the HDF5 were further manipulated to build a matrix for a 30-day post discharge readmission model. A random forest model was built to predict a patient's 30-day readmission risk. While the model's results were not predictive (AUC = 0.53) the modeling approach and pipeline can be applied to data in the Common Data Model (version 5.0). This opens up the possibility of rigorously comparing predictive performance of readmission models across different datasets.

Links:

TransformDBtoHDF5ML: https://github.com/jhajagos/TransformDBtoHDF5ML
MappingOHDSI2HDF5: https://github.com/SBU-BMI/MappingOHDSI2HDF5
Notebooks: https://github.com/SBU-BMI/MappingOHDSI2HDF5/tree/master/notebooks

Presented at the 2017 OHDSI Symposium (October 17th)
https://doi.org/10.6084/m9.figshare.5503813.v1

References:

1. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 2016, ocw042. http://doi.org/10.1093/jamia/ocw04
2. Miotto, R, Li, L., Kidd, BA, & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6, 26094.
3. Choi, E., Bahadori, MT, Schuetz, A, & Stewart, WF. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.Retrieved from https://arxiv.org/pdf/1511.05942.pdf
4. Fernando Pérez, Brian E. Granger, IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/-June 2007.