Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned

Sulev Reisberg, Ph.D.
Estonia & University of Tartu

- Small Eastern European country
- Member of European Union and NATO
- Population 1.4 M
- “Estonia is the world's most digitally advanced society” (WIRED)
- University of Tartu - the oldest and largest university in Estonia
Estonian OMOP people Health Informatics research group at the institute of computer science (University of Tartu) approx 15 p 2-3 people at Tartu University Hospital 2-3 people at STACC (SME) 2-3 people at Quretec (SME) 2-3 people at Estonian Biobank 4-5 others...
Our recent paper

http://doi.org/10.1093/jamiaopen/ooad100

Research and Applications

Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned

Marek Oja, PhD¹, Siri Tamm, MSc¹, Kerli Mooses, PhD¹, Maarja Pajusalu, MSc¹, Harry-Anton Talvik, MSc¹,², Anne Ott, MSc¹, Marianna Laht, MD¹, Maria Malk, MSc¹, Marcus Loo, MSc¹, Johannes Holm, MSc¹, Markus Haug, MSc¹, Hendrik Šuvalov, MSc¹, Dage Särg, MSc¹,², Jaak Vilo, PhD¹,², Sven Laur, PhD¹, Raivo Kolde, PhD¹, Sulev Reisberg, PhD¹,²

¹Institute of Computer Science, University of Tartu, 51009 Tartu, Estonia, ²STACC, 51009 Tartu, Estonia
*Corresponding author: Marek Oja, PhD, Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia (marek.oja@ut.ee)

Author Contributions: Dr M. Oja and S. Tamm are considered co-first authors and Dr R Kolde and Dr S Reisberg are considered co-last authors of this work. In addition, they had full access to all the data in the study and take responsibility for the integrity of data and accuracy of the data analysis.

Abstract

Objective: To describe the reusable transformation process of electronic health records (EHR), claims, and prescriptions data into Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM), together with challenges faced and solutions implemented.

Materials and Methods: We used Estonian national health databases that store almost all residents' claims, prescriptions, and EHR records. To develop and demonstrate the transformation process of Estonian health data to OMOP CDM, we used a 10% random sample of the Estonian population (n = 150,824 patients) from 2012 to 2019 (MAIT dataset). For the sample, complete information from all 3 databases was converted to OMOP CDM version 5.3. The validation was performed using open-source tools.

Results: In total, we transformed over 100 million entries to standard concepts using standard OMOP vocabularies with an average mapping rate 95%. For conditions, observations, drugs, and measurements, the mapping rate was over 90%. In most cases, SNOMED Clinical Terms were used as the target vocabulary.

Discussion: During the transformation process, we encountered several challenges, which are described in detail with concrete examples and solutions.

Conclusion: For a representative 10% random sample, we successfully transferred complete records from 3 national health databases to OMOP CDM and created a reusable transformation process. Our work helps future researchers to transform linked databases into OMOP CDM more efficiently, ultimately leading to better real-world evidence.

Lay Summary

Health data can be found in various sources and formats, making it challenging for researchers. To address this issue, one possible approach is...
Data sources & context of the paper

**Public and private healthcare providers in Estonia**
- Inpatient care providers (hospitals)
- Outpatient care providers
- Family doctors (GP)
- Pharmacies

**National operational health databases**
- CSV: Claims
- CSV: Digital prescriptions
- XML: EHR documents

**Context of this paper**
- Random 10% sample for 2012 - 2019 de-identified but linked by patient pseudonym
- Transforming data to OMOP CDM

**Combined dataset in OMOP CDM**
Overview of ETL

**Extract**
- ICD-10 Diagnoses
- NCSP Surgical procedures
- Local codes for procedures, measurements, drugs, visits
- ATC Drugs
- Local drug product codes
- ICD-10 Diagnoses
- ICD-10 Diagnoses
- SNOMED CT
- LOINC Lab tests
- TNM and cancer stages
- Free text in clinical notes

**Transform**
- Use mappings from Athena
  - Yes
    - Mapping available in Athena?
      - No
        - Use Usagi for manual mapping
          - Use NLP
    - No
      - Create mapping table
        - Clinical validation

**Load**
- Load data into OMOP CDM using SQL and Bash scripts
  - OMOP CDM
    - SNOED CT
    - RxNorm
    - LOINC
    - CANCER MODIFIER
    - UCUM

**Validate**
- Data Quality Dashboard
- CdmInspection package
- ATLAS (ACHILLES)
- Clinical validation
Reviewer 2

"Starting with just 10% makes sense, but it'd be great to have stats for the other 90% too."
• Extensive table of challenges + examples + solutions
• Dataset used in a number of health studies
• A separate paper will be submitted that focuses on data extraction from source documents

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Table 4. Main challenges and solutions of the current work</th>
</tr>
</thead>
<tbody>
<tr>
<td>The same health event is represented in several source datasets without a clear link between them, potentially leading to duplicates.</td>
<td><strong>Example</strong> The same diagnosis code for a patient may be recorded in an EHR, claims, and prescription files. However, it may be difficult to link these documents to a single event due to the absence of a unique identifier for the case.</td>
</tr>
<tr>
<td>No clear guidelines for choosing target vocabulary when multiple standard OMOP vocabularies are available. Additionally, there are no roadmaps indicating which standard vocabularies may no longer be considered standard for OMOP CDM in the near future.</td>
<td><strong>Solution</strong> The same observation code for a patient may be recorded in an EHR, claims, and prescription files. However, it may be difficult to link these documents to a single event due to the absence of a unique identifier for the case.</td>
</tr>
<tr>
<td>Hard to keep manual mapping files up to date as the standard target concepts change over time when updating the vocabularies.</td>
<td>The same observation code for a patient may be recorded in an EHR, claims, and prescription files. However, it may be difficult to link these documents to a single event due to the absence of a unique identifier for the case.</td>
</tr>
</tbody>
</table>

- OMOP: Open Medical Data Program
- CT: Concept Type

- Transforms each record as they are (even if duplicates) but adds the provenance information to the record so one can use it when making cohorts.
- Use the target vocabulary you are more familiar with. Keep in mind that what constitutes a standard OMOP vocabulary may change over time.
- Whichever updating the vocabulary, recheck the mappings in Usual before running the transformation. Usual automatically creates the list of nonstandard mappings so one can fix them before the actual data transformation.
- When working with historical codes, always check the most recent target to ensure the event is the same code.
THANK YOU!

https://health-informatics.cs.ut.ee