Adopting the OMOP Oncology CDM at the Helsinki University Hospital

Valtteri Nieminen, Alexey Ryzhenkov, Johanna Sanoja, Salma Rachidi, Juho Lähteenmaa, Joonas Laitinen, Samu Eränen, Johanna Niklander, Anna Kuosmanen, Oscar Brück, Pasi Rikala, Anna Virtanen, Marianna Niemi, Tomi Mäkelä, Eric Fey, Kimmo Porkka

Background

Helsinki University Hospital (HUS) has harmonized most of its available structured clinical data—including diagnoses, treatments, medications, procedures, and lab tests—into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). In the current implementation of the HUS OMOP database, events are often connected only by the individual patient. For oncology studies, a higher level of abstraction is required, with clinical events aggregated to represent, for example, disease episodes and treatment regimens.

The OHDSI Oncology working group's Oncology Extension defines conventions and a data model to represent cancer data as part of the OMOP CDM. Cancer is represented in *episodes* where a *primary diagnosis* with details such as biomarkers, genetic properties, spread and morphology are expressed as *condition modifiers*.

The key challenge for hospitals in adopting the Oncology extension's cancer model is integrating data from various source systems to represent cancer in an accurate and consistent way. This task is complicated by hospital information systems being designed for care and not research. Along with source systems' data models being heterogenous, oncology-specific data may not be coded clearly in a medical vocabulary or available in a structured form, necessitating pre-processing and interpretation. Treatment pathways of cancers also vary and specialized units have localized practices.

Here we report our progress and experiences from the ongoing effort to adopt the OMOP Oncology CDM and convert oncology data to OMOP at HUS. HUS is Finland's largest university hospital with a catchment area of circa 2.2. million people, and hosts Finland's only Organisation of European Cancer Institutes, (OECI) accredited Comprehensive Cancer Centre.

Methods

We assembled an interdisciplinary working group consisting of clinicians, data scientists, analysts, engineers, and OMOP Oncology CDM experts. First, a prioritized list of variables that would enable meaningful oncology studies was defined. To find the best available source, a data-analyst, together with a clinical expert, identified the data sources. These sources were evaluated, and source-to-concept mappings and a ETL plan developed iteratively together by the working group.

Results

Below we present our progress of adopting the Oncology CDM, developing required data-processing pipelines, initial mappings, available oncology data, and ETL.

Somatic mutation data

We implemented a process to map and load genomics data to OMOP for two data available sources: 1) semi-structured geneticist's statements of somatic mutations and 2) data already in a standard genomic format, such as variant-call-factor (VCF) files. In both cases, we leveraged *ClinGen* - a knowledge base of clinical relevance of genes to arrive at a synonym that can be mapped to OMOP Genomic vocabulary.

The semi-structured statement text varies depending on the mutation panel in question. We developed

a parser to extract, the NM-transcript (e.g NM_003016.4) representing a target gene, the alteration (e.g c.284C>A) and the *variant allele frequency* (VAF). The NM_transcript and gene alteration were concatenated to the Human Genome Variation Society (HGVS) format. A URL was then created to programmatically query the ClinGen database (e.g., reg.genome.network/allele?hgvs=NM_003016.4:c.284C>A). The query returns a JSON format file, from which a synonym for the mutation in the OMOP Genomics vocabulary (Feb. 2024 release) was fetched to construct a source-to-concept mapping. This approach was inspired by the KOIOS-software, but we streamlined the approach to only execute the query and fetch the synonymous concept. For VCF files we used KOIOS to produce mappings.

Treatment regimens

At HUS, drug prescriptions and administrations are recorded at ingredient level and data that would link treatment cycles and regimen are not available. Therefore, regimen have to be derived from the combinations and timings of individual drug administration records. To achieve this, we piloted the distance-based OHDSI ARTEMIS regimen detection tool in lung cancer with promising results. Future work will concern the definition of regimen for other cancer types. The next step is to define the local regimens together with clinicians.

Primary diagnosis, episodes and initial diagnosis

In determining the initial diagnosis date and primary diagnosis required for defining the OMOP Oncology CDM *episodes* we referred to the European Network of Cancer Registries (ENCR) recommendations for incidence date. The ENCR recommendations determine a six step, declining priority order where we target the first one; the date of first histological or cytological confirmation of malignancy. This then dictates that episodes start from the date of the biopsy and the primary diagnosis will be the resulting diagnosis of a pathologist's examination of the sample. Our granularity target for the diagnosis is the same as the Oncology Working group's optimal target; the ICD-O-3 level.

Oncology-specific pathology data

Oncology-specific pathology data in HUS is stored in cancer-specific forms created by clinicians. These forms contain cancer-specific information such as, anatomical location, morphology/histology, and stage. Unfortunately, the forms are not expressed in a standardized medical vocabulary but created by clinician expert groups. The form used depends on indication and the sampling method (resection, needle-biopsy etc.). To utilize these data, cancer specific working groups created intermediate mappings (a custom source vocabulary for each form) field-by-field and created a preprocessing pipeline before ETL. So far, granular mappings and ETL for lung, breast, and prostate cancer have been developed.

Conclusion

Adopting the OMOP Oncology CDM for hospitals is feasible and holds great promise for performing oncology studies that are not possible using data from other oncology data sources such as cancer registries. However, mapping hospital data can be challenging, because abstract, high-level information readily available in registries designed to encompass only oncology data (such as disease and treatment episodes) are usually not readily available in hospital clinical data and have to be extracted from free-text notes, semi-structured data, and/or derived from the combination of lower level, granular records. From our experiences, we highlight the approach inspired by the KOIOS software to mapping genomic data as well as using the ENCR recommendations as the backbone for determining the initial diagnosis and episodes.