Exporting and Running OHDSI Generated Cohort Definitions in a Secure Enclave

Janos Hajagos[1]
[1]**Stony Brook University**

Background

The OHDSI CDM (Common Data Model) is increasingly becoming the de facto standard for representing EHR (Electronic Health Record) data for research. Two large NIH (National Institutes of Health) sponsored projects the N3C[1] (National COVID Cohort Collaborative) and the All of Us Research Program[2] have adopted the OHDSI CDM for storing and representing multi-site EHR data. Due to the restrictive nature of these enclaves it has been difficult to adopt OHDSI tooling like ATLAS to work in these environments. The users of these enclaves have not had to the full benefit of tools like the Atlas's web interface and cohort generation to accelerate the speed and reproducibility of their work. This paper describes a simple manual method for exporting and running OHDSI tooling generated cohorts in the N3C enclave.

Methods

A subset of JSON (JavaScript Object Notation) cohort definitions were imported from the OHDSI Phenotype library and other online sources into a local ATLAS (v2.13). Cohorts were manually exported to the SPARK SQL dialect. The SQL was processed using a Python script with the sqlparse library[3]. The following changes were made: removing vendor specific SQL statements, lower casing the query, and using the sqlparse library to format the query in a more human readable format.

Using the Code Workbook (Figure 1) feature in the N3C enclave the linked code blocks for defining the cohort were manually generated. Above each translated query is the referenced table used by the query which guides the user to make sure all tables are referenced. Unlike a traditional database, table names are scoped at the cell level and need to be imported. Generated SQL were pasted into the code cell and executed. The transformed cohort definition was run against the entire N3C enclave on 6/15/2023 which included data from 78 sites and a total of 19.6 million individuals and 26 billion rows[4].

Results

A cohort definition for COVID inpatients was translated into a code workbook in the N3C enclave[5]. No structural changes were made to the SQL and it was executed against the entire N3C population. The cohort definition included evaluation of COVID test positivity and this required querying the measurement table (12.6 billion rows). Total time to create workbook was 30 minutes and query execution time was 36 minutes.

Figure 1. Translation of an exported JSON cohort from ATLAS into a Code Workbook in the N3C enclave environment.

Conclusion

The goal of this work was to determine the feasibility of translating cohort definitions into a format that can be run in the N3C secure enclave. The user did not need to directly translate the SQL or understand the generated code. This should allow non-technical users to develop clinically meaningful cohorts in Atlas and then export them to a format that a programmer can work with. This work builds on recent support for the SPARK SQL dialect by the OHDSI community. While the process here is manual it should be possible in the future to use a JSON intermediary format to accelerate the code translation.

References

1. Bennett, T. D., Moffitt, R. A., Hajagos, J. G., Amor, B., Anand, A., Bissell, M. M., Bradwell, K. R., Bremer, C., Byrd, J. B., Denham, A., DeWitt, P. E., Gabriel, D., Garibaldi, B. T., Girvin, A. T., Guinney, J., Hill, E. L., Hong, S. S., Jimenez, H., Kavuluru, R., … Rutter, J. (2021). Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data From the US National COVID Cohort Collaborative. *JAMA Network Open*, *4*(7), e2116901–e2116901. https://doi.org/10.1001/JAMANETWORKOPEN.2021.16901
2. Investigators, A. of U. R. P. *et al.* The “All of Us” Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
3. Hajagos, J (2023). Translate OHDSI Cohort to Enclave Definition Project https://github.com/jhajagos/TranslateOHDSICohortToEnclaveDef
4. N3C (2023). The National Covid Collaborative Dashboard (https://covid.cd2h.org/dashboard/) Accessed on 6/16/23
5. https://raw.githubusercontent.com/OHDSI/CureIdRegistry/main/cohort.json