A Data Integration Platform Exploiting the OMOP CDM: An Application in Molecular Cardiology

Elisa Maria Zini, PhD[1], Valentina Tibollo, MS[2], Matteo Gabetta, PhD[3], Andrea Mazzanti, MD[2], Carlo Napolitano, MD, PhD[2], Silvia G. Priori, MD, PhD[2], Riccardo Bellazzi, PhD[1,2]

[1] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[2] IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Italy
[3] Biomeris s.r.l., Pavia, Italy

Background

The use of practice-based data to inform research is one of the main pillars for enabling precision medicine in the Learning Healthcare System (LHS) cycle[1].

- Since 2000: the **Molecular Cardiology Unit (MCU) of Istituti Clinici Scientifici Maugeri (ICSM)** in Pavia has developed a **pathology registry** to improve the knowledge on genetic diseases causing life-threatening arrhythmias in a structurally normal heart (**inherited arrhythmogenic diseases**). The MCU is also using a **Laboratory Information Management System (LIMS)**[2] for the results of the genetic analyses.
- Currently: upgrade of the MCU platforms for the management and visualization of the information coming from clinical and research activities, to provide more coherent and consistent views to clinicians and researchers.

Goal of this work: integration of the different data sources used in the MCU into the OMOP CDM format (version 5.3.1) to guarantee the reproducibility of the methodologies and to adhere to the FAIR Data Principles[3].

Data Sources and Transformation to OMOP CDM

The main data source includes more than 10000 patients, with an average of 20 records of diagnoses and observations (e.g., cardiac events) per patient, and more than 34000 visits.
The data sources used in the MCU are shown in Figure 1, together with the workflow for data integration and transformation into the OMOP CDM.

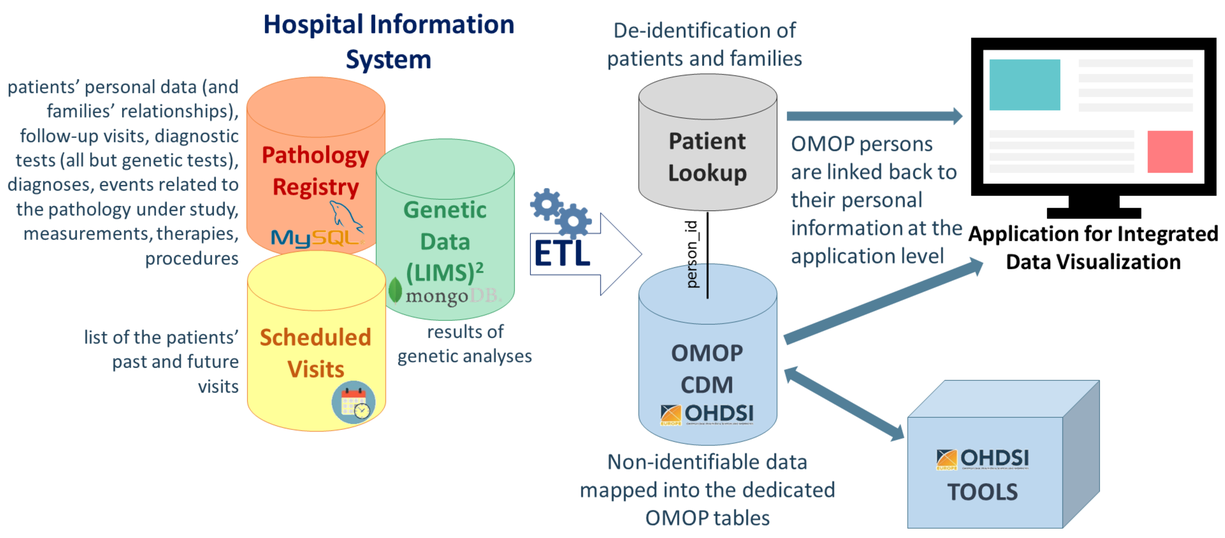

Figure 1. Workflow for data integration and transformation into the OMOP CDM.
ETL routines are expected to run on a weekly basis.

Work in progress:

- scheduled visits → OMOP *visit_occurrence* table
- genetic data → Genomic CDM extension (G-CDM) proposed by the Genomic CDM Subgroup[4]

Results and Challenges

The results of the transformations are shown in Table 1. We faced some challenges during the process:

- Proprietary vocabularies in data sources → manually mapped to the standard vocabularies.
  Issue: some concepts that appear in the same source table belong to different OMOP domains.
- Necessity to include negative results: in all our sources an absence of records implies lack of knowledge, while negative results are stated as such → partially solved using observations and their qualifiers.
- Cases of information loss[5]:
  - birth date not available for 68 patients out of 10170 → these patients were excluded from the import procedures; **10102 patients were included in the OMOP CDM.**
  - patients with status = "dead" but without a date of death were included, but without details about their death.

Other achievements so far: identification of missing data and inconsistencies in the source data (such as events registered before the date of birth or after the date of death) running Achilles on our OMOP-compliant database → inclusion of data preprocessing steps in the ETL, ensuring uniform data cleansing for subsequently collected data

Source data	# records	OMOP CDM Table	# records (%)
Family of a patient (family ID)	10102	observation	10102 (100%)
Relationships between each patient and the proband of their family	5352	fact_relationship	5352 (100%)
Visits	32413	visit_occurrence	32413 (100%)
Diagnoses	10225	condition_occurrence	9853 (96.36%)
		observation	372 (3.64%)
Cardiac Events (cardiac arrest or syncope)	3376	condition_occurrence	3376 (100%)
Triggers of cardiac events	3182	condition_occurrence	1002 (31.49%)
		observation	2180 (68.51%)
Causality relationship between event and trigger	3182	fact_relationship	3182 (100%)
Therapy at the event time	6304	drug_exposure	627 (9.95%)
		observation	5677(90.05%)
Deaths	669	death	587 (87.74%)
Autopsies of sudden deaths	40	procedure	38 (95%)
Measurements (weight and height from different source tables)	11045	measurement	11045 (100%)

Table 1. Results of the data sources transformation into the OMOP CDM.

In yellow : the concepts mapped to different OMOP domains.
In red : the source domains with information loss.
In green : the source domains correctly and completely reported into the OMOP CDM format.

References

1. Tenenbaum JD, Avillach P, Benham-Hutchins M, et al. An informatics research agenda to support precision medicine: seven key areas. J Am Med Inform Assoc JAMIA. 2016;23(4):791–5.
2. Mantra – Biomeris Official Site [Internet]. [cited 2019 Mar 21]. Available from: https://biomeris.it/mantra/
3. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.
4. OHDSI Genomic CDM Subgroup [Internet]. [cited 2019 Mar 21]. Available from: http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:genetics-sg
5. Voss EA, Boyce RD, Ryan PB, et al. Accuracy of an automated knowledge base for identifying drug adverse reactions. J Biomed Inform. 2017 Feb 1;66:72–81.

Contact: elisamaria.zini01@universitadipavia.it