Name:	Michel Van Speybroeck
Affiliation:	Janssen Pharmaceutica N.V.
Email:	mvspeybr@its.jnj.com
Presentation type (s):	**Poster**

Assessment of the conversion of ten European Databases to the OMOP CDM mapping and evaluation of the use of OHDSI tools in that process.

Michel van Speybroeck[1], Johan van der Lei, PhD [2], Lars Halvorsen[1] ,Myriam Alexander, PhD [13], Glen James, PhD [13] , Lara Tramontan, PhD [3,4], Leonardo Méndez-Boo, MD MPH [5,6], Rients van Wijngaarden, MSc [7], Rosa Gini, PhD [8], Miguel A. Mayer PhD [9], Lars Pedersen, PhD [10] , Alessandro Pasqua Msc [11], Sulev Reisberg MSc[12], Johan van der Lei, PhD [2], Peter R. Rijnbeek, PhD[2]

[1]Janssen Pharmaceutica NV, Beerse, Belgium ; [2]Erasmus MC, Rotterdam, The Netherlands; [3]Arsenàl.IT, TV, Italy; [4]SoSePe, PD, Italy; [5]Direcció de Sistemes d'Informació, Institut Català de la Salut, Spain;[6] Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona, Spain; [7]STIZON, Utrecht, The Netherlands; [8]Agenzia regionale di sanità della Toscana, Florence, Italy; [9] Research Programme on Biomedical Informatics (IMIM-UPF), Barcelona, Spain; [10] Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus N, Denmark; [11]Genomedics Srl, Florence, Italy; [12]University of Tartu, Estonia; [13] GSK, Uxbridge, United Kingdom

Abstract

Ten databases were mapped to the OMOP Common Data Model (CDM) in the context of the European Medical Information Framework (EMIF) project. OHDSI tools including White Rabbit / Rabbit in a Hat, jCDM builder, Usagi and Atlas / Achilles were used. The diversity in data structures and use of different vocabularies requires that a rigorous quality assessment process is put in place. This process and the current status of the quality assessment is discussed. Achilles is an important component for evaluation of the database conversion and is also a valuable tool for database characterization. The EMIF stakeholders have evaluated Achilles and results are presented.

Introduction

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.

Database	Country / Region	Population Size	Type	Mapping Status
Agenzia regionale di sanita della Toscana (ARS	Italy / Tuscany	5. 10^6	Administrative	Completed
Aarhus University Hospital Database	Denmark	2.3 10^6	Administrative	Completed
Health Search IMS Health LPD	Italy	1.6 10^6	Primary care	Completed
Integrated Primary Care Information (IPCI)	Netherlands	2.8 10^6	Primary care	Completed
Pedianet	Italy	0.4 10^6	Pediatric data	In Progress

Database	Country / Region	Population Size	Type	Mapping Status
Pharmo	Netherlands	8.4 10^6	Primary care	Completed for cohort
Information System of Parc de Salut Mar (IMASIS)	Spain	1.4 10^6	Hospital data	In Progress
The Information System for the Development of Research in Primary Care (SIDIAP)	Spain / Catalonia	6.4 10^6	Primary care	In Progress
The Health Informatics Network (THIN)	United Kingdom	12 10^6	Primary care	Completed
Estonian Genome Center at the University of Tartu (EGCUT)	Estonia	52 10^3	Biobank	Completed

Figure 1: Overview of the 10 databases mapped to the OMOP CDM

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. Critical success factors for a productive and efficient mapping process include:

1. Database research readiness: the source databases have been developed and curated to cater for different types of research use to different extents. 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. Project management should span the entire cycle from specification through ETL development to evaluation and production deployment. Breaking the project up and / or assigning resources upon availability can lead to extended timelines and the need to rebuild knowledge at the different steps
3. Vocabulary mappings: establishing the vocabulary mappings (assuming that not all mappings are available) is the most resource intensive step. Except for the simplest cases, it's recommended to set realistic goals with associated timings (e.g. map the top 20% of lab tests, covering 80% of all occurrences. From the corresponding absolute count, an accurate effort estimate can be made)

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

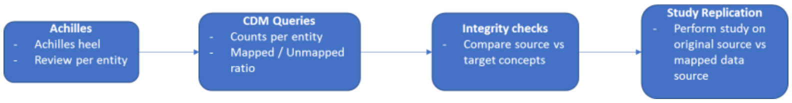
Achilles
- Achilles heel
- Review per entity
CDM Queries
- Counts per entity
- Mapped / Unmapped ratio
Integrity checks
- Compare source vs target concepts
Study Replication
- Perform study on original source vs mapped data source

Figure 2: Proposed flow for assessment of the mapping

Firstly, the Achilles evaluation includes a review of the Achilles heel output and a qualitative check on the different entities where counts for the top concepts are checked against the distribution in the source data. In a second step – the CDM queries – a set of queries is launched against the different databases to provide the following information per entity: the number of mapped/unmapped concepts and the associated number of occurrences, the counts on drug mappings and associated levels and the top 100 codes of unmapped drugs. Figure 3 illustrates this for the drug mappings

Database	Ingredient	Clinical Drug Comp	Clinical Drug Form	Clinical Drug	Unmapped
AUH	5%	11%	12%	72%	
ARS	81%				19%
Health Search IMS	100%				
IPCI	35%	4%	1%	56%	4%
Pedianet	100%				

Figure 3: Drug level mapping. % based on record count (extended version covering all datasources will be presented on the poster)

In some cases, however, source data might not be mapped to the CDM as the corresponding CDM entity might be unknown. As an example, a source system might have a table containing all 'observations' but the syntactic mapping might point to measurements, observations or condition_occurrence. The third step is intended to measure this potential gap. This is an activity that is still to be performed
Finally, the replication of (part of) a study- is still to be performed. The conditions and restrictions and representativeness

Evaluation of Achilles

The standalone version of Achilles (version 1.3) was reviewed by 26 users, covering researchers as well as database owners. The evaluation was performed against the THIN database. User experience was generally very positive with 66% qualifying it as good or excellent and 31% as OK and 4% as poor.

Additional features that users would recommend included the export capability for the tables and graphs , the development of a print functionality , The addition of a database summary description on the landing page which would contain information about the terminology systems used (especially on the tree maps with hierarchical information) , the possibility to see the frequency distribution per person of a particular entity and the ability to search using local vocabularies.

Some users were concerned about possible miss-interpretation of data and were suggesting the capability to add annotations to graphs plus a way to access information on how the data has been generated. EMIF is actively engaged in the further development to implement some of these features. The full report is available through http://forums.ohdsi.org/t/emif-evaluation-of-achilles/1964

Conclusion

It's recognized that harmonization towards the OMOP CDM across different database types and with a wide heterogeneity in vocabulary systems is a significant enabler for performing large scale research. Applying a solid process in the mapping of the data sources and subsequent quality assessment, sharing best practices and optimization of the tools and ensuring the proper organizational context were found to be key for achieving success.