OMOP Cohort creation and Deidentification Steps

The following scripts are to be run on a sites full OMOP dataset in order to prepare the relevant data for sharing with the VIRUS registry. Each script should be run on the same server as the OMOP data but can be customized to run on the preferred Database and Schema.

01 – Run the Cohort Creation Script

- Filename: 01_CURE_ID_Cohort.sql
- Purpose: This script creates a cohort of patients for the CURE ID registry. The patient list is saved in the cohort table, along with other useful data elements.

02 – Run the CURE ID Subsetting

- Filename: 02_CURE_ID_All_Tables.sql
- Purpose: This script takes your OMOP dataset and generates a copy of key tables that have been filtered down to only include people and records related to the CURE ID registry.

03 – Run the Deidentified Data DDL

- Filename: 03_DE_ID_CDM_Table_ddl.sql
- Purpose: Generate the necessary tables for the de-identified version of the CURE ID Cohort
- Customization: By default this script will create tables in a schema titled “deident,” however this can be set to whatever value you desire.

04 – Run the Deidentification Script

- Filename: 04_DE_ID_script.sql
- Purpose: This script creates a copy of the Cohort and removes identifying characteristics to prepare the data for sharing with the VIRUS registry. The steps taken are as follows:
  - Reassignment of Person IDs: Person IDs are regenerated sequentially from a sorted copy of the Person table. These new Person IDs are carried throughout the CDM to all tables that reference it.
  - Date Shifting:
    - Each person is assigned a random date shift value between -186 and +186 days. All dates for that person are then shifted shifted by that amount.
    - Birthdays: After date shifting a person’s birthday, the day is then set to the first of the new birth month. If the person would be > 89 years old then they are assigned a random birth year that would make them 90-99 years old.
  - Date Truncation:
    - A user-defined Start and End date are used to exclude any date shifted data that falls outside of the target date range (E.G. Procedures, conditions occurrences, etc. Does not include Birthdates).
  - Removal of other identifiers:
    - Other potentially identifying datapoints are removed from the dataset such as `location_id`, `provider_id`, `and care_site_id`

05 – Run the Quality Checks (optional)

- Filename: 05_DE_ID_Quality_Checks.sql
- Purpose: This script checks basic metrics for each table in the deidentified dataset to ensure the previous scripts were successful. This does require a human to read the results and does not have built in pass/fail checks.