Getting your NAACCR data into the OMOP CDM

The previous documentation provided a framework for helping decide whether you should ETL your source oncology data into the OMOP CDM, offering both the advantages and the tradeoffs of doing so. This section provides instructions for how to accomplish the ETLing of your NAACCR data if you choose to move forward.

Goal of this documentation:

1. To be used as a vademecum by sites that wish to take their source oncology data, i.e. a cancer registry, and import them into the OMOP CDM.
2. For the ETLer to understand the choices that have been already made–and already embedded into the ETL–and the choices that are left to the ETLer himself/herself.

Important caveats for who can (and who cannot) use the instructions that follow:

This guideline is for you if:

1. Your have a NAACCR dataset in a ***fixed width format***.
    a. If your NAACCR dataset is in an XML format, please see *THIS GUIDELINE*.
    b. NAACCR datasets of either version 15, 16 or 18 are the versions that have fixed width format.
2. The NAACCR dataset in the ETLer's possession has been pulled with only **RECORD TYPE = 'C'.**
    a. Column 1 of your institution's NAACCR data table is labeled RECORD TYPE and each row can take a value of either I, C, A, U, M or L.
    b. The dataset with which you initiate the ETL workflow should have only rows where RECORD TYPE = 'C'.
3. Your institution has an established OMOP CDM, preferably >= version 5.3.

If 1, 2 and 3 of the above are true, you are in good shape (as far as this guideline is concerned) and may continue.

The ETL script itself – what the ETLer needs to know:

NAACCR data elements that will be ETLed into the OMOP CDM:

1. AJCC TNM Staging
2. Grading
3. Metastases
4. Date of Initial Diagnosis

Vocabularies used by the ETL script:

1. Cancer Modifier vocabulary

2. ICD-O-3
3. SNOMED

OMOP tables that are the target tables of the ETL script:

1. **CONDITION_OCCURRENCE** table: The ‘Diagnosis’ and the ‘Date of Initial Diagnosis’ concepts in your source data will be mapped, via concept_IDs in the Cancer Modifier vocabulary, to the CONDITION_OCCURRENCE table. The Diagnosis concept represents a topography (where) and histology (what) diagnosis pairing.
2. **MEASUREMENT** table: The staging, grading and metastases data points in your source data will be mapped, via concept IDs in the Cancer Modifier vocabulary, to the MEASUREMENT table. These concepts represent modifications to the cancer diagnosis record in the CONDITION_OCCURRENCE table.

In order to execute the OHDSI oncology ETL SQL script, your NAACCR data must first be transformed from a wide format to a long format. The ETL script won’t be successful if the data are not in this starting format. In order to get your NAACCR data into this format, please follow the instructions in the link provided here: https://github.com/OHDSI/OncologyWG/tree/master/NaaccrParser. This link takes you to the repository of the OHDSI R package, called NaaccrParser, that accomplishes this data transformation.

Information for execution of the NaaccrParser R package:

1. The **‘record_id_index**’ can take any arbitrary value. This variable is used as a schema of sorts to retain linkage between patients and their corresponding observations. This is needed since the R script is essentially flipping an EAV structure from wide format to long format (wide format = one row per patient with many columns; long format = many rows per patient with fewer columns).

Once the NaaccrParser R package has been successfully executed against your source dataset, you are now ready to execute the ETL script that then ingests the data from this intermediate dataset into the OMOP CDM.