Rapid Generation of Synthetic Data to the OHDSI CDM

Janos G. Hajagos
Stony Brook University

Background

The OHDSI community has long supported the use of synthetic data for educational, engineering, and testing purposes. Synthea is a tool for generating synthetic data which allows the development of custom modules in a visual interface and allows the modelling of a disease process [1]. We have created a Docker container which accelerates the mapping of Synthea data to the OHDSI CDM. The conversion of Synthea to OHDSI CDM relies on an Apache Spark based mapper which uses an intermediary format called PSF (Prepared Source Format). To demonstrate the utility of the approach we generated synthetic COVID-19 infection data [2], mapped it to the OHDSI CDM, and successfully applied an existing cohort identification algorithm.

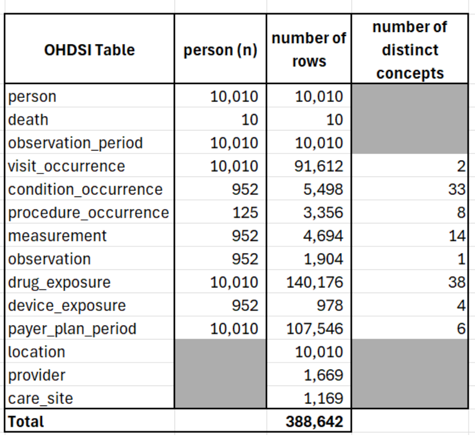
Methods

A 10,000 alive patient population was generated using the Synthea (version 3.2.0) COVID-19 module and outputted to a CSV file format. The data was then mapped to the intermediary PSF in the Apache Parquet Format and then mapped to the OHDSI CDM version 5.4 with the vocabulary release v20230831. The data was then loaded into a Microsoft SQL Server (MSSQL) 2022. The cohort selection script from the Cure-ID project [3] was applied to the population. The whole pipeline was executed in a Docker environment running on a Windows 11 Professional (64Gb of Ram) laptop. The instructions for running the pipeline and the Docker recipe are available on GitHub [4].

Results

The 10,000 patients generated in Synthea took 67 seconds, the mapping to PSF took 22 seconds, staging concept tables took 228 seconds, mapping to OHDSI CDM took 252 seconds, staging concept tables in MSSQL took 984 seconds, staging clinical tables took 20 seconds, and inserting data into OHDSI CDM tables took 243 seconds. Once the concept tables have been processed and staged the total pipeline time is 580 seconds (9.6 minutes) which is reduced from 1,816 seconds (30.2 minutes) for the initial run.

The Synthea tool generates demographically realistic population estimates based on location in the United States. In total the male to female ratio in the data is 4,860 males to 5,150 females. In total 952 synthetic patients tested as COVID-19 positive via documented PCR test results, with 125 being hospitalized, and out of those 8 being put on a mechanical ventilator. A total of 125 patients met population criteria for being hospitalized for COVID-19 as defined in the Cure-ID registry cohort selection algorithm.



OHDSI Table	person (n)	number of rows	number of distinct concepts
person	10,010	10,010	
death	10	10	
observation_period	10,010	10,010	
visit_occurrence	10,010	91,612	2
condition_occurrence	952	5,498	33
procedure_occurrence	125	3,356	8
measurement	952	4,694	14
observation	952	1,904	1
drug_exposure	10,010	140,176	38
device_exposure	952	978	4
payer_plan_period	10,010	107,546	6
location		10,010	
provider		1,669	
care_site		1,169	
Total		**388,642**	

Figure 1. Details on synthetic COVID-19 data generated and mapped to the OHDSI CDM

Conclusion

Synthea has a rich set of modules for generating synthetic patient data. It also has a web-acceeessible GUI tool for users to build and customize modules [5]. The focus of this work is to build an automated Docker pipeline to map PSF generated from Synthea. While an existing and well maintained ETL exists for Synthea [6] the alternative mapper was developed to document how the PSF can be used across a range of different sources. This work demonstrates that Synthea mapped to the OHDSI format can be used to test and debug a cohort selection algorithm. The combination of customizing Synthea and rapid mapping should lead to better engineered data pipelines for OHDSI CDM databases as software engineers do not need access to primary patient data and test data can be distributed with the pipeline.

References

1. Walonoski, J. et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J. Am. Méd. Inform. Assoc. 25, 230–238 (2018).
2. Walonoski, J. et al. SyntheaTM Novel coronavirus (COVID-19) model and synthetic data set. Intell Medicine 1, 100007 (2020).
3. https://github.com/OHDSI/CureIdRegistry
4. https://github.com/jhajagos/PreparedSource2OHDSI/tree/main/map/prepared_source/synthea/docker (2024)
5. https://synthetichealth.github.io/module-builder/
6. https://github.com/OHDSI/ETL-Synthea