Evaluating the Conversion of EHR data into OMOP CDM for Type 2 Diabetes Mellitus Cohort: Insights for Data Consistency

Siriraj Informatics and Data Innovation Center

Presenter: Burin Boonwatcharapai

Email: burin.boo@mahidol.edu

Background

In observational research, establishing well-defined cohorts based on phenotypes is a critical step to ensure data quality for subsequent analyses. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) offers a promising solution by standardizing data structure, simplifying access, and enabling multi-institutional collaborations. However, concerns remain regarding the quality of data transformation from hospital electronic health records (EHR) to the OMOP CDM.

Building on the methodology of Candore et al. [1], which involves comparing analysis results between established cohorts and those generated from the OMOP CDM, our study aims to validate the quality of this data transformation. Specifically, we leverage our previously created Type 2 diabetes mellitus (T2DM) cohort from EHR data prior to its conversion to the OMOP CDM.

Method

The study was conducted at Siriraj Hospital, an academic health center in Bangkok, Thailand. Its EHR data were recently transformed into the OMOP CDM. To evaluate the quality of this data transformation, we created two cohorts: one derived from the original EHR database and the other from the OMOP CDM.

We applied the same T2DM cohort definition to both datasets, including patients aged 18 and above who were identified using ICD-10 diagnosis codes, laboratory values, or prescriptions for diabetes medications. The cohort spanned from June 1, 2013, to September 30, 2023. These criteria were selected for their relevance in assessing data quality across multiple domains. For patients meeting multiple criteria, the date of the first occurrence was used as the inclusion date. Detailed inclusion and exclusion criteria are presented in Figure 1.

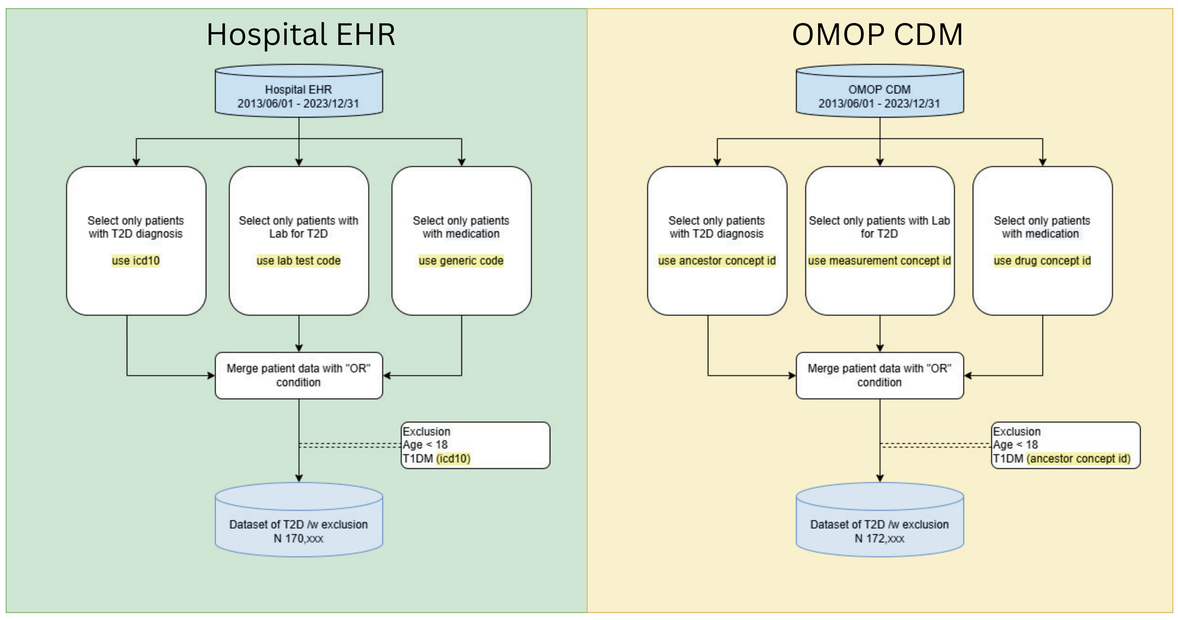

Figure 1: Inclusion & Exclusion criteria for Siriraj Type 2 Diabetes Mellitus cohort

The two cohorts were generated using SQL queries on an MSSQL database. We compared the number of patients meeting each criterion in both the original and OMOP-transformed datasets (Figure 2). Key outcomes following the inclusion date were also evaluated, with cumulative incidences calculated and differences summarized (Figure 3).

Result

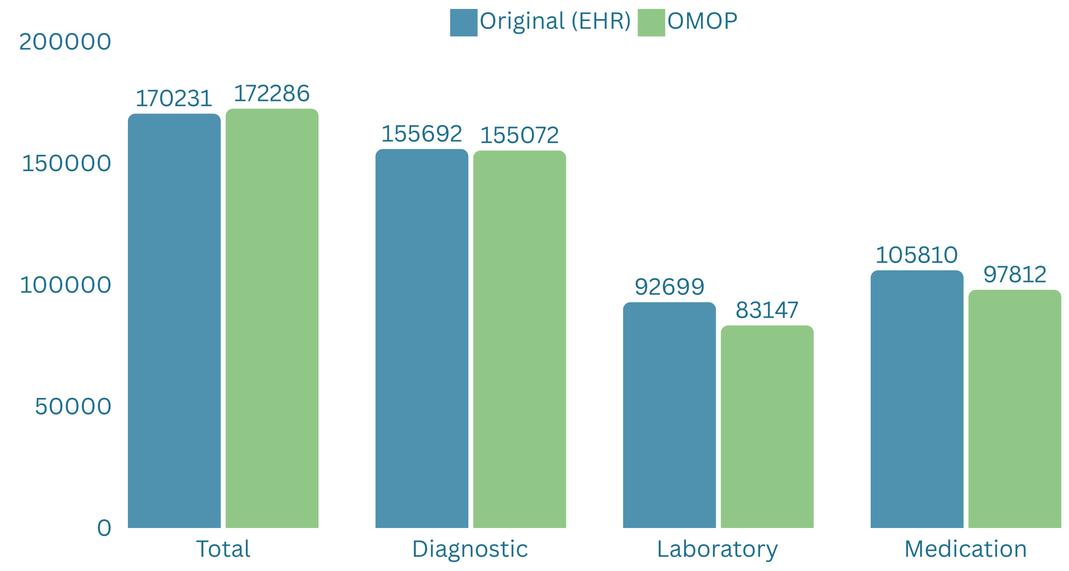

Figure 2: Comparison of Patients Meeting Inclusion and Exclusion Criteria in Original EHR database and OMOP Cohorts

The total number of patients in the original cohort was 170,231, compared to 172,286 in the OMOP cohort, representing a +1.207% difference. Figure 2 presents the frequency of each criterion met.

Overall, most outcomes showed marginal variations between datasets, with differences ranging from -7.65% to +10.71%. Discrepancies between cohorts stemmed from three main factors:

1. **Vocabulary Differences:** Variations in code structures and concepts complicated cohort definitions. For example, identifying T2DM required generalizing ICD-10 codes starting with “E11*,” whereas SNOMED-CT required four specific codes with complex hierarchical logic.
2. **Vocabulary Mapping Challenges:** Differences in laboratory and medication data arose from transitioning Siriraj Hospital’s local codes to OMOP standard vocabularies (e.g., LOINC, RxNorm). Our goal of mapping 95% of transaction frequency left some codes unmapped, which impacted data consistency. Completing the vocabulary mapping process will address this limitation.
3. **Differences in Starting Time Points Between Datasets:** In our OMOP transformation process, we included data from 2010 onward, whereas the EHR system recorded diagnosis data starting in 2000. When applying Type 1 Diabetes exclusion criteria, we filtered out patients with any diagnosis of T1DM across the entire dataset. This time range discrepancy contributed to differences in patient numbers.

Outcome Comparison Between Original and OMOP		
Outcome	CI difference ($CI_{omop} - CI_{original}$)	% Difference
Diabetic Retinopathy	+0.23	+2.53%
Chronic Kidney Disease	-0.59	-2.98%
Cardiovascular Disease	-0.86	-5.31%
Osteoporosis	-0.40	-4.71%
Bone Fracture	-0.11	-3.77%
Peripheral Vascular Disease	-0.15	-2.61%
Essential Hypertension	-2.07	-5.73%
Dyslipidemia	-2.76	-7.65%
Death	+0.48	+10.71%

*CI: Cumulative Incidence

Figure 3: Outcome Comparison Between Original and OMOP datasets with Percentage Difference

The outcome comparison showed small differences in cumulative incidence (CI) for each outcome (Figure 3). Notable findings include:

- Mortality: Largest positive difference (+0.48% CI, representing a 10.71% relative increase in the OMOP dataset)
- Dyslipidemia: Reduction of 2.76% CI (7.65% relative decrease)
- Essential hypertension: Reduction of 2.07% CI (5.73% relative decrease)
- Other outcomes: Slight declines in incidence rates

Burin Boonwatcharapai, Krittaphas Chaisutyakorn and Natthawut Adulyanukosol
Siriraj Informatics and Data Innovation Center, Siriraj Hospital, Mahidol University, Thailand