Standardizing Registry CDISC SDTM data to the OMOP Common Data Model

--Parag Shiralkar[1], Suraj Pujari[1], Jyoti Maini1, Sanket Kalyankar[1], Sophia Zilber[2]

[1]Sumptuous Data Sciences, LLC, www.sumptuous-ds.com

[2] Cure Mito Foundation, www.curemito.org

Introduction:

In technologically enabled healthcare systems, the interoperability of standard data models such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) and the Clinical Data Interchange Standards Consortium (CDISC) is critical. This project is a systematic data transformation effort that ensured the semantic interoperability of concepts from CDISC SDTM to OMOP-CDM table mapping and variable-to-variable mapping. It emphasises the importance of uniform data models in dealing with data management difficulties in healthcare and enhancing data quality. This project used the patient foundation data from Cure Mito Foundation patient Registry. Cure Mito Foundation is a non-profit organization aimed to support the patients with Leigh Syndrome and to promote acceleration of therapeutic development to cure this disease. This patient registry data includes data from patients suffering from Leigh Syndrome, a type of Mito Chondrial disease. We had this patient registry data already converted to CDISC SDTM model. Our effort was to establish interoperability of this data with use of CDISC SDTM standards along with the OMOP-CDM.

Background:

Tackling the complexities and limitations of existing healthcare data management systems made it necessary for many to consider standard data models, including OMOP CDM and CDISC. This includes different data formats, non-standard coding systems, as well as poor intra-operability between databases. Standardization is a foundation to make these hurdles much less troublesome and on the other side improve data quality, interoperability and research outcomes, especially in healthcare. This project specifically focuses on converting data from the Leigh Syndrome Patient Registry of the Cure Mito Foundation, addressing the unique challenges associated with rare disease data management.

Methods:

The project started with the development of a complete CDM result schema and used tools such as Rabbit-in-a Hat and White Rabbit to map SDTM domains to CDM tables. Primarily, we Developed the CDM schema to determine which SDTM domain will be transformed into the CDM Table. This was then paired with extremely careful mapping variable-to-variable, where SDTM domain variables were being aligned with particular CDM fields to ensure 100% accuracy of the data presentation.

Additionally, a detailed CDM specification was crafted in Excel, and key vocabularies including SNOMED, LOINC, and RxNorm were downloaded for semantic enrichment.

Concept IDs were assigned with the help of the Athena tool and through complex programming in R Studio, we compiled concept IDs and joined these with the CDM data frame. Rigorous adherence to established rules ensured data consistency and accuracy throughout the process. The OMOP CDM version 5.4 was used as the framework for this conversion.

Overview:

Our project involved establishment of interoperability of CDISC SDTM with OMOP CDM for the patient registry data based on Leigh Syndrome, a rare disease impacting children. The project results clearly showed the powerful effects of a standardized data model on healthcare data integration. Conversion of Leigh Syndrome Patient Registry data from CDISC SDM to OMOP CDM was successful using the transformation process. This integration facilitated comprehensive data analyses, providing deeper insights into the disease and enhancing research capabilities.

The Key results of this CDISC SDTM to OMOP-CDM include:

1) Improved Data Consistency and Quality: Standardized vocabularies like SNOMED, LOINC, and RxNorm ensured consistent data representation and reduced ambiguities, leading to higher data quality.

2)Enhanced Interoperability: The use of OMOP CDM allowed for better interoperability between diverse data sources, enabling more robust data analysis and sharing across different healthcare systems.

These outcomes significantly contributed to improved insights, decision support, and research outcomes in healthcare, marking a substantial leap forward in data-driven healthcare practices. Specifically, the conversion of the Leigh Syndrome Patient Registry data enabled robust analysis and research capabilities, opening new avenues for understanding the interoperability of SDTM and OMOP CDM.

Discussion:

Although the project achieved remarkable results, it also encountered difficulties like resource-intensive mapping procedures and the intricacy of mapping specific data items. Mapping concept IDs programmatically presented a unique difficulty because, in Athena, concept IDs could only be acquired for exact matches, requiring manual assignment for other data.

It was also challenging to combine all the different domains. In order to solve this, datasets were combined vertically using the stacking method. Automating mapping chores, improving data validation techniques, and encouraging ongoing cooperation with domain experts to improve data standardization procedures are some areas that still require development. Furthermore, the CDISC

and OMOP CDM standards will continue to be improved and updated, which will increase their usefulness and efficacy in healthcare data management.

Conclusion:

The numerous obstacles related to healthcare data management can be overcome in large part by implementing standardized data models like OMOP CDM and CDISC. The Leigh Syndrome Patient Registry data was successfully integrated into the OMOP CDM framework as part of this project, demonstrating the standards' transformative power. The study showed notable gains in data quality and interoperability despite obstacles including manual concept ID assignments and integrating datasets with overlapping variables.

In addition to advancing research on rare diseases, this project's success establishes a standard for data integration initiatives in the future, which will ultimately lead to the development of more efficient and data-driven healthcare ecosystems. The techniques and insights discovered here will provide a strong basis for current and upcoming initiatives that aim to fully utilize standardized healthcare data.