Siriraj Informatics and Data Innovation Center

Enabling i2b2 on OMOP CDM Cohort Data semi-automatically by using Atlas and SQLMesh

Presenter: Natpatchara Pongjirapat Email: natpatchara.pon@mahidol.edu

Background

Informatics for Integrating Biology and the Bedside (i2b2) is an open-source software platform widely used in clinical research for querying and analyzing de-identified patient data. In December 2023, i2b2 version 1.8.0 expanded its support for the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), a standardized format for observational health data.

Our institution utilizes i2b2 for research feasibility assessments of our Electronic Health Record (EHR) data, which is formatted in OMOP CDM. We have also developed a data pipeline that restricts data access to specific cohorts. This approach enables a semi-automated workflow for creating OMOP data cohorts and integrating them into i2b2.

Method

Our framework utilizes the OMOP Common Data Model (CDM) for initial analysis. To begin, we define and construct a cohort. This cohort consists of three columns: patient number, cohort start time, and cohort end time. Researchers can define cohorts within Atlas. Once the query is obtained, the code is executed to establish the cohort within our database.

Following cohort creation, we develop an i2b2-compatible view. We achieve this by utilizing i2b2-on-OMOP and the ENACT Ontology V4.1. To ensure adaptability for future data model changes, we implement SQLMesh to generate corresponding data models.

```
MODEL (
  name i2b2.condition_view,
  kind VIEW
);

SELECT
  visit_occurrence_id AS ENCOUNTER_NUM,
  PERSON_ID AS PATIENT_NUM,
  condition_concept_id::VARCHAR(50) AS CONCEPT_CD,
  COALESCE(provider_id::VARCHAR(50), '@'::VARCHAR) AS provider_id,
  condition_start_datetime AS START_DATE,
  condition_end_datetime AS END_DATE,
  '@'::VARCHAR(100) AS modifier_cd,
  1 AS INSTANCE_NUM,
  NULL::VARCHAR(50) AS valtype_cd,
  NULL::VARCHAR(50) AS location_cd,
  NULL::VARCHAR(255) AS tval_char,
  NULL::DECIMAL AS NVAL_NUM,
  NULL::VARCHAR(50) AS valueflag_cd,
  NULL::VARCHAR(50) AS units_cd,
  NULL::DOUBLE PRECISION AS CONFIDENCE_NUM,
  NULL::VARCHAR(50) AS SOURCESYSTEM_CD,
  NULL::TIMESTAMP AS UPDATE_DATE,
  NULL::TIMESTAMP AS DOWNLOAD_DATE,
  NULL::TIMESTAMP AS IMPORT_DATE,
  NULL AS OBSERVATION_BLOB,
  NULL::INT AS upload_id,
  NULL::INT AS quantity_num,
  condition_SOURCE_concept_id AS SOURCE_CONCEPT_ID,
  condition_source_value AS SOURCE_VALUE,
  'CONDITION' AS DOMAIN_ID
FROM
  omop.CONDITION_OCCURRENCE
WHERE
  PERSON_ID IN
  SELECT PERSON_ID FROM omop.dm_cohort;
```

Figure 1: SQLmesh code example for create i2b2 data models.

SQLMesh is an open-source data transformation framework that enables version control and automated schema evolution for SQL-based data models. This facilitates the creation and updating of cohort-specific schemas, allowing for flexible handling of different cohorts and streamlining the process of integrating new data models into our i2b2-compatible views.

Additionally, we automate the creation and population of i2b2-compatible schemas using a Python script.

Result

Our methodology leverages i2b2 for user-friendly data querying and Atlas, in conjunction with Python and SQLMesh, for cohort generation. In our experiments, Atlas demonstrated its versatility as both a cohort definition tool and a SQL code generator. When combined with SQLMesh's ability to manage multiple data environments, this proved effective in filtering OMOP data into distinct patient groups.

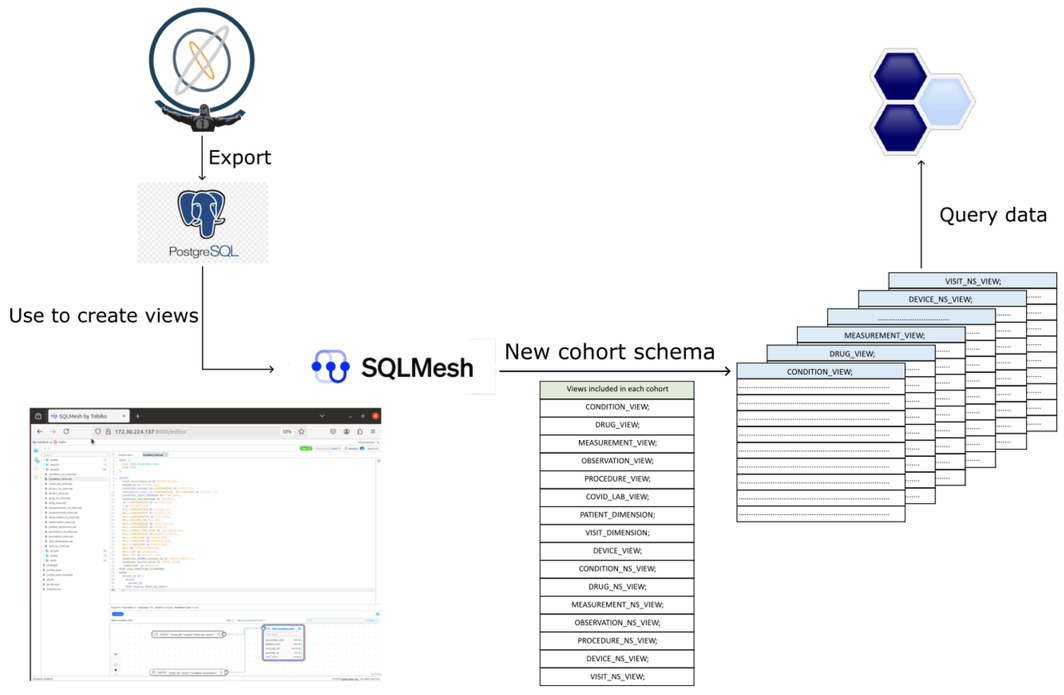

Figure 2: Overview of our proposed framework

By isolating each cohort into separate environments, i2b2 can be configured to access cohorts on an as-needed basis. This approach ensures controlled data access while minimizing the overhead associated with data provision. The integration of Python further enhances the reproducibility of the process, reducing costs for future cohort generation.

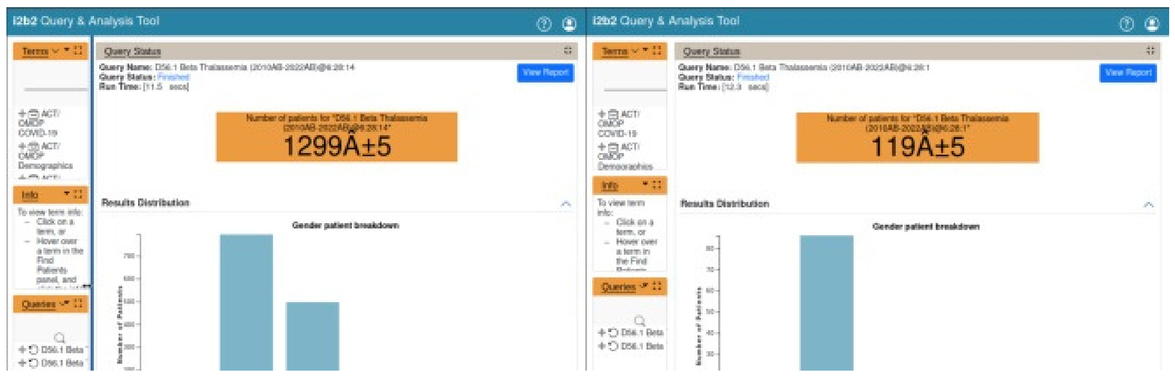

Figure 3: Example cohort comparing query performed on our de-identified data and on subset of our data. left: the query performed on the whole de-identified data, right: the query performed on the demo cohort.

The diabetes cohort, serving as a case study for our framework, demonstrates the potential of this approach. The varying query results validate the effectiveness of our method in separating data based on the defined cohorts.

Despite its promising potential, the proposed framework is still in its early stages, and further testing is needed to evaluate its advantages.

Our experiments also revealed certain challenges, particularly concerning the adaptation of the OMOP ENACT Ontology to our data.

All code used will be released in the future in our github.
To visit scan the qr code or go to https://github.com/sidataplus