Transforming the EBMT dataset to OMOP-CDM 5.3 has been challenging. We have encountered technical and vocabulary challenges.

Ongoing Transformation of the EBMT Registry to the OMOP CDM 5.3

Background: In April 2021, EBMT was granted funds from the EHDEN – Data Partner Call to map its registry data onto the OMOP CDM. Members of the EBMT are centres and individuals, active in the field of transplantation of any kind of haematopoietic cells, or any other organisation involved in the care of donors and recipients of haematopoietic cells.

Results & Discussion

- ETL design is completed (see Figure 1) and is being tested and iterated on.
- ETL design driven by data size (> 4 million records)
- Mappings through SQL queries (memory-efficient)

Encountered Issues

- Handling special characters in free text
- Custom concepts for very specific collected items
- Technical challenges due to DB size
- Genomic concepts
- Dates

Methods

The EBMT data is transformed to OMOP-CDM 5.3 through a custom-designed Extract-Transform-Load (ETL) process written primarily in the Python (v3.10) programming language. All transformations (source to PERSON, source to DEATH, etc.) are implemented in SQL and performed sequentially in an order predetermined based on table constraints and transformation dependencies.

Challenges: We have encountered technical challenges as well as vocabulary mapping challenges along the way. Dedicated functions and tools were developed to handle the technical hurdles. Where needed, custom vocabulary concepts are being created to ensure that the granularity of the source data can be maintained. We look forward to finalizing the work and collaborating with other EHDEN data-partners.