Standardizing European sarcoma registry data to the OMOP Common Data Model

Maaike van Swieten[1], Vittoria Ramella[2], Anna Alloni[2], Matteo Gabetta[2], Peter Prinsen[1], Chiara Attanasio[1], Espen Enerly[3], Siri Larønningen[4,] Roberto Lillini[5], Paolo Lasalvia[5], Joanna Szkandera[6], Stefan Janisch[6], Andreas Muth[7], Emelie Styring[8], Julien Bollard[9], Annalisa Trama[5], Gijs Geleijnse[1]

[1] Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands
[2] Biomeris, Italy
[3] Department of Research, Cancer Registry of Norway, Oslo, Norway
[4] Department of Registration, Cancer Registry of Norway, Oslo, Norway
[5] Fondazione IRCCS Istituto Nazionale dei Tumori INT, Milan, Italy
[6] Medical University of Graz, Graz, Austria
[7] Sahlgrenska University Hospital, Goteborg, Sweden
[8] Lund University, Lund, Sweden
[9] Centre Léon Bérard, Lyon, France

Background

In Europe, survival is worse for rare cancers than for common cancers[1]. Research in rare cancers is hampered by low patient numbers, dispersed clinical data and tumor samples and a limited number of experts per rare cancer diagnosis. A better understanding of the biology, optimal detection and diagnosis and treatment options may lead to improved patient care and outcomes. The ERN EURACAN[2] (European Reference Network for Rare Adult Solid Cancers) was established to address this issue. EURACAN is a network of European healthcare professionals and centers of excellence that share their knowledge and experience in order to develop and implement the best possible care for rare cancers. EURACAN addresses 10 out of the 12 families of rare cancers (75% of all rare cancers), including sarcomas - malignant tumors that develop in bone and/or soft tissue. The Blueberry project, funded by the Dutch Cancer Society (KWF), is part of EURACAN and aims to develop a blueprint for a harmonized sarcoma registry based on the Observation Medical Outcomes Partnership Common Data Model (OMOP-CDM). We include sarcoma data from heterogeneous data sources. Once the data is harmonized, we will execute a first simple use case. In the future, we will implement technology that allows for a federated execution of our use cases.

Methods

The Blueberry consortium consists of seven data partners, with both clinical registries from sarcoma expert centers and population-based registries. The data harmonization process includes four steps: data mapping, data transformation, data quality assessment and data analysis.

We first focused on mapping the core sarcoma data elements as identified by clinical experts. These data elements cover general patient demographics, risk factors and comorbidities, information about the diagnosis of the primary tumor and treatment, and follow up information where available. For this phase, we used the Observational Health Data Sciences and Informatics (OHDSI) tools[3] "White Rabbit", "Rabbit in a Hat" and "Usagi". Custom concept mappings were developed and shared within the Blueberry network for data elements that could not be mapped to standard OMOP concepts. Data mappings were verified with clinical experts to ensure that information from the source database was correctly transformed and the information loss was minimized. Once the data mappings were completed, we implemented extract-transform-load (ETL) pipelines tailored to each of the data sources, and loaded the harmonized data into the OMOP database. Then, we assessed the quality of the data transformation using OHDSI's "Data Quality Dashboard" (DQD) and "Achilles" tools[3]. Adjustments and improvements to the mapping and ETL implementation were performed based on the results of the data quality checks for completeness, conformance and plausibility of the harmonized data. This process was repeated until the output of DQD- and Achilles-based assessments matched our

expectations. When satisfied with the data quality of each of the data sources, we also assessed the quality of all the data within the Blueberry network and reviewed whether the data were 'fit-for-purpose' to execute a simple use case.

Results

The Blueberry project has made significant progress in harmonizing sarcoma data using the OMOP-CDM. Currently, four of the seven data partners have some or all identified core data elements in the OMOP data format, including patient demographics, diagnoses and first line treatments.

The Blueberry project has already highlighted several challenges in data harmonization such as differences in data definitions across different source systems. We are using a collaborative approach to validate the data mappings and reduce inconsistencies, involving clinical experts and data mapping experts. We are also running data quality checks on each of the data sources as well as at the network level using the new ARES package. We identified data quality issues, such as missing data due to missing concepts in the standard OMOP vocabularies and heterogeneity of data related to coding and formatting. To address the gaps in the OMOP vocabulary, we are creating custom concepts for the missing concepts, such as some of the ICDO3 diagnosis codes, to be used by all partners in the project.

Conclusion

The Blueberry project is a collaborative effort to develop a harmonized, federated sarcoma registry based on the OMOP-CDM, using OHDSI tools and methods for data harmonization and quality assessment. The project has made significant progress in mapping, transforming, and assessing the quality of data from four different data sources. Given that data harmonization is an iterative process, we will continue to review and improve this process as the remaining data sources are added. Furthermore, we anticipate to share the results of our first simple use case that will be executed during our Blueberry-study-a-thon in May, using the four OMOP data sources. These results will allow us to evaluate the use of oncology data stored in the OMOP CMD in an European collaboration.

References:

1 Gatta et al., "Burden and Centralised Treatment in Europe of Rare Tumours."
2 https://euracan.eu/
3 https://www.ohdsi.org/software-tools/