# Transforming the Optum® Enriched Oncology module to OMOP CDM

**Dmitry Dymshyts[1], Clair Blacketer[2]**
**[1]Janssen Research & Development, Raritan, NJ; [2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL**

## Background

The Optum® Enriched Oncology Data set is a group of tables that can supplement the Optum® de-identified Electronic Health Record dataset. It contains additional cancer-specific information on a subset of the EHR population that has at least one solid or non-solid tumor diagnosis. In addition to structured data obtained from medical records, the enriched oncology tables provide clinical information extracted from provider notes using natural language processing (NLP) machine learning models (for example, tumor progression, histology, etc.).

Optum oncology data initiatives include enriching Optum EHR data by extracting essential information from the oncology patient's medical records and making it usable for researchers. Specific oncology concepts important for understanding the progression of the disease are often not available in structured formats, particularly the tumor, node, and metastasis (TNM) values, stage information and biomarkers.

As of 2022, there are approximately 1.9 million patients with at least one solid tumor ICD-9 or ICD-10 diagnosis included in the data set. Given the potential for research, the dataset was converted to the OMOP CDM to allow integration with standard OHDSI methods and tools.

## Methods

The data mapping was done with help of the White Rabbit and Rabbit-in-a-hat tools.[1] The concept mapping was done by a semantic analysis of source concepts. Data and concepts mapping is based on the OMOP Oncology Working group guidelines.[2] Please see the mapping logic below:

Table 1. *Source values mapping to OMOP CDM*

| **source table** | **target domain** | **target vocabulary** |
|---|---|---|
| Histology | Condition | SNOMED, ICDO3 |
| Topography | Condition | SNOMED, ICDO3 |
| Laterality | Condition | SNOMED, ICDO3 |
| Behavior (in situ, malignant or benign) * | Condition | SNOMED, ICDO3 |
| summary stage | Measurement | Cancer Modifier |
| metastasis location | Measurement | Cancer Modifier |
| TNM | Measurement | Cancer Modifier |
| tumor grade | Measurement | Cancer Modifier |
| characteristics: advanced, carcinomatosis, extensive, infiltrative, invasive, localized, etc | Measurement | Cancer Modifier |
| Biomarkers** | Measurement | OMOP genomic, LOINC, SNOMED |
| evaluation system: Binet Stage, Durie/Salmon Stage, ECOG performance status, FIGO Stage, Gleason, Gleason score | Measurement | Cancer Modifier |
| Tumor size*** | Measurement | Cancer Modifier |
| treatment regimen | Episode | HemOnc |
| treatment response | Observation | no mapping - not supported by the vocabulary model |
| Tumor progression | Observation | no mapping – will be in Episode table in a future |

*Notes:*

*Behavior (in situ, malignant or benign) and laterality are the part of the Diagnosis (Condition), not a Cancer modifier (Measurement). This mimics the structure of the ICD-O-3 and SNOMED vocabularies used as standard vocabularies for Oncology diagnoses in OMOP. So, the source condition concept is a result of precoordination of Histology, Topography, Behavior and Laterality.

** If the mutation status result is reflected as expression level of immunostaining (“0”, “1+”, “2+”, “3+”), it is mapped to negative, equivocal, or positive, depending on the type of biomarker, for example in EGFR “2+” stands for positive and in ERBB2 “3+” stands for Positive.

*** If there are several dimensions given, the ETL will identify the largest one and put it in a CDM as a “*Largest Dimension of Tumor”,* while other dimensions are reflected by “*Dimension of Tumor”* concept.

- To distinguish the data derived from the Enhanced oncology module from the core OPTUM EHR data, type_concept *“Standard algorithm from EHR”* was introduced.

Since the data are derived using an NLP engine, some cleaning is necessary.

We filtered out the entries where “in situ” and “invasive” characteristics exist at the same day. These modifiers belong to the cancer diagnoses with the same or semantically close (for example, “ductal carcinoma” and “carcinoma”) description, so in most of the cases “in situ” and “invasive” characteristics are related to the same cancer in a same day, which looks like an error in the data. In situ means that cancer cells haven’t spread to nearby tissue, while invasive means the opposite - the cancer has spread beyond the layer of tissue in which it developed and is growing into surrounding, healthy tissues.

We filtered out the entries where numeric and narrative biomarkers results are inconsistent, for example numeric result = “+1”, but the narrative result = “positive mutation” in ERBB2/HER2 measurement. A score of "1+" suggests that there is a low level of HER2 protein present in the cells. This low level is considered within the normal range, and so the cancer is unlikely to respond to therapies that target HER2. Therefore, a "1+" score is usually interpreted as a negative result for HER2 overexpression.

Event tables were deduped if at the same date there was the same information (condition_source_value in Conditions, combination of measurement_source_value, value_as_number, value_source_value in Measurement).

Since source concepts are text entries, they require a manual mapping to a standard terminology. These mappings were challenging sometimes, as shown in Table 2 and Table 3. In table 2, the text entry to map was “**Erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu)”.** For this biomarker, the gene amplification is measured by Fluorescent in situ hybridization and protein expression – by immunohistochemistry.

Since the method is unknown, we map to the more generic, “gene variant measurement” concept, which subsumes Gene Amplification and Protein Expression measurements.

Table 2. *“**Erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu)”** mapping and alternative options*

| Source concept | Target concept name | Target vocabulary id |
|---|---|---|
| Erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu) | ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement | OMOP Genomic |
| Erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu) | ERBB2 Gene Amplification measurement | OMOP Genomic |
| Erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu) | ERBB2 Protein Expression measurement | OMOP Genomic |

*Note, the mapping we end up with is highlighted in green.*

Table 3. ***“Estrogen receptor (ER)”** mapping and alternative options*

| Source concept | Target concept name | Target vocabulary id |
|---|---|---|
| Estrogen receptor (ER) | ESR1 Protein Expression measurement | OMOP Genomic |
| Estrogen receptor (ER) | Estrogen Receptor (ER) Assay | NAACCR |
| Estrogen receptor (ER) | Estrogen receptor assay (ERA) | SNOMED |
| Estrogen receptor (ER) | ESR2 (estrogen receptor 2) gene variant measurement | OMOP Genomic |
| Estrogen receptor (ER) | ESR1 (estrogen receptor 1) gene variant measurement | OMOP Genomic |

*Note, the mapping we end up with is highlighted in green.*

Why this mapping was chosen:

Technically, “Estrogen Receptor (ER) Assay” and “Estrogen receptor assay (ERA)” from NAACCR and SNOMED, can be our targets as well, but it is agreed that the OMOP Genomic vocabulary is a priority. It will be fixed on the vocabulary level leaving the only one concept standard. Also, it is a Protein Expression, because the test identifies the presence of the receptors, i.e., proteins; and the estrogen receptor (ER) is the common name of the ER1.

**Results**

This way the OPTUM EHR OMOP CDM gains additional rows with more detailed cancer information.

Table 4. Increase of rows in CDM.

| OMOP Table | rows count | persons count |
|---|---|---|
| Condition | 45 484 874 | 1 832 403 |
| Measurement | 48 859 999 | 1 300 329 |
| Episode | 1 377 880 | 549 473 |
| Observation | 2 666 756 | 390 411 |

*Note, these persons already exist in Core OPTUM EHR set, but they gain additional cancer information now.*

**Conclusion**

Current work is focused on data structure and concepts mapping.

The conversion enriches the Optum EHR data with cancer modifiers, treatment regimen and more detailed diagnoses, which brings more insights to the observational research on cancer.

Since the source data comes from medical notes, it might have inaccuracies. The entries with contradictory information were filtered out: invasive and in-situ at the same day, inconsistent numeric and narrative biomarkers results. More of such metrics to be developed by comparing the Onco-module contents with the core Optum EHR data.

The Treatment response and Tumor progression are currently stored in the Observation table. Having it in OMOP enables further analysis and integration of these values into the Episode table.

**References**


1. WhiteRabbit [Internet]; Available from https://github.com/OHDSI/WhiteRabbit
2. OHDSI Oncology Workgroup [Internet]; Available from https://github.com/OHDSI/OncologyWG/wiki/