Reproducible ETL process to transform Estonian health data to OMOP CDM

INTRO:
- Estonia needs a research database where all health data is standardized and ready to use for observational research.

METHODS:
- 10% random sample of the Estonian population (n=149K patients) from 2012 to 2019
- Dataset included three national data sources:
  - insurance claims (n=6.2M)
  - digital prescriptions (n=9.8M)
  - electronic health records (EHR) (n=97M)
- OMOP CDM v5.3
- Technologies used:
  - Git for version control
  - PostgreSQL for database
  - Python (luigi), SQL, bash for ETL pipelines
  - GATE/NLTK for NLP
  - Translation of local vocabularies to standard vocabularies

RESULTS:
- All three different data sources were combined successfully into one OMOP CDM. With this, we have a full view of patient data over the observation period.
- Process is reproducible and used for different datasets and projects in Estonia:
  - Asthma specific dataset
  - COVID specific dataset
  - Estonian Biobank health data
  - Participation in network studies:
    - Prostate cancer study - PIONEER

Statistics on the ETL procedure to convert Estonian health data to OMOP CDM

<table>
<thead>
<tr>
<th>Source vocabulary</th>
<th>Target vocabulary</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOMESCO Classification of Surgical Procedures (NCSP)</td>
<td>ICD10</td>
<td>226,987</td>
<td>97.1%</td>
</tr>
<tr>
<td>LOINC</td>
<td>SNOMED-CT</td>
<td>6,561,336</td>
<td>92.6%</td>
</tr>
<tr>
<td>RxNorm</td>
<td>SNOMED-CT</td>
<td>14,831,884</td>
<td>94.9%</td>
</tr>
<tr>
<td>RxNorm (post-hoc)</td>
<td>SNOMED-CT</td>
<td>7,568,562</td>
<td>77.6%</td>
</tr>
<tr>
<td>RxNorm (post-hoc)</td>
<td>SNOMED-CT</td>
<td>7,568,562</td>
<td>77.6%</td>
</tr>
</tbody>
</table>

Cleaning of raw data
ETL process
Validation

Contact: marek.oja@ut.ee

Acknowledgements: This work was supported by the Estonian Research Council grants number 8319, RE04-EK-85, the European Union through the European Regional Development Fund (project number 2018E013), and the Estonian Biobank. The whole conversion was carried out in the High Performance Computing Center of the University of Tartu.