IMPROVEMENT of CANCER DIAGNOSIS REPRESENTATION in OMOP CDM

1. BACKGROUND

In a source data, neoplastic disorders can be encoded in multiple coding systems including ICD-9CM, ICD-10, ICDO, and ICD-0-3. SNOMED is uniformly used to represent diagnosis in OMOP CDM. Currently, only ICD-9CM, ICDO, and ICDO-10CM concepts are mapped to SNOMED. Moreover, SNOMED does not always provide granularity sufficient to represent cancer diagnoses and available in ICD-0-3 coding. We propose a solution for representing ICD-0-3 in OMOP without changes of the existing Common Data Model (CDM) or conventions. Such a solution needs to add the granularity of the histology and topography information ICD-0-3 can provide, and to have these represented in a single concept. In addition, these Concepts have to be incorporated into the hierarchical system of SNOMED while preserving its description logic.

2. METHODS

To represent cancer diagnosis in the OMOP CDM Condition domain (Condition Occurrence), we propose to perform a pre-coordination of the ICD-0-3 axes, topography and histology, to a single concept representing unique cancer diagnoses. We propose to map each combination to a corresponding SNOMED concept with respective anatomic site and morphology attributes. If a concept does not exist in SNOMED, we propose to create an OMOP standard concept.

Concept pre-coordination

In ICDO, cancer diagnosis represented by a combination of ICD-0-3 topography and histology.

ICD-0-3 topography example: in C18.7, the C18 indicates that the site is the colon and the 7 indicates that the sub-site is the sigmoid colon.

ICD-0-3 histology example: in 8103, the “81” indicates the adenocarcinoma, “0” indicates malignant bone, “3” indicates malignant bone.

We propose to represent a combination of ICD-0-3 topography and histology as one pre-coordinated concept and map it to a pre-coordinated SNOMED concept.

Demand-driven generation of concepts

ICD-0-3 codes are highly specific and granular.

ICD-0-3 codes exist in SNOMED.

ICD-0-3 codes exist in OMOP v10.82.

SNOMED codes exist in SNOMED.

ICD-0-3 codes exist in OMOP v10.82.

3. RESULTS

We added 6,554 ICD-0-3 combinations taken from the real patient data out of 50,000 possible combinations reported by SEER. 2,698 of these combinations had the SNOMED equivalent and 3,046 did not.

We represented these missing pre-coordinated concepts by new OMOP standard concepts. The coded combinations were based on combinations occurring in one of our first available data set. When tested on another data set, the coverage was 75%.

Example of cancer registry data ETL using ICD-0-3 vocabulary

<table>
<thead>
<tr>
<th>Code</th>
<th>Condition Occurrence</th>
<th>Condition Occurrence</th>
<th>Condition Occurrence</th>
<th>Condition Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2222</td>
<td>Malignant tumor of breast</td>
<td>11/15/2016</td>
<td>&lt;cancer registry&gt;</td>
<td>KID:3-1, C50.9</td>
</tr>
<tr>
<td>2222</td>
<td>&lt;SNOMED: Malignant tumor of breast&gt;</td>
<td>11/15/2016</td>
<td>&lt;cancer registry&gt;</td>
<td>KID:3-1, C50.9</td>
</tr>
<tr>
<td>2255</td>
<td>&lt;SNOMED: Malignant tumor of breast&gt;</td>
<td>12/7/2016</td>
<td>&lt;cancer registry&gt;</td>
<td>KID:3-1, C50.9</td>
</tr>
<tr>
<td>2255</td>
<td>&lt;SNOMED: Malignant tumor of breast&gt;</td>
<td>12/7/2016</td>
<td>&lt;cancer registry&gt;</td>
<td>KID:3-1, C50.9</td>
</tr>
</tbody>
</table>

4. CONCLUSION

We extended OMOP standardized vocabularies to support more granular representation of cancer diagnoses.

To extend the coverage of mappings and improve the quality, we are presently converting several data sets containing ICD-0-3 cancer codes into OMOP CDM.

To perform mappings of the entire SEER reported set, we will improve mappings of ICD-0-3 histology and topography attributes to respective SNOMED attributes. We will formally propose that new pre-coordinated concepts are implemented in SNOMED and improved mappings between ICD-0-3 and SNOMED become part of UMLS.

The remaining challenge is to reconcile in OMOP heterogeneous cancer coding granularity inherited from the source systems. We intend to use SNOMED hierarchy to tackle this challenge.