OMOP CDM Oncology Module at Work

Rimma Belenkaya, Michael Gurley, Christian Reich, Dmitry Dymshyts, Jeremy Warner, Robert Miller, Andrew Williams, RuiJun Chen

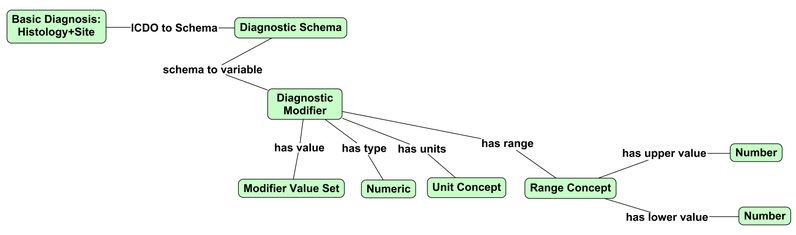
Background

Observational research in cancer requires substantially more detail to represent diagnoses, treatments, and outcomes than most other therapeutic areas. Cancer diagnosis is defined by a constellation of histology, site, stage, grade, genetic biomarkers. Cancer treatments are administered in defined order and cycles, and cannot be fully described by individual medications. At the same time, clinically and analytically relevant representation of cancer diagnoses, treatments, and outcomes requires data abstraction (e.g. recurrence, remission, end-of-life events, chemotherapy regimens, treatment cycles, response to treatments) that is not readily available in the source data and has not been traditionally supported in OMOP CDM. Here, we introduce a new Cancer Module of the OMOP CDM, which allows for both the required granularity and abstraction of cancer data to support transformation from the source data and standardized analytics. We tested the Module in EHR and Cancer Registry data against a number of typical use cases.

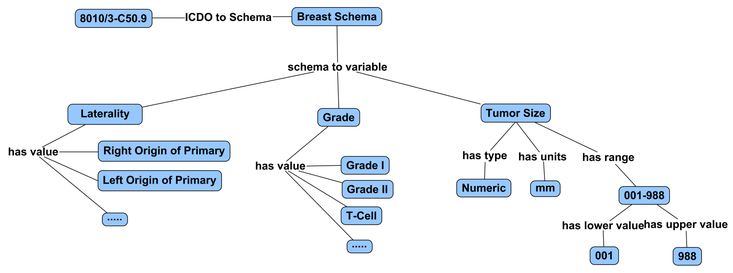
Methods

Cancer diagnosis as a complex model

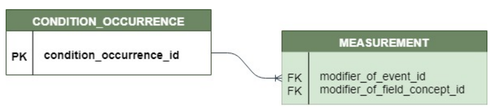
Cancer Diagnosis Model in the OMOP Vocabulary



Example of Cancer Diagnosis Model in the OMOP Vocabulary



Cancer diagnosis representation in the OMOP CDM



- **Precoordinated concept** of cancer **Histology+Site** is stored in **Condition_Occurrence**
- **Diagnostic modifiers** are stored in **Measurement** and **linked to** the **Condition_Occurrence** record

Example of cancer diagnosis in the OMOP CDM

Histology+Site diagnosis in **Condition_Occurrence**

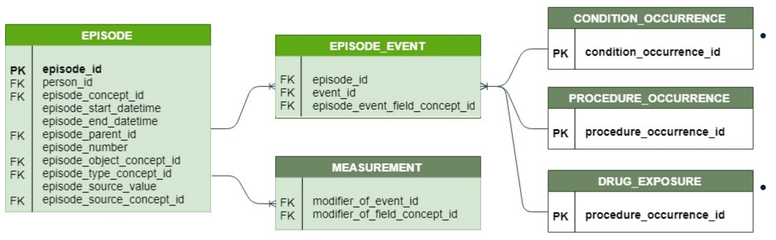
condition_occurrence_id	123456789	
person_id	1	
condition_concept_id	**4116071**	← **SNOMED concept 'Carcinoma of breast'**
condition_start_datetime	June 9, 2019	
condition_type_concept_id	32535	
condition_source_value	**8010/3-C50.9**	← **Precoordinated concept of ICD-O Histology & Site**
condition_source_concept_id	44505310	

Grade modifier in **Measurement**

measurement_id	567890
person_id	1
measurement_datetime	June 9, 2019
measurement_concept_id	**35918640** ← **NAACCR concept 'Grade Pathological'**
measurement_date	June 9, 2019
value_as_concept_id	**35922509** ← **NAACCR concept 'G3: High combined histologic grade (unfavorable); SBR score of 8-9 points'**
measurement_type_concept_id	32534 ← **OMOP concept 'Tumor registry'**
measurement_source_value	3844 ← **NAACCR code for 'Grade Pathological'**
measurement_source_concept_id	35918640 ← **NAACCR concept 'Grade Pathological'**
value_source_value	breast@3844@3 ← **NAACCR code for 'G3: High combined histologic grade (unfavorable); SBR score of 8-9 points'**
modifier_of_event_id	**123456789** ← **Value of the respective condition record condition_occurrence_id**
modifier_field_concept_id	**1147127** ← **Concept for 'condition_occurrence.condition_occurrence_id'**

Disease and treatment abstraction

Disease and treatment episodes in the OMOP CDM



- New **Episode** model added to the OMOP CDM
- **Episodes** are represented by **episode_concept_id** and **episode_object_concept_id:** '*Treatment Regimen*'-of-'*Paclitaxel and Bevacizumab*'
- **Episodes** and **lower level events** (e.g. conditions, drugs) are **linked** via **Episode_Event** table

Example of disease and treatment episodes in the Episode table

'First occurrence'*-of-*'Carcinoma of breast'

episode_id	12345	
person_id	1	
episode_concept_id	**32528**	← **OMOP concept 'First disease occurrence'**
episode_start_datetime	June 9, 2019	
episode_object_concept_id	**4116071**	← **SNOMED concept 'Carcinoma of breast'**
episode_type_concept_id	32535	← **OMOP concept 'Tumor registry'**

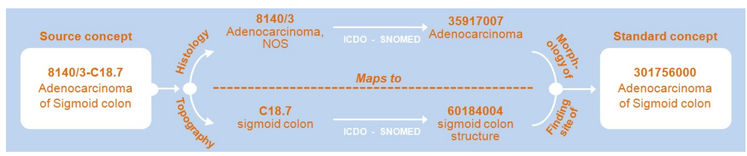
'Treatment regimen'*-of-*' Paclitaxel and Bevacizumab'

episode_id	12346	
person_id	1	
episode_concept_id	**32531**	← **OMOP concept 'Treatment Regimen'**
episode_start_datetime	July 9, 2019	
episode_parent_id	12345	← **Foreign key to the disease Episode record**
episode_object_concept_id	35804255	← **HemOnc concept Paclitaxel and Bevacizumab'**
episode_type_concept_id	32545	← **OMOP concept 'Episode algorithmically derived from EHR'**

New vocabularies

ICD-O **Gold standard for encoding pathology diagnosis, Histology + Site**

- ICD-O **Histology + Site** were **precoordinated** and **mapped to SNOMED**
- **Unmapped concepts** were designated as **standard**

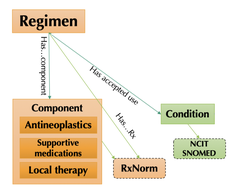


NAACCR **Data dictionary** for recording **US Tumor Registry** data

- Used as a template for **linking diagnostic modifiers to** their respective **cancer type**
- Served as a **foundation for generic ETL** for ingesting Tumor Registry data into OMOP CDM

HemOnc **Ontology of cancer medications**

- **Connects individual drug ingredients** to **chemotherapy regimens**
- **Classifies cancer treatments**
- Provides **diagnostic indication for each treatment**
- Used to **automate chemotherapy regimen extraction** from individual drugs



Database instantiation and ETL

- Developed **vocabulary-driven ETL** for data conversion from Tumor Registry
- **Converted EHR** and **Registry data** into OMOP CDM
- **Derived First Cancer Occurrence Episode** from lower level events
- **Derived First Treatment Course Episode** including **Chemotherapy Regimens** from lower level events

Results

We converted EHR and Registry data from four participating institutions using uniform vocabulary-driven ETL.

We derived First Cancer Occurrence and First Treatment Course using NAACCR and HemOnc vocabularies.

We achieved 95% of coverage for the diagnoses reported in the source data by the Standardized Vocabularies, the remaining 5% representing rare cancers.

We tested the following clinical characterization use cases:

1. Survival from initial diagnosis
2. Time from diagnosis to treatment
3. High-level treatment courses for 1st cancer occurrence
4. Derivation of chemotherapy regimens from atomic drugs

Conclusions

We incorporated foundational structural and semantic support into the OMOP CDM to represent clinical cancer disease and treatment data. This significantly improves specificity of cancer cohort definitions. Introduction of disease and treatment abstractions supports key clinical characterization use cases.

Future work on the Oncology Module will include:

- Adding domains for genomics, imaging and outcomes
- Improving ICD-O-3 to SNOMED mapping precision
- Mapping of NAACCR data dictionary concepts to SNOMED, using Nebraska Lexicon
- Improving precision of chemo regimen identification.

Memorial Sloan Kettering Cancer Center 1884 · Northwestern University · IMS Health & Quintiles are now IQVIA · ODYSSEUS DATA SERVICES INC · VANDERBILT · Tufts CTSI Tufts Clinical and Translational Science Institute · COLUMBIA Columbia University Department of Biomedical Informatics