An evaluation of the transformation of a large German EHR database to the OMOP CDM

Andreas Ochs[1], Milou Brand[2], Jack Brewster[2], Methodios Typou[1], Meda Sandu[1], Joe Maskell[1], Meghan Pettine[2], Atif Adam[2], George Kafatos[1]
(1) Amgen Ltd., Uxbridge, United Kingdom; (2) IQVIA, London, United Kingdom

Background

The utilization of real-world data from electronic health records (EHR) is essential in advancing healthcare research and improving patient outcomes. In Germany, the Germany Disease Analyzer (GDA) database includes comprehensive EHR data from general practitioners and specialists. Homogenizing such databases is becoming pivotal in conducting large multi-country observational studies. This study aims to evaluate the quality of transforming the GDA database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format. By validating the conversion and mapping of medical native codes to standard OMOP concepts, we can ensure the harmonized version of GDA can be utilized alongside other OMOP data sources to produce robust real-world evidence.

Methods

Study Design and Study Population

This is a retrospective cohort study leveraging data from the GDA database. The GDA data covers approximately 40.9 million individuals from a representative sample of around 1,300 GPs and 1,200 specialists in Germany (around 4-5% of the German population).

Study Period

The study period spans from January 1, 2013, to September 30, 2022, analyzed in 2-year periods.

Patient Eligibility

Patients aged 18 years or older at the index date were included.

Index Date and Baseline Period

The index date for each patient and each confirmed condition was defined as the earliest record of the relevant condition, where the patient satisfies all inclusion (i.e., 18 years old) within each 2-year period. The baseline period extends from one year before the index date up to and including the index date. This study is cross-sectional, with no follow-up beyond the index date.

Studied Conditions

Using a string search of medical terms we created ICD-10 code lists for the following conditions of interest:

- myocardial infarction (MI)
- stroke
- type 2 diabetes mellitus (T2DM)

Data Analysis

The primary goal of this study is to assess the quality of the transformation of the native GDA database and its native codes into the OMOP CDM and non-standard and standard concepts. This involves calculating and comparing code and patient counts, demographics, and relevant clinical characteristics of selected disease groups at different transformation stages (native codes, non-standard concepts, standard concepts).

The quality of the transformation was assessed along three metrics:

1. the number of diagnosis codes identified at each transformation stage,
2. the number of unique diagnosis codes at each transformation stage, and
3. the number of patients identified with defined conditions at each stage.

Each metric was utilized to assess the concordance at the different stages of the transformation:

- Concordant Results: indicate the full overlap between the native codes and OMOP concepts (non-standard or standard).

- Discordant Native Code Results: indicate results are present using native codes but not by OMOP concepts (non-standard or standard).
- Discordant OMOP Concept Results: indicate results are present using OMOP concepts (non-standard or standard) but not by native codes.

Statistical analysis was performed using SAS, SQL, and R (version 4.2.1).

Results

Before embarking on comparisons between the different transformation stages, a small number of fictitious patients were removed from the source GDA dataset.

The results vary between the conditions studied (Table):

- **MI:** Perfect concordance was observed across all 2-year increments for all three metrics and between both transformation steps from native codes to non-standard and non-standard to standard concepts.
- **Stroke:** Perfect concordance was observed for all 2-year increments for the two metrics of unique number of codes and number of patients for the transformation step from native codes to non-standard concepts. Minor discordance was observed across all 2-year increments for the metric of number of diagnosis codes between native codes and non-standard concepts. This discrepancy merits further investigation, which might also have impact on the transformation between non-standard and standard concepts. Therefore, no results are reported for this transformation step. Perfect concordance was observed across all 2-year increments for the number of patients from non-standard to standard concepts.
- **T2DM:** Perfect concordance was observed for all 2-year increments for the two metrics of unique number of codes and number of patients for the transformation step from native codes to non-standard concepts. Minor discordance was observed across all 2-year increments for the metric of number of diagnosis codes between native codes and non-standard concepts. This discrepancy merits further investigation before results for this metric also can be shared for transformations between non-standard and standard concepts. Minor discordances for all 2-year increments were observed between non-standard and standard concepts in the metrics of number of patients and number of unique codes.

Discussion

Continued Work

Ongoing efforts are focused on evaluating the discordant results observed in stroke and T2DM, specifically analyzing the factors contributing to these variabilities. Tentative conclusions can be suggested for the observed discordance in T2DM at the stage of transformation from non-standard to standard concepts: this is likely observed due to the mapping of the non-standard concepts to standard concepts when it comes to differentiating between Type 1 Diabetes Mellitus and T2DM.

Impact of Work

The transformation of the GDA EHR database into the OMOP CDM format demonstrates high consistency, completeness, and clinical plausibility across the analyzed conditions at all transformation stages and metrics, though more investigation is required for some transformation stages. The results showed minimal discrepancies, which varied depending on the condition studied and the stage of transformation. These findings validate the effectiveness of the OMOP CDM for harmonizing EHR datasets in Germany, ensuring accurate and comparable data for health outcomes research. The robust transformation enables researchers to confidently utilize

harmonized data for observational studies, facilitating improved healthcare strategies and patient care.

Table: results of transformation by metric, transformation stage and condition

	Number of diagnosis codes		Unique number of diagnosis codes		Number of patients	
	Native to non-standard	**Non-standard to standard**	**Native to non-standard**	**Non-standard to standard**	**Native to non-standard**	**Non-standard to standard**
Myocardial infarction	Perfect match	Perfect match	Perfect match	Perfect match	Perfect match	Perfect match
Stroke	Minor discordance	No results shown[1]	Perfect match	Perfect match	Perfect match	Perfect match
Type 2 Diabetes	Minor discordance	No results shown[1]	Perfect match	Minor discordance	Perfect match	Minor discordance

[1] results are not shown as further investigation of minor discordance at native to non-standard transformation stage is required

Conflict of interest statements and financial disclosure

AO/MS/JM/GK are Amgen employees/shareholders.
MT is contracted employee of Amgen.
MB/JB/MP/AA are IQVIA employees.