Lessons Learned from Implementing OMOP ETLs Across EHR and Registry-Based Data Sources

Renata Silva[1], Sofia Gomes[1] , Mariana Canelas-Pais[1,2,3], Ana Sá-Sousa[1], Tiago Taveira-Gomes[2,4]

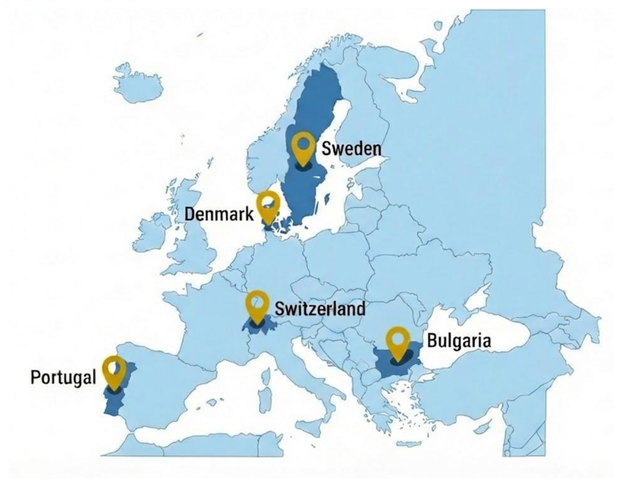
[1]MTG Research and Development Lab, Porto, Portugal
[2]Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Porto, Portugal
[3]RISE-Health - Centre for Health Technologies and Services Research, Porto, Portugal
[4]SIGIL Scientific Enterprises, Dubai, United Arab Emirates
renata.silva@mtg-research.com

Background

Data diversity between Electronic Health Records (EHRs) and clinical registries is a major barrier to multi-center research[1,2]. ETL design is strongly influenced by the structural and semantic characteristics of the source system[3].

This work compares ETL experiences across European EHRs and registries, to identify methodological challenges and support cross-border interoperability initiatives (e.g. European Health Data Space - EHDS).

Methods



Review of 17 OMOP-CDM ETL Implementations

6 Clinical Registries (ex: oncology, rheumatic disease)

vs

11 EHR Systems (e.g: primary care, inpatient care)

Results

Registry ETLs: focuses on the structural reshaping of denormalized datasets and the reconstruction of implicit clinical meaning from disease-specific variables to resolve fragmented data lineage

EHR ETLs: prioritizes navigating complex normalized schemas and interpreting workflow-driven semantics to extract relevant clinical events from high-volume transactional data

Table 1. Comparative characteristics observed during OMOP CDM ETL development across registry-based and EHR-based clinical data sources

Dimension	Registry	EHR
Data Structure	Wide, partially denormalized; JSON/XML attributes	Normalized relational schemas across linked tables
Vocabulary Availability	No standardized vocabularies; source-derived mappings	Controlled vocabularies with reference tables
SQL Transformation Patterns	Unpivot, UNION ALL, GROUP BY	Multi-table JOINs and filtering (WHERE)
Data Lineage Traceability	Fragmented lineage from manual curation	Strong lineage via transactional models
Relational Consistency	No referential integrity constraints;	Referential integrity enforced; Infrequent inconsistencies
Source Data Origin	Periodic data extracts or reporting datasets	Transactional clinical systems
Data Semantics Preservation	Reconstruction of registry-specific semantics	Workflow-driven data semantics
Access and Integration	Structured data exports	APIs, DB connections, vendor-mediated extractions
Temporal Data Representation	Simplified or aggregated time context	Reconstruction of clinical timelines
Historical System Evolution	Variable additions and schema drift	System migration and legacy workflow logic

CDM: Common Data Model; DB: Database; EHR: Electronic Health Record; OMOP: Observational Medical Outcomes Partnership.

Conclusion

To build scalable pipelines, OMOP-CDM ETL designs must be specifically tailored to the unique structural, semantic, and data-generation characteristics of each source system.

References

1. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining Data From Electronic Health Records [Internet]. Agency for Healthcare Research and Quality (US); 2019 [cited 2026 Feb 5]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551878/
2. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol [Internet]. 2021 Nov 2;21(1):238. Available from: https://doi.org/10.1186/s12874-021-01434-3
3. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open [Internet]. 2021 July;4(3):ooab001. Available from: http://dx.doi.org/10.1093/jamiaopen/ooab001