Process of Conversion of Ukrainian Medical Data to OMOP CDM Format

Bohdan Khilchevskyi, MD[1,2], Denys Kaduk, MD[1,4],Maksym Trofymenko, MD[1], Polina Talapova, MD, PhD[1,2], Tetiana Nesmiian MD, LLM[1,3], Max Ved[1], Inna Ageeva[1], Pavlova Olga, MD, PhD[4], Holovko Tetiana, PhD[4,5], Shevchenko Natalia, MD, PhD, DSc[4,5]

[1] IT company SciForce, Kharkiv, Ukraine
[2] Kharkiv National Medical University, Kharkiv, Ukraine
[3] Kharkiv National Pedagogical University named after H. S. Skovoroda, Kharkiv, Ukraine
[4] V. N. Karazin Kharkiv National University, Kharkiv, Ukraine
[5] Cardiorheumatology Department, State Institution "Institute of Health Protection of Children and Adolescents of the National Academy of Medical Sciences of Ukraine", Kharkiv, Ukraine

Background

Standardizing healthcare data is essential for enabling large-scale, multinational research and improving patient outcomes. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) provides a comprehensive framework for harmonizing diverse healthcare datasets. While Ukrainian medical data is still developing and undergoing digitalization, converting it into the OMOP CDM format is a crucial step towards enhancing data interoperability and supporting international collaborative research.

Objective

This project aims to convert Ukrainian medical data, characterized by its unique structure and regional nuances, into the OMOP CDM format. The goal is to demonstrate the feasibility of integrating this data into the Observational Health Data Sciences and Informatics (OHDSI) network, thereby enhancing data interoperability and supporting international collaborative research.

Methods

The process involved two main components:

Data Mapping: Identification and mapping of Ukrainian medical terminologies and codes to OMOP standardized vocabularies[1] such as SNOMED, LOINC, and UCUM. This ensures data consistency and accuracy for analysis. The mapping process itself was multifaceted, considering the data's origin and the presence of unique regional terms. Key steps included:

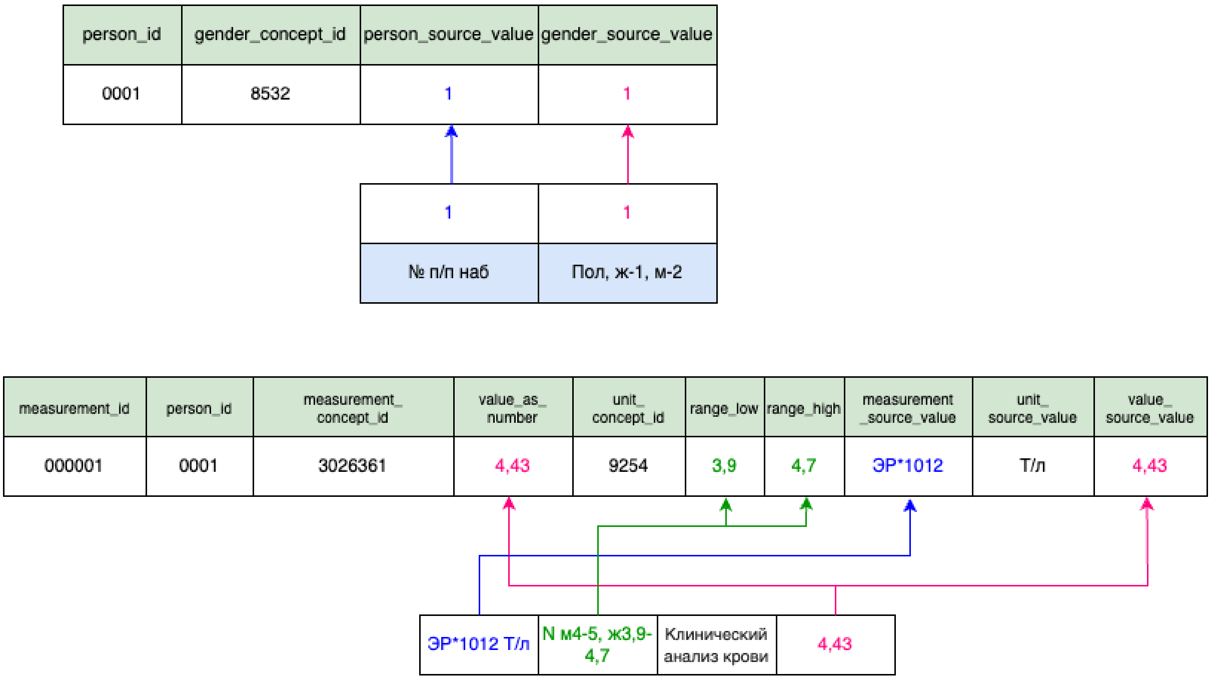
- **Translation and Abbreviation Expansion:** The original terms in the native language were translated into English by Google Translate and validated by clinical experts. Abbreviations were expanded during translation to improve clarity. For example, "КА" became "CA (coefficient of atherogenicity)" and "ФВпж" became "Right ventricular ejection fraction."
- **Data Structure and Categorization:** We worked with a column named "translated_source_name" containing the translated medical terms. To enhance precision, we added two new columns:
  - **source_code:** This assigns a unique code (UKR_Data_001 to UKR_Data_181) to each original term for easy reference.
  - **category_translation:** This categorizes terms into meaningful groups like "Echocardiography," "Serological Examination of Blood," or "Ultrasound of Knee Joints." This distinction is crucial for concepts with identical names but different

meanings. For instance, "Examination Result" could refer to various ultrasound examinations (liver, gallbladder, joints). Categorization allows for accurate identification of the specific examination type. We even created new categories not present in the original data, such as "Coagulogram," "Symptoms," and "Objective Status."

- **Mapping with Standardized Vocabularies:** Our primary focus was on selecting the most appropriate standard concept, but we also prioritized using the most widely recognized vocabularies for optimal flow and interoperability:
    - **SNOMED CT:** Used for concepts related to conditions, observations, anatomical sites, etc.
    - **LOINC:** Utilized for measurable quantities that came with specific units of measurement embedded within the source data (e.g., weight, heart rate, lab values).
    - **UCUM:** Provided standardized units of measurement when necessary.
- **SSSOM Mapping Precision Metadata**[3]**:** The mapping process employed specific columns to ensure thorough standardization and analysis:
    - **predicate_id:** This indicates the relationship between the original term and the mapped term in the standard vocabulary. It uses values like "skos:exactMatch" for perfect matches, "skos:broadMatch" for broader concepts, and "skos:noMatch" if no suitable mapping exists.
    - **confidence:** This reflects the certainty of the mapping, highlighting the degree of match between the original and standardized terms. It is represented by a score between 0.0 and 1.0, where 1 denotes total confidence, 0.8 indicates 80% of confidence, and 0.5 signifies 50% of mapping certainty.

ETL (Extract, Transform, Load) Process: The development of ETL pipelines to handle unstructured data is always a thought process. It was considered to develop ETL using CDM 5.4 version[2]. Basically, the source is a table where the field name is considered in most cases as measurement, observation, condition, or procedure details, and the value of the field are treated as a result. The process of ETL is provided in a Figure 1

Figure 1. ETL illustration



person_id	gender_concept_id	person_source_value	gender_source_value
0001	8532	1	1

1	1
№ п/п наб	Пол, ж-1, м-2

measurement_id	person_id	measurement_ concept_id	value_as_ number	unit_ concept_id	range_low	range_high	measurement _source_value	unit_ source_value	value_ source_value
000001	0001	3026361	4,43	9254	3,9	4,7	ЭР*1012	Т/л	4,43

ЭР*1012 Т/л	N м4-5, ж3,9-4,7	Клинический анализ крови	4,43

Results

The conversion successfully standardized a small dataset of Ukrainian medical data, including patient demographics, diagnoses, and procedures. The process uncovered several challenges, such as dealing with non-standardized local codes and addressing data gaps. As a continuous process, the next step is to create a proper validation process with data visualization for further refinement.

This detailed mapping approach ensures the accuracy and consistency of the Ukrainian medical data, preparing it for in-depth analysis using tools like OHDSI ATLAS[4]. This standardized data allows for robust findings and valuable contributions to international digital medicine conferences.

Conclusion

This case demonstrates the feasibility and benefits of converting regional healthcare data to the OMOP CDM format. The standardized Ukrainian dataset can be integrated to the OHDSI research network and can be an example for Ukrainian data holders. Future efforts will focus on refining the process and we are hoping to expand the scope to include more healthcare institutions across Ukraine.

References

1. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc. 2024 Mar;31(3):583-590. doi: 10.1093/jamia/ocad247.
2. OHDSI. OMOP CDM v5.4. Available from: https://ohdsi.github.io/CommonDataModel/cdm54.html. Accessed 21 June 2024.
3. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database (Oxford). 2022; doi: 10.1093/database/baac035.
4. OHDSI. ATLAS: A collaborative web-based tool for cohort building, visualization, and data analysis. Available from: https://github.com/OHDSI/Atlas. Accessed 21 June 2024.