Memorial Sloan Kettering Cancer Center

Extending OMOP CDM to Support Observational Cancer Research

Michael Gurley
Rimma Belenkaya

Content

This document contains two parts:

Challenges

- **Reconciliation of cancer data from heterogeneous sources**
  - Quality: Completeness and Accuracy
    - Cancer Registries: complete for 1st occurrence (except for SEER states); high quality (golden standard)
    - Electronic Medical Records: complete; variable quality
    - Clinical Trials: complete; high quality
  - Encoding: Variations and Granularity
    - Cancer Registries: ICD-O; internal NAACCR vocabulary
    - Electronic Medical Records, ICD-9/10; free text
    - Clinical Trials: CDISC; custom coding

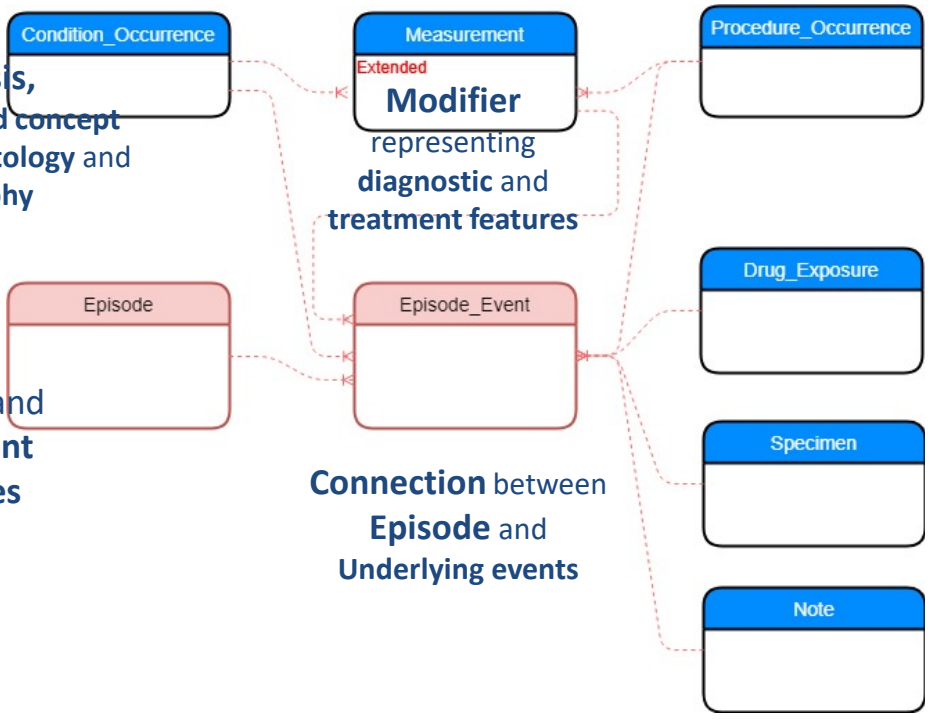
- **Gaps in semantic standards**
  - NAACCR is not mapped to any terminology
  - CAPs, synoptic pathology reports do not have complete terminology coverage
  - Existing drug classifications are not specific to oncology
  - Drug regimen semantic representation is not complete

- **Absence of abstraction layer representing clinician's/researcher's view**
  - Disease and treatment episodes, outcomes
  - Connection between higher level abstractions and lower level events
  - Prediction of: response to treatment, overall and disease free survival, time to relapse, end of life event

Overall Approach

- **Cancer diagnosis**
  - Represent cancer diagnosis as a combination of **histology** (morphology) + **topography** (anatomy)
- **Modifiers**
  - Diagnostic and treatment features that vary between different cancer diagnoses and treatments are represented as modifiers and explicitly linked to the respective diagnosis or treatment
  - Examples of diagnosis modifiers are stage, grade, laterality, foci, tumor biomarkers. These diagnostic features are assessed when a patient is first diagnosed and also (possibly) for each cancer recurrence. Repeated measurements of the same modifier (lymph node invasion) may be recorded. Different modifiers may be recorded on different dates
  - Examples of treatment modifiers are surgery laterality, radiotherapy dosage and frequency.
- **Disease and treatment episodes**
  - Disease and treatment abstractions are modeled as **episodes**, a new CDM construct that can be used to represent other abstractions such as episode of care.
  - These episodes may be derived algorithmically pre- or post-ETL or extracted from the source data directly. In addition to the regular OMOP type_concept_ID, we propose to store references to the derivation algorithms in the vocabulary.
  - Disease episodes include first occurrence, remissions, relapses, and end of life event.
  - Treatment episodes include treatment course, treatment regimen, and treatment cycle.
  - One set of “verified” modifiers is associated with each disease/treatment episode.

Cancer Representation in OMOP CDM



Diagnosis,
pre-coordinated concept
representing **histology** and
topography

Modifier
representing
diagnostic and
treatment features

Underlying events

Disease and
Treatment
Episodes

Connection between
Episode and
Underlying events

Cancer Representation in OMOP CDM

- **Cancer diagnoses** are stored in CONDITION_OCCURRENCE as pre-coordinated concepts combining histology and topography.
- **Cancer treatment events** are stored in the PROCEDURE_OCCURRENCE and DRUG_EXPOSURE tables.
- **Disease** and **treatment episodes** (e.g. first cancer occurrence, treatment regimen hormonal therapy) are represented in the new EPISODE table.
- **Links between the disease** and **treatment episodes and the underlying events** (conditions, procedures, drugs) are stored in the new EPISODE_EVENT table.
- **Additional diagnostic** and **treatment features** are stored in the MEASUREMENT table as modifiers of the respective condition, treatment, or episode. MEASUREMENT table is extended to include a reference to the condition, treatment, or episode record.

Cancer Diagnosis Record

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**4200514 ("Adenocarcinoma of sigmoid colon")**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**
condition_source_concept_id	**36517865 ("Adenocarcinoma of sigmoid colon")**
condition_source_value	**Histology 8140/3; Topography C18.7**

SNOMED

ICD-O, collapsed

Cancer Diagnosis in OMOP Vocabulary

1. New pre-coordinated concept representing combination of two ICD-O axes, histology and topography

CONCEPT

Field	Content
concept_id	36517865
concept_name	Adenocarcinoma of sigmoid colon
concept_code	8140/3- C18.7
vocabulary_id	ICDO3
standard_concept	

CONCEPT_RELATIONSHIP

Field	Content
concept_id_1	36517865
relationship_id	Has associated histology
concept_id_2	4290838 ("Adenocarcinoma, NOS")

CONCEPT_RELATIONSHIP

Field	Content
concept_id_1	36517865
relationship_id	Has finding site
concept_id_2	4244588 ("Sigmoid colon")

2. Existing pre-coordinated SNOMED concept linked to the same histology and topography axes

CONCEPT

Field	Content
concept_id	4200514
concept_name	Adenocarcinoma of sigmoid colon
concept_code	301756000
vocabulary_id	SNOMED
standard_concept	S

CONCEPT_RELATIONSHIP

Field	Content
concept_id_1	4200514
relationship_id	Has associated histology
concept_id_2	4290838 ("Adenocarcinoma, NOS")

CONCEPT_RELATIONSHIP

Field	Content
concept_id_1	4200514
relationship_id	Has finding site
concept_id_2	4244588 ("Sigmoid colon")

3. Mapping between the new pre-coordinated source concept and a standard SNOMED concept

CONCEPT_RELATIONSHIP

Field	Content
concept_id_1	36517865
relationship_id	Maps to
concept_id_2	4200514

Advantages of Using Histology-Topography Pre-coordinated Concepts

- **Reflect source granularity** in Cancer Registries and Pathology reports
- **Consistent with OMOP CDM**
  - representation of diagnosis as one concept
  - usage and extension of SNOMED
- **Support consistent queries** along histology and topography axes at different levels of hierarchy regardless of source representation

Episodes

- **Disease** and **treatment abstractions** are modeled as episodes
- **Disease abstractions include**: first occurrence, remissions, relapses, and end of life event.
- **Treatment abstractions include**: treatment course, treatment regimen, and treatment cycle.
- These **abstractions may be derived** algorithmically pre- or post-ETL or extracted from the source data directly.
  - In addition to the regular OMOP type_concept_ID, we propose to store **references to the derivation algorithms** in the vocabulary.

EPISODE and EPISODE_EVENT Tables

Field	Required	Type	Description
episode_id	yes	integer	A unique identifier for each Episode.
person_id	yes	integer	A foreign key identifier to the Person who is undergoing the Episode. The demographic details of that Person are stored in the PERSON table.
episode_concept_id	yes	integer	A foreign key that refers to a standard Episode Concept identifier in the Standardized Vocabularies. Examples of an Episode Concept can be: Treatment Regimen, Treatment Cycle, Disease First Occurrence, Remission, Relapse, Episode of Care
episode_start_datetime	yes	date	The date and time on which the Episode was started.
episode_end_datetime	yes	date	The date and time on which the Episode was ended.
episode_parent_id	no	integer	A foreign key that refers to a parent Episode entry representing an entire episode if the episode spans multiple cycles.
episode_number	no	integer	An ordinal count for an Episode that spans multiple times
episode_object_concept_id	yes	integer	A foreign key that refers to a concept identifier in the Standardized Vocabularies describing disease, treatment, or other abstraction that the episode describes. For example, 'Breast Carcinoma' or 'Chemotherapy'.
episode_type_concept_id	yes	integer	A foreign key that refers to a standard Episode Type Concept identifier in the Standardized Vocabularies reflecting the provenance of the episode derivation. It may reference a derivation algorithm, sources such as cancer registry, EMR, etc.
episode_source_value	no	varchar(50)	The source code for the Episode as it appears in the source data. This code is mapped to a standard episode Concept in the Standardized Vocabularies and the original code is, stored here for reference.
episode_source_concept_id	no	integer	A foreign key to a Episode Concept that refers to the code used in the source.

Field	Required	Type	Description
episode_id	yes	integer	A foreign key identifier to the Episode that the Episode Event belongs to.
visit_occurrence_id	no	integer	A foreign key identifier to the visit_occurrence record for which an episode is recorded.
condition_occurrence_id	no	integer	A foreign key identifier to the condition_occurrence record for which an episode is recorded.
procedure_occurrence_id	no	integer	A foreign key identifier to the procedure_occurrence record for which an episode is recorded.
drug_exposure_id	no	integer	A foreign key identifier to the drug_exposure record for which an episode is recorded.
device_exposrue_id	no	integer	A foreign key identifier to the device_exposur record for which an episode is recorded.
measurement_id	no	integer	A foreign key identifier to the measurement record for which an episode is recorded.
specimen_id	no	integer	A foreign key identifier to the speciment record for which an episode is recorded.
observation_id	no	integer	A foreign key identifier to the observation record for which an episode is recorded.
note_id	no	integer	A foreign key identifier to the note record for which an episode is recorded.
cost_id	no	integer	A foreign key identifier to the cost record for which an episode is recorded.

Disease Episode Records

EPISODE

Field	Content
episode_id	**4325345**
person_id	**John Smith**
episode_concept_id	**First Occurrence**
episode_start_datetime	**February 14, 1996**
episode_end_datetime	**November 18, 1996**
episode_object_concept_id	**Adenocarcinoma of sigmoid colon**
episode_type_concept_id	**Algorithm #123**

EPISODE_EVENT

Field	Content		
episode_id	**4325345**	**4325345**	**4325345**
condition_occurrence_id	**9900145**	**9900850**	
procedure_occurrence_id			**456774870**
drug_exposure_id			
specimen_id			
note_id			

CONDITION_OCCURRENCE

Field	Content
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**Adenocarcinoma of sigmoid colon**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**

CONDITION_OCCURRENCE

Field	Content
condition_occurrence_id	**9900850**
person_id	**John Smith**
condition_concept_id	**Adenocarcinoma of sigmoid colon**
condition_occurrence_start_datetime	**September 15, 1999**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**EMR**

PROCEDURE_OCCURRENCE

Field	Content
procedure_occurrence_id	**456774870**
person_id	**John Smith**
procedure_concept_id	**Intravenous chemotherapy**
procedure_occurrence_start_datetime	**November 1, 1996**
procedure_occurrence_end_datetime	**November 18, 1996**
procedure_occurrence_type_concept_id	**EMR**

Treatment Episode Records

EPISODE

Field	Content
episode_id	9900850
person_id	John Smith
episode_concept_id	Treatment Regimen
episode_start_datetime	August 1, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	
episode_number	
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	Cancer Registry
episode_source_value	Chemotherapy
episode_source_concept_id	C25 (NAACCR ID)

EPISODE

Field	Content
episode_id	9900851
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	August 1, 1996
episode_end_datetime	August 28, 1996
episode_parent_id	9900850
episode_number	1
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE

Field	Content
episode_id	9900852
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	October 15, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	9900850
episode_number	2
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE_EVENT

Field	Content		
episode_id	9900851	9900851	9900851
condition_occurrence_id			
procedure_occurrence_id			
drug_exposure_id	9900145	9900146	9900147
device_exposure_id			
observation_id			
specimen_id			
note_id			

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900145
person_id	John Smith
drug_concept_id	Cyclophosphamide
drug_exposure_start_datetime	August 1, 1996
drug_exposure_end_datetime	August 1, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Cyclophosphamide 1000 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900146
person_id	John Smith
drug_concept_id	Doxorubicin hydrochloride
drug_exposure_start_datetime	August 4, 1996
drug_exposure_end_datetime	August 4, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Doxorubicin Hydrochloride 50 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900147
person_id	John Smith
drug_concept_id	Dexamethasone acetate
drug_exposure_start_datetime	August 7, 1996
drug_exposure_end_datetime	August 7, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Dexamethasone acetate 8 MG/ML Injectable

Vocabulary Extensions for Episodes

- Add ‘Episode’ domain and concepts for episode_concept_ID
  - Examples of concepts: ‘First Disease Occurrence’, ‘Treatment Regimen’.
- Add episode type concepts in the 'Type Concept' domain:
  - Examples of concepts: ‘Algorithmically-derived episode pre-ETL ‘.
- Add new ‘Procedure/Treatment’ domain and concepts for episode_object_concept_id
  - Base on NAACCR/SEER treatment variables
  - Examples of concepts: ‘Chemotherapy‘, ‘External beam, photons’
- Add cancer specific treatment classification (Drugs, Surgical, Radiothearpy)
  - Source: Observational Research in Oncology Toolbox (OROT) classification vocabulary
- Add treatment regimen specifications
  - Source: HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals

Advantages of Using Episodes

- Supports levels of abstraction that are **clinically** and **analytically relevant**
- Supports **explicit connection between a disease/treatment abstraction** and **lower level events** (conditions, procedures, drugs) that are linked to this abstraction
- **Persists provenance of episode derivation** (e.g. directly from source data, algorithmically)
- Is **generalisable** to:
  - abstraction of **other chronic diseases**
  - Representation of **episode of care** (Gowtham, to be continued)

Modifiers

- Modifier are **similar to measurements** in that they require a standardized test or some other activity to generate a quantitative or qualitative result.
- Modifiers **are not independent measurements**: they add specificity to cancer diagnosis, treatment, or episode.
  - For example, LOINC 44648-4 'Histologic grade' may modify cancer diagnosis of “Tubular carcinoma” recorded in CONDITION_OCCURRENCE.
  - Therefore, although modifier_of_event_id and modifier_of_table_concept_id are not required fields, they must be populated for modifiers.
- **Repeated modifier records** (lymph node invasion) may be associated with one or multiple condition occurrence records.
- Modifiers for the same condition record **may be recorded on different dates**.
- **One set of “verified” modifiers** must be associated with a disease or treatment episode.

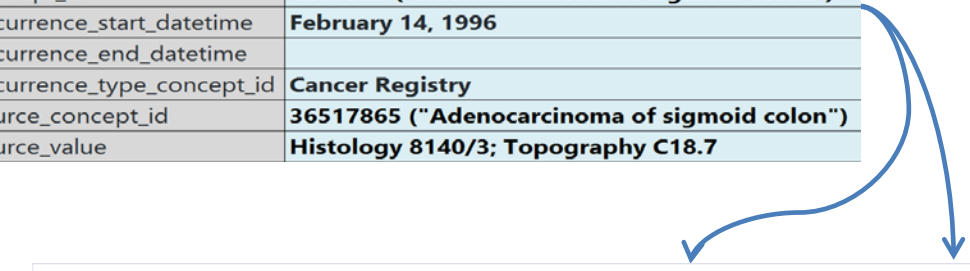
Modifiers: Extension of MEASUREMENT Table

Field	Required	Type	Description
modifier_of_event_id	No	integer	A foreign key identifier to the event (e.g. condition, procedure, episode) record for which the modifier is recorded.
modifier_of_field_concept_id	No	integer	The concept representing the table field concept that contains the value of the event id for which the modifier is recorded (e.g. Condition_Occurrence.condition_occurrence_id).

- Pros:
  - Not a new redundant structure.
  - Some modifiers can be recorded independent of a diagnosis.
- Cons:
  - Nullable foreign key.

Diagnosis Modifier Records

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**4200514 ("Adenocarcinoma of sigmoid colon")**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**
condition_source_concept_id	**36517865 ("Adenocarcinoma of sigmoid colon")**
condition_source_value	**Histology 8140/3; Topography C18.7**

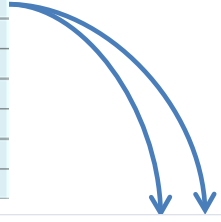


MEASUREMENT		
Field	**Content**	
measurement_id	**9996687687**	**9996687687**
modifier_of_event_id	**4325345**	**4325345**
modifier_of_field_concept_id	**condition_occurrence.condition_occurrence_id**	**condition_occurrence.condition_occurrence_id**
measurement_datetime	**February 23, 1996**	**February 23, 1996**
measurement_type_concept_id	**Pathology Report**	**Cancer Registry**
measurement_concept_id	**8084230 ("Tumor perforation presence by microscopy")**	**8084234 ("Size. Maximum dimension in tumor")**
value_as_concept_id	**5084235 ("Present")**	
value_as_number		**5.5**
unit_concept_id		**1014231 ("centimiters")**
measurement_source_concept_id		**C65 (NAACCR)**
measurement_source_value	**Tumor perforation present**	**Maximum tumor size**
value_source_value	**Yes**	**5.5 cm**

Treatment Modifier Records

EPISODE

Field	Content
episode_id	**9904650**
person_id	**John Smith**
episode_concept_id	**Treatment Regimen**
episode_start_datetime	**February 23, 1996**
episode_end_datetime	**March 23, 1996**
episode_parent_id	
episode_number	
episode_object_concept_id	**Radiotherapy Treatment**
episode_type_concept_id	**EMR**
episode_source_value	**Radiotherapy**
episode_source_concept_id	



MEASUREMENT

Field	Content	
measurement_id	**3216687687**	**3216687688**
modifier_of_event_id	**9904650**	**9904650**
modifier_of_field_concept_id	**episode.episode_id**	**episode.episode_id**
measurement_datetime	**February 23, 1996**	**February 23, 1996**
measurement_type_concept_id	**EMR**	**EMR**
measurement_concept_id	**5084230 ("Fractions")**	**5084234 ("Radiation dose")**
value_as_concept_id		
value_as_number	**43**	**7740**
unit_concept_id		**1014237 ("cgy")**
measurement_source_concept_id		
measurement_source_value	**prescribed_fractions**	**prescribed_dose_cgy**
value_source_value	**43**	**7740 cgy**

Vocabulary Extensions for Modifiers

- Add NAACCR and Nebraska Lexicon vocabularies and mappings between the two
  - Nebraska Lexicon Project terminology is modeled within SNOMED Observable entity hierarchy. Terminology sets have been completed by clinical and genomic biomarkers for breast and colorectal cancers.
  - For those cancer types not yet covered in Nebraska Lexicon, use North American Association of Central Cancer Registries (NAACCR) data dictionary concepts.
  - Mappings should be created between NAACCR and Nebraska Lexicon concepts to support ETL for standardized concepts.

Advantages of Using Modifiers

- Reflect **granularity of source data** in Cancer Registry and Pathology Synoptic Reports
- Attribute-Value structure supports **representation of any number and type of features**
- Support **explicit connection to histology/topography**

Summary

- Support research and analytic use cases
- Maximize the use of existing OMOP CDM constructs and conventions
- Reuse and extension of existing standards
  - ICD-O, SEER, NAACCR, CAP, Nebraska Lexicon
- Align cancer use case with other conditions
- Support efficient queries

Next Steps

- Ratify CDM extensions.
- Implement required terminologies in vocabulary tables.
- Develop and publish ETL instructions on Github repo.

Future Work

- Genomic Data
  - Create structures and vocabularies to house genomic results next to clinical data.
- Recurrence/Progression Detection
  - Support the open-sourcing and free exchange of algorithms to derive recurrence/progression from low-level clinical events.
- Imaging Data
  - Create structures and vocabularies to house imaging data. Support the open-sourcing and free exchange of algorithms to detect features from the combination of raw imaging data and imaging meta-data.

OHDSI Oncology CDM Extension Proposal

DETAILED PROPOSAL

OHDSI Oncology CDM Extension Proposal

Background

1. Challenges related to representation of cancer diagnosis in source data.
   Cancer diagnosis is usually recorded in two sources: Cancer/Tumor Registries and electronic medical records (EMR). In Tumor Registries, cancer diagnoses are abstracted from pathology reports and EMR. Pathology-based diagnosis is coded in ICD-O representing a combination of cancer histology and topography. ICD-O is considered a gold standard to annotate cancer diagnosis. EMR-based diagnosis is coded in ICD-9/10. Although diagnosis in Cancer Registries is most accurate and granular, in most (non-SEER) states it is only recorded for the first cancer occurrence and not recorded for recurrent cancer. In EMRs, cancer diagnosis is recorded as regular billing and problem list diagnoses using ICD-9/10 coding that is less granular than ICD-O. Therefore, tracking of cancer diagnosis at the same level of granularity throughout the course of disease is a serious challenge.

   In addition to cancer histology and topography, there are other cancer features that define cancer diagnosis and determine outcomes and treatments. Similar to histology and topography, these features are abstracted from pathology reports into Cancer Registries for the first cancer occurrence (with the exception for SEER states where all cancer occurrences are recorded in Cancer Registries). Unlike histology and topography, these additional features are rarely available in EMR in a structured form. They may be present in pathology systems. Deriving and reconciling one “clean” set of cancer attributes for each cancer occurrence is critical and challenging.

   Identifying cancer occurrences is another challenge because, with the exception of SEER states that explicitly record cancer recurrences, they are not available in a structured form in EMRs. Patterns of ICD9/10 diagnoses after initial diagnosis do not reliably track the recurrence status of a patient’s cancer diagnosis.

2. Challenges of cancer diagnosis representation in OMOP CDM and Vocabulary.
   Since in the source data cancer diagnosis is represented in ICD-O for the first occurrence and in ICD-9/10 for all occurrences, the challenge is to connect these two representations. OHDSI’s standard for diagnosis representation is SNOMED. ICD-9 and ICD-10 have been successfully mapped to SNOMED. One challenge is to validate existing mappings between ICD-O and SNOMED CT and propose new SNOMED CT coding to cover all the existing ICD-O histology and topography combinations. The other challenge is to reconcile SNOMED diagnosis resulted from mappings from ICD-O and ICD-9/10.

   Cancer histology-topography must be linked to other key diagnostic features like stage, and grade. Although there are a few common features for many cancer types (e.g. stage, grade), their values vary and other features (e.g. tumor size, laterality, biomarkers) are specific to each cancer type. There are presently two attribute-value tables that may house cancer diagnostic features, Observation and Measurement. However, none of them has an explicit

linkage to Condition_Occurrence. Implications on extending either of these tables to support the linkage and store modifier tables must be carefully evaluated.

From the vocabulary stand point, many cancer diagnosis features are covered in LOINC and SNOMED CT. The challenge is to create a set of these concepts for each cancer type (e.g. breast, lung, etc.). There is an ongoing initiative that intends to create and align these sets with the ICCR (International Collaboration on Cancer Reporting) data sets. However, at this point, it has only covered a few cancer types.

3. Treatment clinical event welter.
   Oncology treatments often create a clinical event welter that thwarts many analytic use cases. Instead, we want a concept that aggregates lower-level clinical events into a higher-level abstraction. Some examples of clinical event welter:
   - Beam IMRT to left breast 15 Fractions at 267 cGy Dose over 27 days.
     - Spawns 72 entries in the PROCEDURE_OCCURRENCE table across 11 CPT codes.
   - Docetaxel + Carboplatin, 21-DAY cycle, 6 cycles over 115 days
     - Spawns 22 entries in the PROCEDURE_OCCURRENCE table across 5 CPT codes and 50 entries in the DRUG_EXPOSURE table across 13 RxNorm codes.

In the oncology community, there is a widely shared *treatment* concept: tumor registries, practice guidelines, clinical trials databases and oncology analytic platforms all employ the concept of a *treatment*. The *treatment* concept aggregates lower-level clinical events into a higher-level abstraction. But we still want this higher-level abstraction *treatment* concept to connect to lower-level clinical events.

Can we have both the benefit of a higher-level oncology treatment abstraction and connection to lower-level clinical events? Can we find a place for pre-made oncology *treatment* abstractions connected to lower-level clinical events? Conversely, can we derive higher-level oncology *treatments* from lower-level clinical events ? Finally, can we attach unconnected pre-made oncology *treatment* abstractions to lower-level clinical events?

Adding support for the representation of *treatment* abstractions within the OMOP CDM will enable the following use cases:

- Classify each treatment at a level intuitive to oncology professionals/researchers.
- Enumerate how many oncology treatments have been performed on a patient.
- Characterize when each treatment begins and ends.
- Describe when a treatment is “switched”.
- Reuse the grouping of low-level clinical events present in source systems. Not lose *treatments* abstractions during conversion into the OMOP CDM.
- A target for the algorithmic derivation of treatment abstractions when not present in our source systems

- Harmonize EHR/claims database oncology treatment data and tumor registry oncology treatment data.
- Attribute properties to an oncology treatment as a whole.

4. Absence of abstraction layer representing clinician's and researcher's view
   Clinicians and researches view cancer as a chronic disease with a series of disease episodes: first occurrence, remission, relapse, end of life event. Cancer is maintained with a treatment course comprised of one or more treatment modalities, multiple regimens, and cycles. Disease progression is monitored at prescribed intervals and reported as outcomes (e.g stable disease). These abstractions of disease, treatment, and outcomes are rarely available in the source data. Derivation and persistence of these abstractions are critical since they are key variables for prediction of disease progression, disease free and overall survival.

   Currently, Condition_Era and Drug_Era are the structures that house derived condition and drug exposure episodes. However, neither their attributes nor algorithms of their derivation will support complex abstractions of cancer disease and treatment.

I. Overall Approach

- **Cancer diagnosis**
  - In our current approach, we define cancer diagnosis as a combination of **histology** (morphology) + **topography** (anatomy)
- **Diagnosis modifiers**
  - Diagnostic and treatment features that vary between different cancer diagnoses and treatments are represented as modifiers and explicitly linked to the respective diagnosis or treatment
  - Examples of diagnosis modifiers are stage, grade, laterality, foci, tumor biomarkers. These diagnostic features are assessed when a patient is first diagnosed and also (possibly) for each cancer recurrence. Repeated measurements of the same modifier (lymph node invasion) may be recorded. Different modifiers may be recorded on different dates
  - Examples of treatment modifiers are surgery laterality, radiotherapy dosage and frequency.
- **Disease and treatment abstraction layer**
  - Disease and treatment abstractions will be modeled as episodes, a new CDM construct that can be used to represent other abstractions such as episode of care.

- These abstraction may be derived algorithmically pre- or post-ETL or extracted from the source data directly. In addition to the regular OMOP type_concept_ID, we propose to store references to the derivation algorithms in the vocabulary.
- Disease abstractions include first occurrence, remissions, relapses, and end of life event.
- There should be one set of "verified" cancer modifiers associated with each cancer occurrence and relapse.
- Treatment abstractions include treatment course, treatment regimen, and treatment cycle.

II. Representing cancer diagnosis as histology and topography combination

The International Classification of Diseases for Oncology (ICD-0)[1] is a dual-axis vocabulary used to identify cancer topography (anatomic site) and histology (morphology) to track and report cancer incidence, survival and mortality.

The topography code describes the anatomical site of origin of the neoplasm. The code always has a prefix of "C", followed by a three digit number that indicates the site (two digits) and the subsite (one digit), separated by a decimal point. Example: C18.4: the C18 indicates that the site is the colon and the 4 indicates that the sub-site is the transverse colon.

The histology code describes the characteristics of the tumor itself, including its cell type and biological activity. The code is composed of four digits that indicate the cell type or histology and one digit that indicates the behavior. The first four digits are separated from the last (behavior) digit by a forward slash (/). The behavior digit can be: 0 (benign), 1 (uncertain behavior), 2 (carcinoma in situ), 3 (malignant, primary site), 6 (malignant, metastatic site), 9 (malignant, uncertain whether primary or metastatic site), Examples: Squamous cell carcinoma in situ, NOS = 8070/2, Adenocarcinoma, NOS = 8140/3, Carcinoma, NOS = 8010/3.

Each combination of these two dimensions, histology and topography, rolls up to a unique cancer diagnosis. For example, Carcinoma, NOS (8010/3) and 'Unspecified part of bronchus or lung' (C34.9) rolls to 'Carcinoma of the lung'. However, there is no single code to annotate this unique cancer diagnosis. National Cancer Institute (NCI) Surveillance, Epidemiology, and End Result Program (SEER) provide validation lists for "coherent" topography/morphology or site/histology combinations[2].

To represent cancer diagnosis, the combination of histology and topography, in the OMOP CDM Condition domain (Condition_Occurrence) without changes of the existing structure, we propose to perform a pre-coordinated collapse of the ICD-O axes, histology and topography, to a single OMOP originated concept representing unique cancer diagnosis and preserve linkages between these single codes and the ICD-O axes in the OMOP vocabulary[3].

Our proposed approach will support adherence to the OMOP CDM/Vocabulary conventions:

- One required _concept_id field will be populated in the corresponding domain table, Condition_Occurrence.
- Vocabulary-related attributes are stored in a vocabulary data model in a uniform way

If a new proper SNOMED code is created, the OMOP-originated concept can be easily replaced by it.

A similar mapping approach is used for representing LOINC/HL7 clinical note types and CDO ontology in the OMOP CDM NLP tables.

Detailed implementation

1. Create pre-coordinated source (non-standard) concepts in the Concept table representing unique cancer diagnosis by collapsing the two ICD-O axes, histology and topography into unique concepts using SEER validation lists https://seer.cancer.gov/icd-o-3/. For example, a new pre-coordinated concept derived by collapsing '*Adenocarcinoma, NOS'* (8140/3) and '*Sigmoid colon'* (C18.7) will be '*Adenocarcinoma of sigmoid colon'* (8140/3- C18.7):

Field	**Record**
concept_id	36517865
concept_name	Adenocarcinoma of sigmoid colon
concept_code	8140/3- C18.7
vocabulary_id	ICDO3

2. Create mapping between the new pre-coordinated source concept and a standard SNOMED concept '*Adenocarcinoma of sigmoid colon'* (301756000) in Concept_Relationship table

Field	Record 1	Record 2
concept_id_1	36517865	4200514
concept_id_2	4200514	36517865
relationship_id	*Maps to*	*Is mapped to*

3. Each pre-coordinated SNOMED concept is linked to morphology/histology ('*Has associated morphology*') and anatomic site/topography ('*Has finding site'*) in the Concept_Relationship table thus supporting analysis along those axes:

concept_id_1	concept_id_2	relationship_id
4200514	4290838	*Has associated morphology*
4200514	4244588	*Has finding site*

Where 4290838 represents a SNOMED concept of '*Malignant adenomatous neoplasm – category*' and 4244588 represents a SNOMED concept of '*Sigmoid colon structure'*.

4. Detailed ETL instructions are provided in the Appendix.

III. CDM Representation and Extension

1. Overview

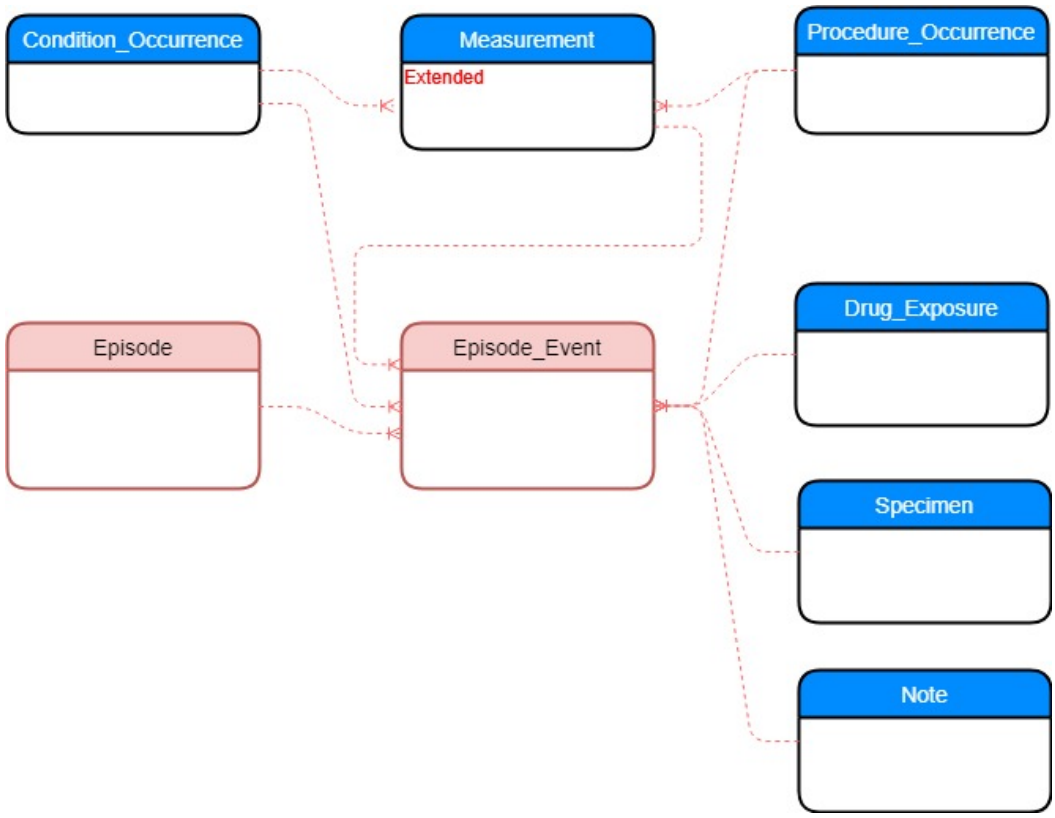


Fig 1. High level ERD

Existing tables are depicted in blue and new tables and relationships in red.

- Cancer diagnoses are stored in CONDITION_OCCURRENCE as pre-coordinated concepts combining histology and topography.
- Cancer treatment clinical events are stored in the PROCEDURE_OCCURRENCE and DRUG_EXPOSURE tables.
- Disease and treatment abstractions (e.g. first cancer occurrence, treatment regimen hormonal therapy) are represented in the new EPISODE table.
- Links between the disease and treatment episodes and the underlying events (conditions, procedures, drugs) are stored in the new EPISODE_EVENT table.

- Additional diagnostic and treatment features are stored in the MEASUREMENT table as modifiers of the respective condition, treatment, or episode. MEASUREMENT table is extended to include a reference to the condition, treatment, or episode record.

2. Representing cancer disease and treatment abstractions as episodes

New EPISODE table

Episode represents disease and treatment abstractions like first disease occurrence or treatment regimen derived algorithmically or extracted directly from the source data. This table can be also used to represent other abstractions such as episode of care.

Field	Required	Type	Description
episode_id	yes	integer	A unique identifier for each Episode.
person_id	yes	integer	A foreign key identifier to the Person who is undergoing the Episode. The demographic details of that Person are stored in the PERSON table.
episode_concept_id	yes	integer	A foreign key that refers to a standard Episode Concept identifier in the Standardized Vocabularies. Examples of an Episode Concept can be: Treatment Regimen, Treatment Cycle, Disease First Occurrence, Remission, Relapse, Episode of Care
episode_start_datetime	yes	date	The date and time on which the Episode was started.
episode_end_datetime	yes	date	The date and time on which the Episode was ended.
episode_parent_id	no	integer	A foreign key that refers to a parent Episode entry representing an entire episode if the episode spans multiple cycles.
episode_number	no	integer	An ordinal count for an Episode that spans multiple times
episode_object_concept_id	yes	integer	A foreign key that refers to a concept identifier in the Standardized Vocabularies describing disease, treatment, or other abstraction that the episode describes. For example, ‘Breast Carcinoma’ or ‘Chemotherapy’.

Field	Required	Type	Description
episode_type_concept_id	yes	integer	A foreign key that refers to a standard Episode Type Concept identifier in the Standardized Vocabularies reflecting the provenance of the episode derivation. It may reference a derivation algorithm, sources such as cancer registry, EMR, etc.
episode_source_value	no	varchar(50)	The source code for the Episode as it appears in the source data. This code is mapped to a standard episode Concept in the Standardized Vocabularies and the original code is, stored here for reference.
episode_source_concept_id	no	integer	A foreign key to a Episode Concept that refers to the code used in the source.

Conventions

- Valid Episode Concepts belong to the 'Episode' domain.
- Valid Episode Type Concepts belong to the 'Episode Type' vocabulary in the 'Type Concept' domain.
- Valid Episode Object Concepts belong to different domains based on the corresponding concept class/vocabulary of the Episode Concept.
    - 'Disease Episode':
        - 'Condition' domain
    - 'Treatment Episode'
        - 'Procedure/Treatment' domain
    - 'Episode of Care Episode'
        - 'Episode of Care' domain.

Vocabulary Extensions to Represent Episodes

1. Add 'Episode' domain.

concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
Disease Episode	Episode	Episode	Disease Episode	S	OMOP generated
Treatment Regimen Episode	Episode	Episode	Treatment Episode	S	OMOP generated
Treatment Cycle Episode	Episode	Episode	Treatment Episode	S	OMOP generated
Episode of Care Episdoe	Episode	Episode	Episode	S	OMOP generated

2. Add 'Episode Type' vocabulary.

concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
Pre-made episode in source system	Type Concept	Episode Type	Episode Type	S	OMOP generated
Algorithmically-derived episode pre-ETL	Type Concept	Episode Type	Episode Type	S	OMOP generated
Algorithmically-derived episode post-ETL	Type Concept	Episode Type	Episode Type	S	OMOP generated

3. Add concepts to new 'Procedure/Treatment' Domain. Based the entries on NAACCR/SEER treatment variables

concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
Chemotherapy	Procedure/ Treatment	Procedure/ Treatment	Drug Treatment	S	OMOP generated
Homonal Therapy	Procedure/ Treatment	Procedure/ Treatment	Drug Treatment	S	OMOP generated
Immunotherapy	Procedure/ Treatment	Procedure/ Treatment	Drug Treatment	S	OMOP generated
External beam, NOS	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
External beam, photons	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
External beam, protons	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
External beam, electrons	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
External beam, neutrons	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
External beam, carbon ions	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Brachytherapy, NOS	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated

concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
Brachytherapy, intracavitary, LDR	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Brachytherapy, intracavitary, HDR	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Brachytherapy, Interstitial, LDR	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Brachytherapy, Interstitial, HDR	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Brachytherapy, electronic	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Radioisotopes, NOS	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Radioisotopes, Radium-232	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Radioisotopes, Strontium-89	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated
Radioisotopes, Strontium-90	Procedure/ Treatment	Procedure/ Treatment	Radiation Therapy Treatment	S	OMOP generated

4. **Placeholder for NAACCR treatment data dictionary items.**

5. **Placeholder for Observational Research in Oncology Toolbox (OROT) classification vocabulary.**

6. **Placeholder for entries in CONCEPT_RELATIONSHIP between 'Procedure/Treatment' domain, NAACCR data dictionary items and OROT classification vocabulary.**

New EPISODE_EVENT table

Episode_Event stores links between cancer disease and treatment episodes and the underlying events (conditions, procedures, drugs, etc.).

Field	Required	Type	Description
episode_id	yes	integer	A foreign key identifier to the Episode that the Episode Event belongs to.

Field	Required	Type	Description
visit_occurrence_id	no	integer	A foreign key identifier to the visit_occurrence record for which an episode is recorded.
condition_occurrence_id	no	integer	A foreign key identifier to the condition_occurrence record for which an episode is recorded.
procedure_occurrence_id	no	integer	A foreign key identifier to the procedure_occurrence record for which an episode is recorded.
drug_exposure_id	no	integer	A foreign key identifier to the drug_exposure record for which an episode is recorded.
device_exposrue_id	no	integer	A foreign key identifier to the device_exposur record for which an episode is recorded.
measurement_id	no	integer	A foreign key identifier to the measurement record for which an episode is recorded.
specimen_id	no	integer	A foreign key identifier to the speciment record for which an episode is recorded.
observation_id	no	integer	A foreign key identifier to the observation record for which an episode is recorded.
note_id	no	integer	A foreign key identifier to the note record for which an episode is recorded.
cost_id	no	integer	A foreign key identifier to the cost record for which an episode is recorded.

Conventions

- One record in the EPISODE_EVENT table represents a link between one episode and one event: only one event reference field is populated. For example, to represent a link between first cancer occurrence stored in the Episode table and initial diagnosis record stored in CONDITION_OCCURRENCE, only episode_id and condition_occurrence_id fields are populated with respective values.
- Some episodes may not have links to the underlying events. For such episodes, EPISODE_EVENT table is not populated.

3. Representing cancer diagnosis features as diagnosis modifiers and treatment features as treatment modifiers.

Extending MEASUREMENT table

The Measurement table may contain cancer diagnosis, treatment, and episode modifiers such

as cancer stage, grade, lymph node involvement, tumor size, tumor biomarkers, radiotherapy total dose and others (numeric or categorical) obtained through laboratory tests, imaging, and pathology reports.

To explicitly link cancer diagnosis, treatment, or episode record to its modifier, we propose to add the following fields to the Measurement table:

Field	Required	Type	Description
modifier_of_event_id	No	integer	A foreign key identifier to the event (e.g. condition, procedure, episode) record for which the modifier is recorded.
modifier_of_field_concept_id	No	integer	The concept representing the table field concept that contains the value of the event id for which the modifier is recorded (e.g. Condition_Occurrence.condition_occurrence_id).

Conventions

- Modifier records are similar to regular measurement records in that they require a standardized test or some other activity to generate a quantitative or qualitative result. However, modifiers are not independent measurements but modifiers which add specificity to cancer diagnosis, treatment, or episode. For example, LOINC 44648-4 'Histologic grade' may modify cancer diagnosis of “Tubular carcinoma” recorded in CONDITION_OCCURRENCE. Therefore, although modifier_of_event_id and modifier_of_table_concept_id are not required fields, they must be populated for modifiers.
- Repeated modifier records (lymph node invasion) may be associated with one or multiple condition occurrence records.
- Modifiers for the same condition record may be recorded on different dates.
- One set of “verified” cancer modifiers must be associated with a disease or treatment episode.
- Valid Concepts for the value_as_concept field normally belong to the ‘Meas Value’ domain.

Vocabulary Extensions to Represent Cancer Diagnosis Modifiers and Treatment Modifiers

OHDSI Oncology CDM Extension Proposal

1. Standardized diagnosis modifier terminology from Nebraska Lexicon Project:

We recommend leveraging standard terminology developed by Nebraska Lexicon Project [4]. This initiative intends to implement CAP (College of American Pathologists) Protocol Templates[5] by providing terminology binding between LOINC and SNOMED CT**.** The majority of the associated terminology development is modeled within SNOMED Observable entity hierarchy. Coded LOINC observables are linked to SNOMED value sets. Terminology sets have been completed by clinical and genomic biomarkers for breast and colorectal cancers.

a. For those cancer types not yet covered in Nebraska Lexicon, we will be using North American Association of Central Cancer Registries (NAACCR) data dictionary concepts to represent cancer diagnostic features.
b. Mappings should be created between NAACCR and Nebraska Lexicon concepts to support ETL for standardized concepts.

2. Placeholder for Standardized treatment modifier terminology from NAACCR data dictionary.

3. Add concepts to the 'Meas Type' vocabulary in the 'Type Concept' domain;

concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code
Cance Registry	Type Concept	Meas Type	Meas Type	S	OMOP generated

OHDSI Oncology CDM Extension Proposal

EXAMPLES OF CANCER DATA IN OMOP CDM

1. Cancer diagnosis record

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**4200514 ("Adenocarcinoma of sigmoid colon")**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**
condition_source_concept_id	**36517865 ("Adenocarcinoma of sigmoid colon")**
condition_source_value	**Histology 8140/3; Topography C18.7**

2. Diagnosis modifier record

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**4200514 ("Adenocarcinoma of sigmoid colon")**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**
condition_source_concept_id	**36517865 ("Adenocarcinoma of sigmoid colon")**
condition_source_value	**Histology 8140/3; Topography C18.7**

MEASUREMENT		
Field	**Content**	
measurement_id	**9996687687**	**9996687687**
event_id	**9900145**	**9900145**
event_domain_id	**CONDITION_OCCURRENCE**	**CONDITION_OCCURRENCE**
measurement_datetime	**February 23, 1996**	**February 23, 1996**
measurement_concept_id	**8084230 ("Tumor perforation presence by microscopy")**	**8084234 ("Size. Maximum dimension in tumor")**
value_as_concept_id	**5084235 ("Present")**	
value_as_number		**5.5**
unit_concept_id		**1014231 ("centimiters")**

3. Diagnosis episode and related event records

EPISODE	
Field	**Content**
episode_id	**4325345**
person_id	**John Smith**
episode_concept_id	**First Occurrence**
episode_start_datetime	**February 14, 1996**
episode_end_datetime	**November 18, 1996**
episode_object_concept_id	**Adenocarcinoma of sigmoid colon**
episode_type_concept_id	**Algorithm #123**

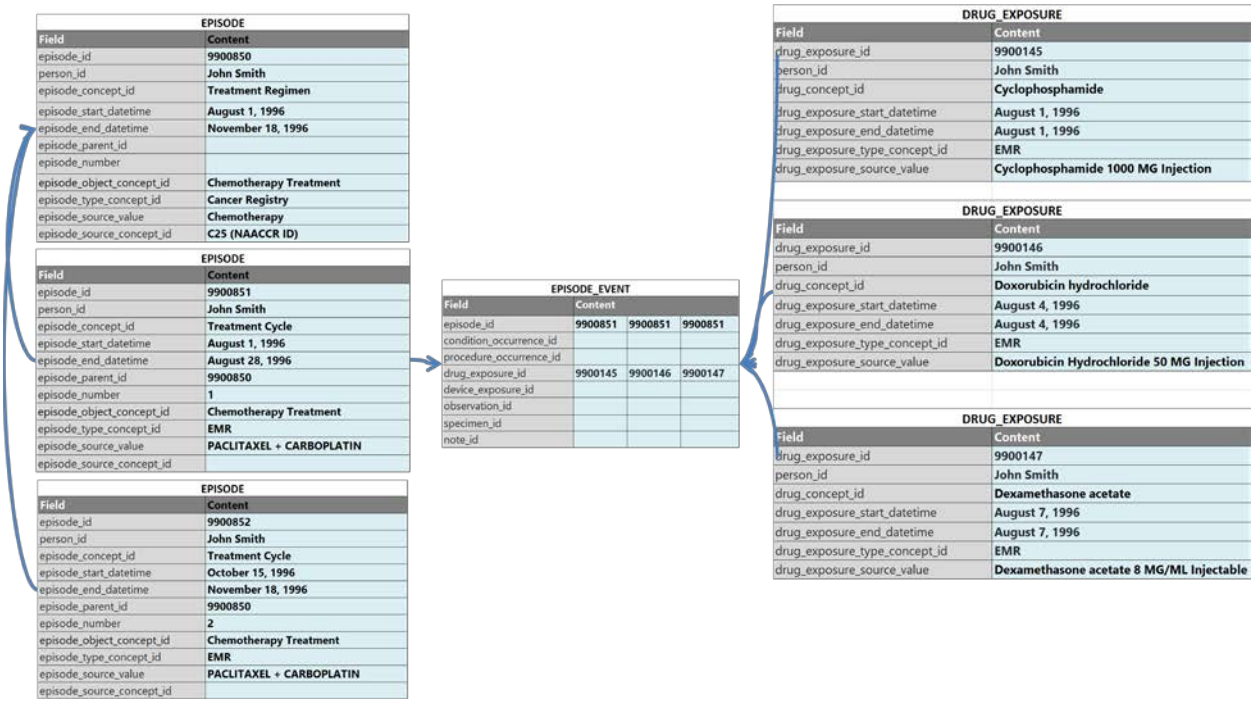
EPISODE_EVENT			
Field	**Content**		
episode_id	**4325345**	**4325345**	**4325345**
condition_occurrence_id	**9900145**	**9900850**	
procedure_occurrence_id			**456774870**
drug_exposure_id			
specimen_id			
note_id			

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900145**
person_id	**John Smith**
condition_concept_id	**Adenocarcinoma of sigmoid colon**
condition_occurrence_start_datetime	**February 14, 1996**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**Cancer Registry**

CONDITION_OCCURRENCE	
Field	**Content**
condition_occurrence_id	**9900850**
person_id	**John Smith**
condition_concept_id	**Adenocarcinoma of sigmoid colon**
condition_occurrence_start_datetime	**September 15, 1999**
condition_occurrence_end_datetime	
condition_occurrence_type_concept_id	**EMR**

PROCEDURE_OCCURRENCE	
Field	**Content**
procedure_occurrence_id	**456774870**
person_id	**John Smith**
procedure_concept_id	**Intravenous chemotherapy**
procedure_occurrence_start_datetime	**November 1, 1996**
procedure_occurrence_end_datetime	**November 18, 1996**
procedure_occurrence_type_concept_id	**EMR**

4. Treatment episode and related event records



EPISODE

Field	Content
episode_id	9900850
person_id	John Smith
episode_concept_id	Treatment Regimen
episode_start_datetime	August 1, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	
episode_number	
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	Cancer Registry
episode_source_value	Chemotherapy
episode_source_concept_id	C25 (NAACCR ID)

EPISODE

Field	Content
episode_id	9900851
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	August 1, 1996
episode_end_datetime	August 28, 1996
episode_parent_id	9900850
episode_number	1
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE

Field	Content
episode_id	9900852
person_id	John Smith
episode_concept_id	Treatment Cycle
episode_start_datetime	October 15, 1996
episode_end_datetime	November 18, 1996
episode_parent_id	9900850
episode_number	2
episode_object_concept_id	Chemotherapy Treatment
episode_type_concept_id	EMR
episode_source_value	PACLITAXEL + CARBOPLATIN
episode_source_concept_id	

EPISODE_EVENT

Field	Content		
episode_id	9900851	9900851	9900851
condition_occurrence_id			
procedure_occurrence_id			
drug_exposure_id	9900145	9900146	9900147
device_exposure_id			
observation_id			
specimen_id			
note_id			

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900145
person_id	John Smith
drug_concept_id	Cyclophosphamide
drug_exposure_start_datetime	August 1, 1996
drug_exposure_end_datetime	August 1, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Cyclophosphamide 1000 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900146
person_id	John Smith
drug_concept_id	Doxorubicin hydrochloride
drug_exposure_start_datetime	August 4, 1996
drug_exposure_end_datetime	August 4, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Doxorubicin Hydrochloride 50 MG Injection

DRUG_EXPOSURE

Field	Content
drug_exposure_id	9900147
person_id	John Smith
drug_concept_id	Dexamethasone acetate
drug_exposure_start_datetime	August 7, 1996
drug_exposure_end_datetime	August 7, 1996
drug_exposure_type_concept_id	EMR
drug_exposure_source_value	Dexamethasone acetate 8 MG/ML Injectable

References

1. ICDO-3 vocabulary
   http://codes.iarc.fr/usingicdo.php
2. SEER histology and topography combinations
   https://seer.cancer.gov/icd-o-3/.
3. Proposed approach for mapping ICDO-3 to SNOMED
   http://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=documentation:oncology:poster2018-improvement_of_cancer_diagnosis_representation_in_omop_cdm3_1_.pdf
4. Nebraska Lexicon
   https://www.unmc.edu/pathology/informatics/tdc.html
5. CAP Protocol Templates
   http://www.cap.org/web/oracle/webcenter/portalapp/pagehierarchy/cancer_protocol_templates.jspx
6. FORDS
   https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/fords%202016.ashx

Appendix A

ETL instructions for mapping ICD-O to SNOMED

COMPLETE ICD-O SOURCE CODES

Cancer diagnoses are usually represented by a combination of ICD-O-3 histology and topography codes. To map this combination to SNOMED follow these steps:

1. Transform diagnosis SOURCE VALUE
    a. Histology code. In the source, it is normally formatted like this: 8070/3, where 8070 is histology type and 3 is tumor behavior. If histology type and behavior are stored separately, concatenate them to get one histology concept, e.g. 8070/3.
    b. Topography code. the source, it is normally formatted like this: C50.2. Be aware of the dot. if the source doesn't have the dot, insert it after the 3d character: C502 -> C50.2. If the source code contains only 3 characters, the dot is not required: C50 -> C50.
    c. Source value. Concatenate histology code and topography code using hyphen: 8070/3-C50.2. This value will be stored in the CONDITION_OCCURRENCE.CONDITION_SOURCE_VALUE field.

2. Extract value of diagnosis SOURCE CONCEPT ID
   Concept ID for the combined histology/topography code is stored in the CONCEPT table. The following SQL shows how to extract its value for the above example:

```
SELECT CONCEPT_ID
FROM CONCEPT
WHERE CONCEPT_CODE = '8070/3-C50.2'
        AND VOCABULARY_ID = 'ICDO3'
```

   The resulting value 36517865 will be stored in the CONDITION_OCCURRENCE.CONDITION_SOURCE_CONCEPT_ID field and will be used in mapping to a standard SNOMED code (next section).

3. Extract value of STANDARD CONCEPT ID
   Source concept ID of the combined histology/topography code is mapped to a standard concept ID in the CONCEPT_RELATIONSHIP table. The following SQL shows how to extract its value for the above example:

```
SELECT CONCEPT_ID_2
FROM CONCEPT_RELATIONSHIP
WHERE CONCEPT_ID_1 = 36517865
        AND RELATIONSHIP_ID = 'Maps to'
```

   The resulting value [36517865] will be stored in the CONDITION_OCCURRENCE.CONDITION_CONCEPT_ID field.

INCOMPLETE ICD-O SOURCE CODES

In some cases when the source data are incomplete, apply the following approach.

1) Tumor behavior is not known
   Use 1 (uncertain behavior) to making your code complete: 8070 -> 8070/1

2) Topography is unknown.
   Use mappings from this file https://seer.cancer.gov/tools/conversion/ICD03toICD9CM-ICD10-ICD10CM.xls (last 3 tabs of this file) to obtain topography if you have ICD-10 code for this diagnosis. Note, if you have long ICD-10CM code, you need to cut it off to have only 5 symbols (including dot): C50.211 -> C50.2
   In case when a patient has several cancer diagnoses, use ICD-10 from the date closest to the ICD-O histology date.

Appendix B

ETL instructions for mapping Treatment abstractions.

The varying levels of grouping/abstraction of lower-level clinical events into TREATMENTS available within source systems will require different ETL strategies.

1. Oncology EHR contains TREATMENT groupings/abstractions natively.
   - No algorithmic derivation necessary. Use OROT to map clinical event codes to TREATMENT concepts. Insert low-level clinical events, the grouping/abstraction structures and the connections between them.
2. EHR records administrations/prescriptions of the drugs in a chemotherapy regimen or each fraction of a radiation therapy treatment. No grouping/abstractions natively.
   - Insert the low-level clinical events. Algorithmically derive TREATMENT abstractions/groupings and connections between them. Use OROT to map clinical event codes to TREATMENT concepts.
3. Tumor Registry records that a chemotherapy regimen or a radiation therapy treatment occurred and an EHR records administrations/prescriptions of the drugs in a chemotherapy regimen or each fraction of a radiation therapy treatment. No grouping/abstractions natively.
   - Insert low-level clinical events and the grouping/abstraction structures. Use OROT to algorithmically derive connections between TREATMENT abstractions/groupings and low-level clinical events.
4. Tumor Registry records that a chemotherapy regimen or a radiation therapy treatment occurred.
   - Insert only into the TREATMENT grouping/abstraction structure.