Full overview of patient cancer episodes in Estonia needs combining multiple data sources in OMOP CDM

Title: Transforming Estonian cancer data to episode table in OMOP

Background: In Estonia, oncology data is scattered across multiple databases. In addition to national claims, prescription and discharge report databases, there are curated registries like the cancer, cancer screening and death registry. Each data source adds a layer of crucial information. Cancer registry details the first-time cancer diagnosis but does not follow up the patient, claims give comprehensive view, but shallow view of the procedures and discharge reports add important detail in unstructured notes. Only by linking all the data sources can we take a comprehensive look at the cancer diagnosis and treatment process. Here we describe extraction of the oncology data for OMOP CDM cancer episodes table from the multiple sources.

Data source
- Estonian Biobank health data (n = 200k)
- Dataset includes:
  - Insurance claims
  - Prescriptions
  - Cancer registry
  - Discharge summaries
  - Death registry
- For this analysis data from 2012 to 2019 was used
- Methodology used converting data to OMOP CDM is described in Oja et. al.

Difficulties encountered
- Source codes are not detailed enough
  - Cases in cancer registry are recorded with time lag.
  - Cancer registry contains information only about the initial diagnosis and treatment, there is no information about the further treatment path.
  - TNM code needed to incorporate ICD-10 codes mapped to SNOMED (Malignant neoplastic disease and descendants) to the algorithm.

Methods

OMOP data
- Insurance claims
- Prescriptions
- Discharge summaries
- Cancer registry
- Death registry

Disease episodes start – first occurrence of cancer diagnosis according to cancer registry using ICD-O-3 diagnosis code. Further development needed to incorporate ICD-10 codes mapped to SNOMED (Malignant neoplastic disease and descendants) to the algorithm.

ICD-O-3* ICD-10 diagnosis * TNM code* stage** Gleason score*

<table>
<thead>
<tr>
<th>Concept logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemotherapy* 60 Chemistry (id=4273629) and all descendants (id=4458073)</td>
</tr>
<tr>
<td>Radiation oncology AND/OR radiotherapy (id=4029715) and all descendants</td>
</tr>
<tr>
<td>Cancer surgery*</td>
</tr>
</tbody>
</table>

Unstructured texts*

Disease Dynamic

All M0 codes defined as “Confined diseases” and M1 codes together with Secondary malignant neoplastic disease as “Metastatic Disease”.

TNM M0 and M1 codes** and ICD-10 codes**

Time lag for collapsing to one episode (days)

Treatment Episodes
- Local service codes*, prescribed drugs*

Disease Extent

Results

Table 1. Preliminary episodes count

<table>
<thead>
<tr>
<th></th>
<th>No of episodes</th>
<th>No of persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease (cancer) Episodes</td>
<td>5789</td>
<td>5471</td>
</tr>
<tr>
<td>Treatment Episodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>6443</td>
<td>3111</td>
</tr>
<tr>
<td>Radiotherapy</td>
<td>3064</td>
<td>2670</td>
</tr>
<tr>
<td>Disease Extent</td>
<td>8386</td>
<td>7977</td>
</tr>
</tbody>
</table>

Figure 1. Intersection of persons between different diagnosis sources (cancer registry, discharge summaries and insurance claims) and extracted episodes (radiotherapy, chemotherapy, disease extent).