Standardizing Leigh Syndrome Patient Registry Data to the OMOP Common Data Model by Using Open Source R

Authors: Parag Shiralkar[1], Sanket Kalyankar[1], Jyoti Maini[1], Sushma Ghanta[1], Radhika Lakkireddy[1], Kasey Woleben[2], Sophia Zilber[2]

1- Sumptuous Data Sciences; 2 - Cure Mito Foundation

Background

OMOP (Observational Medical Outcomes Partnership)Common Data Model (CDM) was developed by the Observational Health Data Sciences and Informatics (OHDSI) community, which provides a standardized structure for organizing healthcare data and facilitating interoperability between different data sources. After standardization, data from disparate sources can be integrated and analyzed more efficiently, enabling comprehensive research and analysis.

Leigh syndrome is a rare, complex, and incurable early onset (typically infant or early childhood) mitochondrial disorder with both phenotypic and genetic heterogeneity, for which generating real-world evidence is challenging. Leigh syndrome global patient registry was developed by the Cure MIto Foundation, a leading patient advocacy organization dedicated to advancing research of Leigh syndrome and empowering and supporting affected families worldwide.

The registry includes two patient-reported surveys: a general survey utilizing the NIH Common Data Elements, and a Leigh Syndrome-specific survey and is hosted on a Coordination of Rare Diseases at Sanford (CoRDS) platform. Our objective is to document the process and outcomes of transforming registry data to the OMOP CDM and highlight challenges and our potential solutions.

Methods

The registry's raw data was scanned using White Rabbit, an application that evaluates the structure and contents of a database in preparation for developing an ETL (Extract, Transform, Load) process. ETL is used to effectively automate the transformations while preserving the quality and integrity of the data.

With the aid of the scan report produced by White Rabbit, we then created mapping specifications using Rabbit in a Hat, an application for interactively designing an ETL to the OMOP Common Data Model. Using OMOP CDM v5.4 IG, we applied mapping standards and used R programming to change the data through value standardization, de-duplication, and normalization to bring it into compliance with OMOP CDM criteria.

Additionally, with Athena's assistance, we gave concept ids and vocabulary to the data. Following data transformation, independent programming utilizing certain R packages was used for validation and quality assurance checks to guarantee data quality and conformance to OMOP CDM criteria.

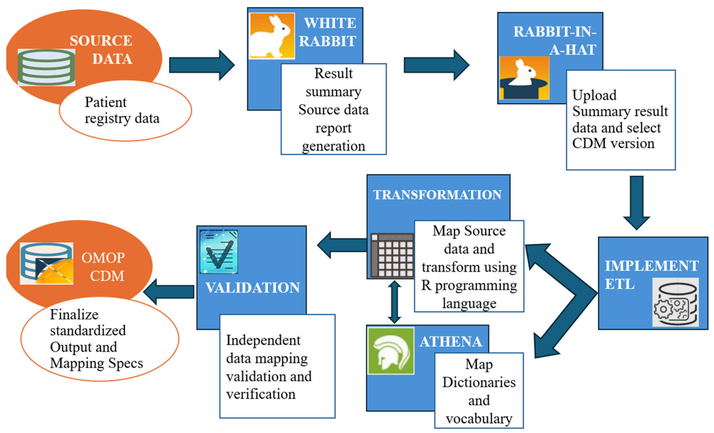
Contact:
parag.shiralkar@sumptuous-ds.com
sophia@curemito.org

Results

The entire registry dataset was evaluated, processed, and standardized to align with the requirements of the OMOP Common Data Model. The output we have developed was in compliance with OMOP CDM v5.4 Implementation guide.

Mapping source registry data to the OMOP CDM and aligning it with standardized vocabulary enhances data availability and usability for further disease research. This standardization eases seamless data integration, enabling robust real-world evidence generation. Our approach overcomes data heterogeneity challenges, paving the way for future research and improved patient outcomes. Additionally, converting source data into OMOP CDM generally enables mapping to other standards like SDTM, facilitating data availability for U.S. FDA submission purposes, although in our case data has already been mapped to SDTM previously.

Figure 1: Process flow of patient registry data transformation to OMOP



Conclusions

In this study, novel approaches for converting raw data from two registry databases to the OMOP CDM were successfully developed and applied with limited deviation and very few records being excluded from mapping. This approach addresses a need for combining real-world data of patients with rare diseases for the purpose of evidence generation and could serve as reference for future researchers wishing to undertake similar data mapping projects.

Acknowledgments: Vani Musikara, Alekhya Ettamsetti, Pallavi Bakare, Suraj Pujari, Danielle Boyce, MPH, DPA

Table 1: Mapping of Leigh syndrome patient registry data to OMOP domains.

Domain	Domain description	Leigh syndrome registry data mapped to domain
PERSON	Identifies each person with demographic details	Demographic information
VISIT_OCCURENCE	Captures details of healthcare encounters or instances of survey completion	Questionnaire date
CONDITION_OCCURRENCE	Records of conditions (diagnoses, symptoms) experienced by a person	Diagnosis, first concerns noticed, symptoms history
PROCEDURE_OCCURRENCE	Records of procedures	Type of genetic and diagnostic tests done
DEVICE_EXPOSURE	Records of medical device use	Devices such as feeding tubes, mobility devices
MEASUREMENT	Records of measurements or tests	Genetic testing results
OBSERVATION	Captures clinical facts about a person obtained in the context of examination, questioning or a procedure	Loss of milestones, caregiver burden, quality of life, family history, healthcare utilization
DEATH	Records of death and cause of death	Death information
SPECIMEN	Information about biological samples	Biospecimen information
LOCATION	Geographic information	Participant country or state
PROVIDER	Information about healthcare providers	Healthcare providers seen by participant
CDM_SOURCE	Contains detail about the source database and the process used to transform the data into the OMOP Common Data Mode	CDM Metadata information such as CDM source name, version of CDM, CDM holder name, CDM release date.

References

Zilber, S., Woleben, K., Johnson, S. C., et al. (2023). Leigh syndrome global patient registry: uniting patients and researchers worldwide. Orphanet Journal of Rare Diseases, 18, 264. https://doi.org/10.1186/s13023-023-02886-0

Shiralkar, P., Bakare, P., Woleben, K., & Zilber, S. (2024). Interoperability of Leigh syndrome patient registry data with regulatory submission standards. Journal of the Society for Clinical Data Management, 4(1). https://doi.org/10.47912/jscdm.244