The European OHDSI Initiative: Why are we here?

Peter Rijnbeek
Department of Medical Informatics
Erasmus MC Rotterdam
Performing multi-database studies in Europe

Our Goals:
• We want to generate real-world evidence in a fully reproducible way
• We want to shorten the time from study conception till publication without sacrificing quality

Our Challenges:
• We have to deal with a large amount of different database structures
• We have many different coding systems in Europe (ICPC, READ, ICD9, ICD10, etc.)
• Our health care systems differ considerably across Europe

We need to find solutions to improve the inter-operability of data

We need a framework for rapid access to and fully reproducible analysis of data
“What's the adherence to my drug in the data assets I own?”

Analytical method: Adherence to Drug

Application to data

Current solution: One SAS or R script for each study

- Not scalable
- Not transparent
- Expensive
- Slow
- Prohibitive to non-expert routine use
We have made progress by developing tools to standardize part of the process.
But we can do better...

- Each data custodian has to create custom data extraction scripts
- This step is often not transparent to the study lead
- It is time consuming to perform the semantic harmonization step
Solution: Data Standardization Enables Systematic Research

- Adherence
- Mortality
- Source of Business
- Safety Signals

OHDSI Tools

OMOP CDM

North America
Southeast Asia
China
Europe
UK
Japan
India
So Africa
Switzerland
Italy
Israel
Standardized data
Network Studies
Networks of networks

Coordinating Center

ISDN
University Medical Center
Inpatient Hospital
Outpatient Hospital

Another Network

Network
The European Medical Information Framework (EMIF) Project

14 European countries combining 57 partners
€56 million worth of resources
3 projects in one
5 year project (2013–2018)
**Goals of EMIF**

**EMIF-Platform Objective**
Develop a framework for evaluating, enhancing and providing access to human health data across Europe, support EMIF-Metabolic and EMIF-AD as well as support research using human health data in general.

### Tool Development
- **Key tool developed—EMIF Catalogue** as data “shop window” to support the platform architecture, also being utilised by other initiatives (DP-UK, EPAD)
- **Workflow, data extraction, harmonisation and aggregation software** developed and being used to support several studies as use cases for research process

### Common Data Model
- Piloting the [OMOP common data model](http://www.commondatamodel.org) in the platform databases to assess its potential towards data harmonization
Why the OMOP-CDM

Our requirements:

• Active open source community: mapping tools, analytical tools etc.
• Solution for semantic harmonization by using standardized vocabularies
• Need to store ALL source data including source vocabularies
• Possibility to refine to database-specific queries
• Multiple technical infrastructures supported
EMIF Adopted the OMOP-CDM and has worked on mapping 10 European databases

Table 1: The 10 European databases that are part of the EMIF initiative and that are mapped to the OMOP CDM.

<table>
<thead>
<tr>
<th>Database</th>
<th>Country / Region</th>
<th>Population Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agenzia regionale di sanità della Toscana (ARS)</td>
<td>Italy / Tuscany</td>
<td>$5 \times 10^6$</td>
<td>Administrative database of Tuscan population</td>
</tr>
<tr>
<td>Aarhus University Hospital Database</td>
<td>Denmark / Northern Region</td>
<td>$2.3 \times 10^6$</td>
<td>Administrative database of Central and North Jutland</td>
</tr>
<tr>
<td>Health Search IMS Health LPD</td>
<td>Italy</td>
<td>$1.6 \times 10^6$</td>
<td>Primary care data of GP’s using the Health Search System</td>
</tr>
<tr>
<td>Integrated Primary Care Information (IPCI)</td>
<td>Netherlands</td>
<td>$2.8 \times 10^6$</td>
<td>Primary care database</td>
</tr>
<tr>
<td>Pedianet</td>
<td>Italy</td>
<td>$0.4 \times 10^6$</td>
<td>Pediatric database</td>
</tr>
<tr>
<td>Pharnmo</td>
<td>Netherlands</td>
<td>$8.4 \times 10^6$</td>
<td>Primary care database</td>
</tr>
<tr>
<td>Information System of Parc de Salut Mar (IMASIS)</td>
<td>Spain</td>
<td>$1.4 \times 10^6$</td>
<td>Hospital database</td>
</tr>
<tr>
<td>The Information System for the Development of Research in Primary Care (SIDIA)</td>
<td>Spain / Cataluna</td>
<td>$6.4 \times 10^6$</td>
<td>Primary care database</td>
</tr>
<tr>
<td>The Health Informatics Network (THIN)</td>
<td>United Kingdom</td>
<td>$12 \times 10^6$</td>
<td>Primary care database</td>
</tr>
<tr>
<td>Estonian Genome Center at the University of Tartu</td>
<td>Estonia</td>
<td>$52 \times 10^3$</td>
<td>Biobank</td>
</tr>
</tbody>
</table>
ETL requires multi-disciplinary team

- Medical knowledge
- CDM Knowledge
- Project Coordination
- ETL Development
- Database / Infrastructure management
- Local
- EMIF

Depending on preferences & available skills, EMIF can take on different roles
OHDSI tools supporting the process

- Analyze Data Source
- Define Mapping Logic
- Develop Mappings (ETL)
- Test ETL
- Deploy ETL
- Acceptance Testing / QA

Evaluate Infrastructure

White rabbit
Rabbit in a hat
github
usagi
Achilles
Assessment of the conversion of ten European Databases to the OMOP CDM and evaluation of the use of OHDSI tools

Michel van Sprycket,1 Lars Halvorsen,2 Myriam Alexander,3,5,6 Glen James,3,5,6 Lara Tramontan,3,5,6 Leonardo Mendez-Boo, MD MPH,3,6 Rients van Wingarden, MSc,3,6 Rosa Giri, PhD,3 Miguel A. Mayer PhD,3 Lars Pedersen, PhD,2 Alessandro Pasqua MSc,1 Susil Reinsberg MSc,1 Johan van der Leij, PhD,5 Peter R. Rijnbeek, PhD,5

1JanssenPharmaceutica NV, Beerse, Belgium; 2Erasmus MC, Rotterdam, The Netherlands; 3Arsenal, IT, TV, Italy; 4Södertörn University; 5Institut Catala de la Salud, Spain; 6Institut Universitar d’Investigacions en Atenció Primària Jordi Gol (IDAP Jordi Gol), Barcelona, Spain; 7STI&ODC, Utrecht, The Netherlands; 8Agencia regional de Sanidad de la Toscana, Florence, Italy; 9Research Programme on Biomedical Informatics (IMIM-UPF), Barcelona, Spain; 10Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark; 11Genomics Srl, Firenze, Italy; 12University of Tartu, Estonia; 13GSK, Uxbridge, United Kingdom

Background

The European Medical Information Framework (EMIF) aims to develop a sustainable platform for the re-use of real world data sources, covering a wide variety of sources: regional healthcare systems, hospital data, primary care data and biobanks. The harmonization of data sources towards the OMOP CDM and the use of OHDSI tools are an important constituent of the EMIF platform. The population data sources that are part of the EMIF initiative are shown in Table 1.

Table 1: The 10 European databases that are part of the EMIF initiative and that are mapped to the OMOP CDM.

<table>
<thead>
<tr>
<th>Database</th>
<th>Country / Region</th>
<th>Population Size</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aarhus University Hospital Database</td>
<td>Denmark / Northern Region</td>
<td>2.3 x10^6</td>
<td>Administrative database of Central and Northern Denmark</td>
<td>Completed</td>
</tr>
<tr>
<td>Health Search IMS Health LPD</td>
<td>Italy</td>
<td>1.6 x10^6</td>
<td>Primary care data of GPs using the Health Search System</td>
<td>Completed</td>
</tr>
<tr>
<td>Integrated Primary Care Information (IPCI)</td>
<td>Netherlands</td>
<td>2.5 x10^6</td>
<td>Primary care database</td>
<td>Completed</td>
</tr>
<tr>
<td>Peritell</td>
<td>Italy</td>
<td>0.6 x10^6</td>
<td>Pediatric database</td>
<td>In Progress</td>
</tr>
<tr>
<td>Pharmaco</td>
<td>Netherlands</td>
<td>8.4 x10^6</td>
<td>Primary care database</td>
<td>Completed for cohort study</td>
</tr>
<tr>
<td>Information System of Parc de Salut Mar (IMAS3)</td>
<td>Spain</td>
<td>1.4 x10^6</td>
<td>Hospital database</td>
<td>In Progress</td>
</tr>
<tr>
<td>The Information System for the Development of Research in Primary Care (S/DAP)</td>
<td>Spain / Catalunya</td>
<td>6.4 x10^6</td>
<td>Primary care database</td>
<td>In Progress</td>
</tr>
<tr>
<td>Estonian Genome Center at the University of Tartu</td>
<td>Estonia</td>
<td>52.8 x10^6</td>
<td>Biobank</td>
<td>Completed</td>
</tr>
</tbody>
</table>

Methods

Mapping to the OMOP CDM

The mapping to the OMOP CDM was based on the best practices as developed by the OHDSI community. Different technologies for the ETL (Java-jCDMBuilder / SQL / Kettle / Python) were used – depending on the party who developed the ETL and / or the technology that was acceptable for the data source.

Assessment of the mapping

Following the mapping of the databases, there is a need to understand the overall ‘quality’ of the mappings and to assess the readiness of the mapped databases to support research questions. The process that is followed is illustrated below.

Evaluation of Achilles

The standalone version of Achilles (version 1.3) was reviewed by 25 users, covering researchers as well as database owners through a structured assessment.

<table>
<thead>
<tr>
<th>Achilles</th>
<th>CDM Queries</th>
<th>Integrity checks</th>
<th>Study Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Achilles heel</td>
<td>- Counts per entity</td>
<td>- Compare source vs target concepts</td>
<td>- Perform study on original source vs mapped data source</td>
</tr>
</tbody>
</table>

Conclusions

The Achilles tool was well appreciated by our user group and suggestions to improve this tool have been made and implemented. Our work on the conversion of European databases to the OMOP CDM showed that it is feasible but requires detailed quality assessments. Extensions of the Standardized Vocabularies are needed to capture all the European data adequately. This work is ongoing. The conversion of the databases will be further assessed and improvements will be proposed in the upcoming period.
Current Challenges: ETL

The following factors were found to be most impactful on overall speed and quality of the ETL:

1. **Source Database research readiness**: The ‘quality’ of the input data structure – and the availability of internal knowledge on how the database is defined - are the primary driver of efficiency and quality of the CDM Mapping.

2. **Strong project management**: superior results in terms of quality and speed can be achieved when resources are allocated and active project management is executed.

3. **Vocabulary mappings**: establishing the vocabulary mappings is the most resource intensive step. It’s recommended to set realistic goals with associated timings (e.g. map the top 20% of lab tests, covering 80% of all occurrences).
Evaluation of translation: Structural Mapping

Did all my source data end up in the CDM?

Lost in translation?

- Can be very good reason for differences: business rules assessment
- Iterative process to optimize the ETL
- No structural CDM limitations encountered so far
Evaluation of translation: Vocabulary Mapping

IPCI Database Example

- High data coverage.
- Term coverage is further improved by extending the Standard Vocabularies, e.g. RxNorm-Extension to accommodate European Drug market.
Rationale for the European OHDSI Initiative

1) The need for data standardization through a common data model is recognized by many stakeholders, including regulators

2) Many projects are being initiated in Europe in which the OMOP-CDM is adopted

3) There is a large interest across Europe to learn more about the OMOP-CDM and OHDSI tools
We are in a perfect storm!!

But...
- We should not all invent the same wheel but collaboratively perform open science
- An active European Community will be able to further improve CDM structure, and will improve vocabulary content to match our European needs
What do we need?

• Implement quality framework for full analytical pipeline to build trust

• Establish ETL conventions (business rules) to further improve the interoperability

• Create a European eco-system to support mapping, including certified SMEs

• Further extend the data network
Goal of the Themis workgroup is to define business rules for ETL to further improve the inter-operability of the data.

For example:
- Patients with multiple values for sex, gender, race
- Potentially contradictory relationship between days supply, quantity, drug exposure end date and sig in Drug Exposure
- Duplicate procedures or visits at the same day
- Medical events after death date.
- Observation Period definitions for Claims and EHR records
- etc
Training a crucial prerequisite for success!

- There is a lot of miss-understanding about the OMOP-CDM
  - A protocol-based versus CDM study
  - Loss of source data
  - Etc.

- We need to train all stakeholders including our medical students
Aims OHDSI-EUROPE

• Stimulate Community Building in Europe
• Stimulate Adoption of the CDM
• Create a platform for sharing ideas and results
• Close collaboration with all stakeholders
• Close collaboration with other chapters
• Training of all stakeholders
Join the Journey

Our ambitions are high but achievable if all disciplines collaborate and share ideas!

We need all of you to keep the OHDSI train on track and running at speed in Europe!