Mapping phenotype data of a biobank to OMOP common data model

Sulev Reisberg  
Kristjan Metsalu  
Harry-Anton Talvik  
Jaak Vilo

Data (Estonian Genome Center)  
Mapping to OMOP  
Results

52K Estonians (5% of population)  
Questionnaire  
Original data  
Biobank

63%  
56,773

Persons

Table

<table>
<thead>
<tr>
<th>Country</th>
<th>Domestic product name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonia</td>
<td>Lozap H</td>
</tr>
<tr>
<td>Hungary</td>
<td>TERVALON HCT 50 mg/12,5 mg filmtabletta</td>
</tr>
<tr>
<td>Latvia</td>
<td>Lozap H 50 mg/12,5 mg apvakortš tablettes</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Lozap H 50/12,5 mg plevele dengtos tablettes</td>
</tr>
</tbody>
</table>

Estonian Genome Center, University of Tartu  
Vice Director, Senior Research Fellow, PhD

Acknowledgements  
Peter R. Rijnbeek, Michel Van Speybroeck, Mairo Puusepp, Tõnu Esko

References


Contact  
sulev.reisberg@ut.ee

52K Estonians (5% of population)

SQL scripts

For setting up an empty database  
For building local copy of OMOP vocabulary  
For building necessary mapping tables  
For mapping input data to OMOP CDM

Mapping challenges

a) Too many options  
Which one is the “correct” OMOP concept for measured “body weight”?

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>3013762</td>
<td>Body weight Measured</td>
<td>Measurement</td>
</tr>
<tr>
<td>3013853</td>
<td>Body weight Measured ~ ante partum</td>
<td>Measurement</td>
</tr>
<tr>
<td>3015644</td>
<td>Body weight Measured ~ postoperative</td>
<td>Measurement</td>
</tr>
<tr>
<td>3019336</td>
<td>Body weight Measured ~ pre dialysis</td>
<td>Measurement</td>
</tr>
<tr>
<td>3022281</td>
<td>Body weight Measured ~ pre pregnancy</td>
<td>Measurement</td>
</tr>
<tr>
<td>3023166</td>
<td>Body weight Stated</td>
<td>Measurement</td>
</tr>
<tr>
<td>3025315</td>
<td>Body weight</td>
<td>Measurement</td>
</tr>
<tr>
<td>3026600</td>
<td>Body weight Estimated</td>
<td>Measurement</td>
</tr>
<tr>
<td>3027348</td>
<td>Body weight special circumstances</td>
<td>Measurement</td>
</tr>
<tr>
<td>3027348</td>
<td>Body weight percentile range Categorizations</td>
<td>Measurement</td>
</tr>
<tr>
<td>4029378</td>
<td>Body weight Set</td>
<td>Observation</td>
</tr>
<tr>
<td>4022831</td>
<td>Body weight AND/OR growth problem</td>
<td>Condition</td>
</tr>
<tr>
<td>4099154</td>
<td>Body weight</td>
<td>Observation</td>
</tr>
</tbody>
</table>

b) Country-level non-standard “ATC” codes for drugs containing several components

c) Product names differ in different countries, no central registry

Example of product names of Lozap H in 4 countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Domestic product name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonia</td>
<td>Lozap H</td>
</tr>
<tr>
<td>Hungary</td>
<td>TERVALON HCT 50 mg/12,5 mg filmtabletta</td>
</tr>
<tr>
<td>Latvia</td>
<td>Lozap H 50 mg/12,5 mg apvakortš tablettes</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Lozap H 50/12,5 mg plevele dengtos tablettes</td>
</tr>
</tbody>
</table>

d) Hierarchy codes, code ranges

Diagnosis O04 does not have a standard mapping, because it is an ICD-10 hierarchy code.

When a range is recorded instead of exact diagnosis, e.g. 110-115 (Hypertensive diseases), there is no standard mapping given in OMOP vocabulary.

e) Missing options

There is a concept for “red-blonde hair”, but not for “blonde hair”.

Lessons learned

- Do not leave unmapped records out from OMOP model, mark them with CONCEPT_ID=0 instead
- Keep your mapping scripts under version control (git)
- Make the whole mapping process easily repeatable (need to run it several times)
- In a secure system one has to build mapping scripts on test data, which is challenging (it is also challenging to provide representative test data)

Samuli Metsalu

Achilles

AchillesWeb

Persons

<table>
<thead>
<tr>
<th>Table</th>
<th>Mapped records</th>
<th>Mapped codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>100%</td>
<td>51,890</td>
</tr>
<tr>
<td>Conditions</td>
<td>99%</td>
<td>402,376</td>
</tr>
<tr>
<td>Observations*</td>
<td>100%</td>
<td>2,448</td>
</tr>
<tr>
<td>Measurements**</td>
<td>100%</td>
<td>107,939</td>
</tr>
<tr>
<td>Drugs</td>
<td>86%</td>
<td>56,773</td>
</tr>
</tbody>
</table>

* All non-mapped observations are shown under Conditions
** Only weight and height

“The visualizations are exactly what we need”

Tõnu Esko

Estonian Genome Center, University of Tartu
Vice Director, Senior Research Fellow, PhD