MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190670

Standardized Observational Cancer Research Using the OMOP CDM Oncology Module

Rimma Belenkaya[a], Michael Gurley[b], Dmitry Dymshyts[c], Sonia Araujo[d], Andrew Williams[e], RuiJun Chen[f], Christian Reich[g]

[a] OHDSI Oncology Workgroup, Memorial Sloan Kettering Cancer Center, New York City, NY, USA,
[b] Clinical and Translational Sciences Institute, Northwestern University, Chicago, IL, USA,
[c] Odysseus Data Science Inc, Cambridge, MA, USA,
[d] Real World Analytics Solution, IQVIA, London, UK,
[e] Maine Medical Center Research Institute, Center for Outcomes Research and Evaluation, Portland, ME, USA,
[f] Biomedical Informatics Department, Columbia University Medical Center, New York City, NY, USA,
[g] Real World Analytics Solution, IQVIA, Cambridge, MA, USA

Abstract

Observational research in cancer requires substantially more detail than most other therapeutic areas. Cancer conditions are defined through histology, affected anatomical structures, staging and grading, and biomarkers, and are treated with complex therapies. Here, we show a new cancer module as part of the OMOP CDM, allowing manual and automated abstraction and standardized analytics. We tested the model in EHR and registry data against a number of typical use cases.

Keywords:
Oncology, Research, Standardized

Introduction

Observational research utilizes secondary use of observational healthcare databases for pharmacoepidemiology, health outcomes, and health services research. A variety of statistical and epidemiological methods produce estimates of desired or adverse effects caused by medical interventions and associated statistical artifacts such as confidence intervals and p-values. Cancer research is no different from general observational research. However, cancer characterization requires much more details such as: histology of the tumor tissue, anatomical sites affected by the disease, tumor size, affected lymph nodes and metastases, tumor markers (conventional and genomic aberrations), and standardized disease grade and staging.

The treatment of cancer is also more complex than most other conditions since it involves surgery, radiotherapy, chemotherapy, targeted therapy, and immunotherapy. Many of these are administered in cycles. In addition, chemotherapy compounds are applied in predefined regimens.

Cancer diagnosis and treatment interventions are complicated healthcare interactions stretching over longer periods of time. To perform observational cancer research, low-level events need to get abstracted into higher level disease episodes, treatments, and outcomes such as initial diagnosis, drug regimen, radiotherapy, surgical resection, response to treatment, overall and disease-free survival, etc.

Systematic and standardized analytics requires data harmonization, which also enables distributed research networks [1]. Both data format and representation (coding) need to be standardized for true federated analytics. Currently, there is no comprehensive data model or standard semantic terminology system covering the cancer domain available in the public domain to support this approach.

ICD-10 is the most common coding scheme for defining the condition in observational data. However, it details mostly anatomical sites of cancer, while histology and tumor attributes are poorly covered. ICD-O-3 is a classification with explicit topology and histology domains, but it is not connected to other terminologies. The two coding systems are not combined to form meaningful conditions and it lacks detailed tumor characteristics. The CAP Cancer Protocols solves the latter problem by defining a mandatory and specific set of data elements for each cancer type, called synoptic reporting. However, this resource does not have a freely available computer-readable representation and does not have a properly defined terminology. The NAACCR Data Dictionary is addressing this issue and also contains terminologies for treatment. However, its data elements are not conceptualized to carry a distinct meaning independent from context, and for most of its data elements, it does not utilize or map to any external existing coding system. For clinical trials, CDISC developed comprehensive Therapeutic Area Standards, but it covers only four cancer types and also lacks relationships to external standards. HL7 FHIR is developing cancer-specific profiles, but only Breast Cancer Staging is available today. The Nebraska Lexicon addresses all of the above issues, creating a freely available ontology embedded in SNOMED-CT, but it is a work in progress, and currently only covers breast, colorectal, lung cancer, and malignant melanoma.

For cancer treatments, the situation is equally inconsistent. RxNorm is the standard vocabulary for the OMOP drug domain, but it is not specific to oncology and does not contain regimens. Both the NCI List of cancer drugs, the SEER*Rx, and the NCI Metathesaurus contain cancer specific drugs and their indication, but not regimens with constituent ingredients normalized to RxNorm. The NCCN has comprehensive drug and regimen information, but nothing is available in the public domain. The SEER OROT contains a comprehensive cancer drug repository with maps to NDC and HCPCS. Finally, HemOnc contains a rich drug and regimen ontology, including indications and links to RxNorm but requires proper life cycle of their concepts.

Here, we introduce the OMOP Cancer Module. It consists of a comprehensive hierarchy of cancer conditions, including topography, histology and tumor attributes, as well as treatments with regimens. It also contains the data model for abstracted

episodes and a mechanism to link them to the lower level clinical detail. We converted four cancer databases into this model and tested it for a number of typical use cases.

Methods

We used the following standardized vocabularies:

1. Cancer diagnosis: We incorporated all ICD-O-3 histology and topography codes and created equivalent or "uphill" links to SNOMED. We combined all histologies and topologies to create standard conditions. Instead of instantiating the entire Cartesian product, we only used reported combinations derived from ICD-O-3 site, ICD-O-3 SEER Site/Histology Validation List, Columbia University Medical Center (CUMC) Cancer Registry, and Northwestern University (NU) Tumor Registry. We mapped the resulting conditions that existed in SNOMED, and connected the rest of the conditions to their histology and topography and linked them to the higher level SNOMED concept.
2. Modifiers: We incorporated and instantiated both the NAACCR Data Dictionary and the Nebraska Lexicon and de-duplicated the resulting corpus. For ambiguous NAACCR concepts, we pre-coordinated the data elements with its site-specific context.
3. Treatment episodes: We instantiated NAACCR surgery and radiation concepts. We incorporated HemOnc drug regimens and classification. We used OROT to connect drug concepts to the underlying source codes.

Episode model

We introduced new Episode and Episode_Event tables to represent disease and treatment episodes and their connection to lower level events (Figure 1).

Database instantiation

We converted Electronic Health Records (EHRs) and registries from four participating institutions into the new model.

Abstraction

We used an automatic abstraction algorithm to derive disease and treatment episodes [2]. We also performed manual abstraction and compared it to the automatic abstraction.

Empirical Model Testing

We compared the manual to the automatic abstraction and characterized the results. We then tested the following use cases using either abstractions.

- Complete history of cancer disease for patients cohort
- Complete list of modifiers for cohort of patients
- Stratification of therapy by high level classes based on vocabulary hierarchy

For each use case, we distributed a standard query through the network of databases and experts validated the results.

Results

We successfully generated the Standardized Vocabularies and converted the data from four participating institutions: NU, CUMC, Maine Medical Center Research Institute, and IQVIA.

We achieved 95% of coverage for the diagnoses reported in the source data by the Standardized Vocabularies, the remaining 5% represented rare cancers. The positive predictive value (PPV) of the automatic abstraction compared to manual chart review and Cancer Registry was 96% for any cancer diagnosis for four cancers, 100% for chemotherapy, 98% for hormone therapy, 100% for immunotherapy, and 86% for radiotherapy.

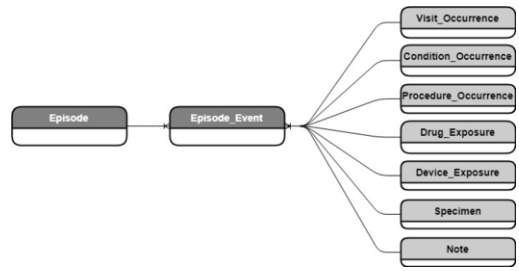

Figure 1– New Episode and Episode_Event tables

Conclusions

We successfully standardized the key aspects of cancer diseases required for observational research. Further work will include: extending Standardized Vocabularies to provide complete coverage of diagnoses and diagnostic modifiers, modeling of genomic data and outcomes, and creating automatic abstraction algorithms that would be utilized in an Open Network of cancer databases.

References

[1] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, and D. Madigan, Characterizing treatment pathways at scale using the OHDSI network, *Proceedings of the National Academy of Sciences of the United States of America* **113** (2016), 7329-7336.

[2] N.M. Carroll, K.M. Burniece, J. Holzman, D.B. McQuillan, A. Plata, and D.P. Ritzwoller, Algorithm to Identify Systemic Cancer Therapy Treatment Using Structured Electronic Data, *JCO Clin. Cancer Informatics*. (2017).

Address for correspondence

Rimma Belenkaya
belenkar@mskcc.org.