Case study: Challenges faced by EMIF in utilising the OMOP CDM

Johan van der Lei
Erasmus Medical Center
Rotterdam

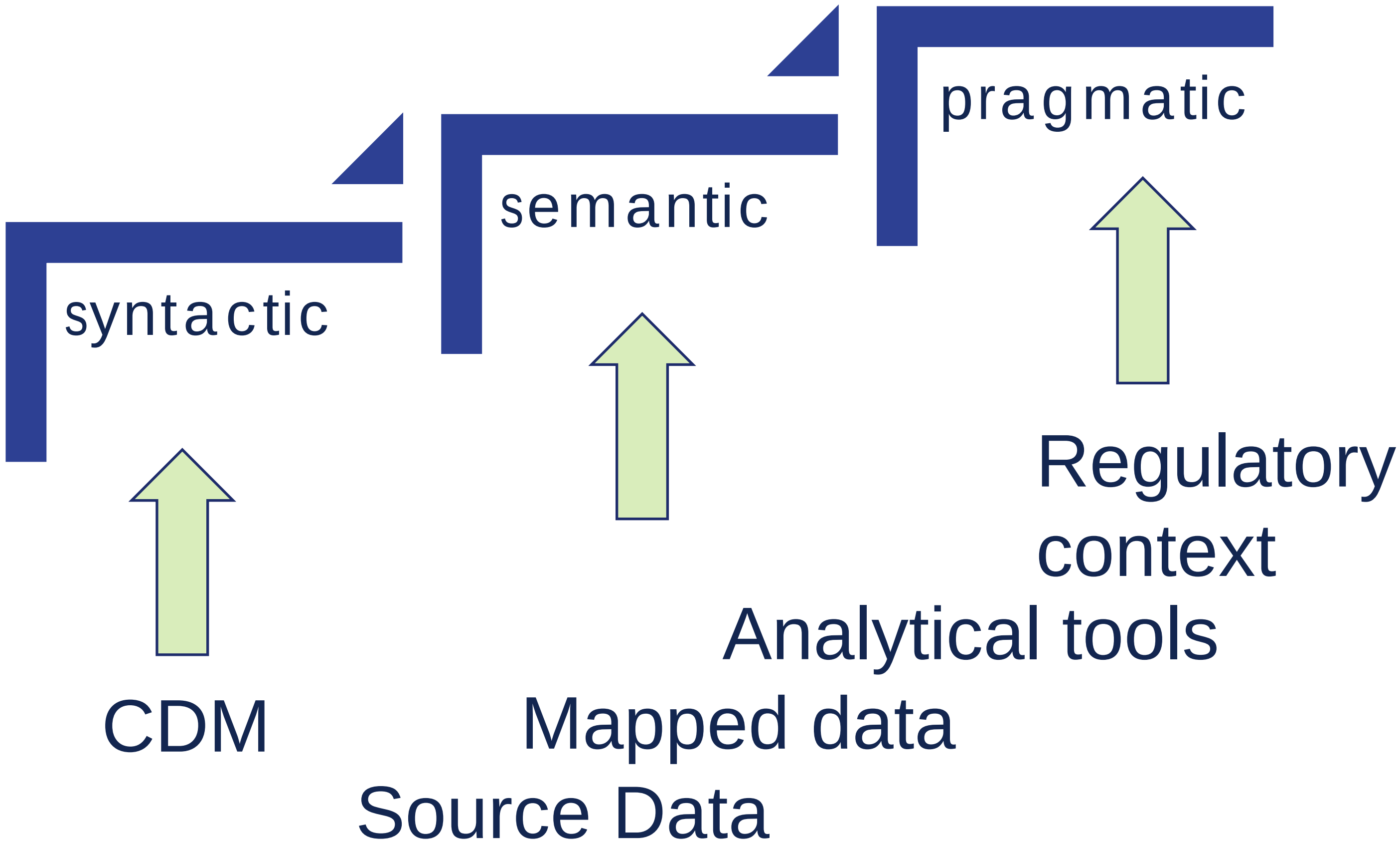
Outline

- Scaffolding
- EMIF and a CDM
- EMIF and the OMOP CDM
- Ongoing activities/challenges

Information

- Syntactic: “grammar”
- Semantic: “meaning”
- Pragmatic: “consequences”

Views on information



Views on information

syntactic

semantic

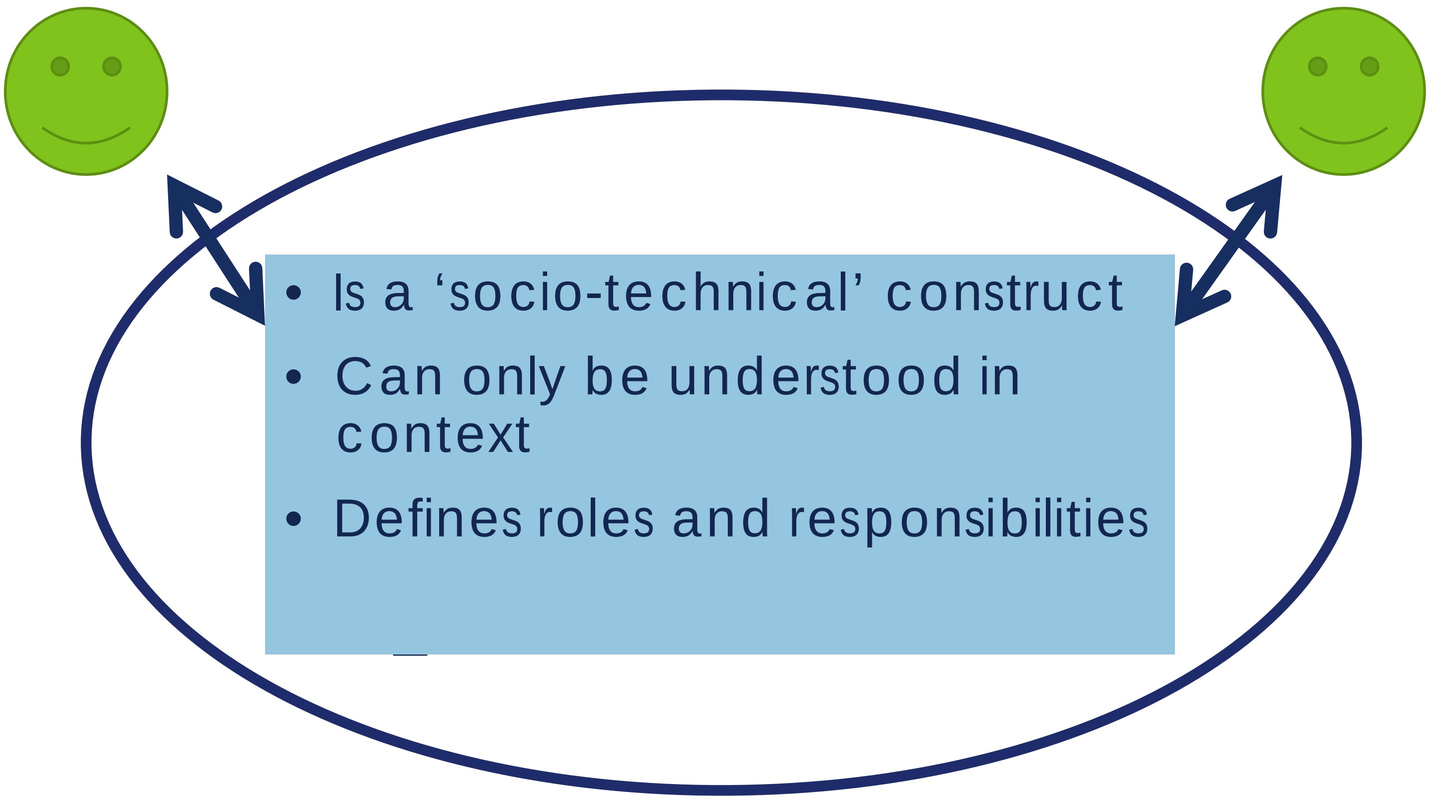
pragmatic

CDM…

- On the syntactic level
- Multiple solutions possible
  - Models are dynamic
- Debate often: semantic and pragmatic
- But that discussion is often independent of a specific model

From medical informatics perspective:

- The question which ***CDM*** to use is probably not the right question……



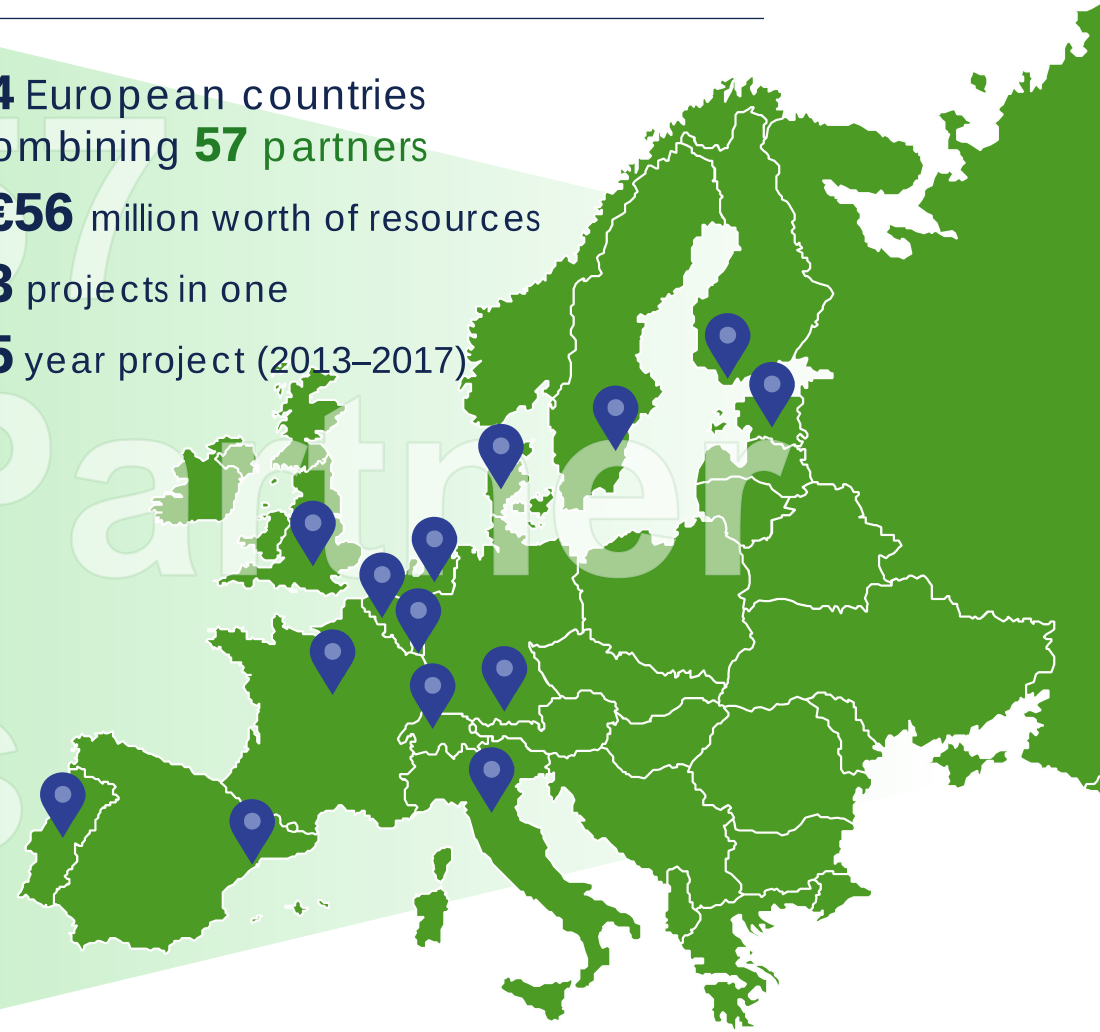
Project overview

ACADEMIC PARTNERS (37)

SME PARTNERS (9)

EFPIA PARTNERS (10)

PATIENT ORGANISATION (1)



14 European countries combining **57** partners

€56 million worth of resources

3 projects in one

5 year project (2013–2017)

Why is EMIF needed?

Potential applications of Real World Data

Discovery

- Biomarker discovery
- Predictive modelling
- Disease insight generation (opportunity identification)

Development

- Trial design and feasibility analysis
- Electronic health record (EHR)-facilitated recruitment
- Prospective cohort selection

Launch/ Post-Launch

- Analysis of treatment pathways
- Collection of clinical and economic evidence
- Ongoing efficiency and safety monitoring

EMIF Setting

- Data from very diverse sources
  - Population based
  - Hospital based
  - Disease specific cohorts
  - Biobanks
- Diverse data
- Broad spectrum of research questions
- Overall purpose: facilitate re-use of data

EMIF and CDM Challenges

- Clear need for a CDM
- Broad spectrum of coding schemes, languages, and settings
- Need to store ALL source data including source vocabularies
- Possibility to escape/refine to study-specific solutions
- Reproducible research: Open, Transparent, Source data, Mappings, Analytical tools
- Flexibility in role transfer (e.g., study coordinator)
- Multiple technical infrastructures

EMIF and OMOP CDM: why?

- No silver bullet…
- but not yet another model !!!
- Diversity of the EU setting: Support for Standardized Vocabularies
- Not limited to specific analytical use case
- Open source
- Multiple platforms
- OHDSI
  - Open collaborative
  - Growing in EU

EMIF Databases being mapped to the OMOP-CDM

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

Database	Country / Region	Population Size	Type	Status
Agenzia regionale di sanita della Toscana (ARS	Italy / Tuscany	$5\ 10^6$	Administrative database of Tuscan population	Completed
Aarhus University Hospital Database	Denmark / Northern Region	$2.3\ 10^6$	Administrative database of Central and North Jutland	Completed
Health Search IMS Health LPD	Italy	$1.6\ 10^6$	Primary care data of GP's using the Health Search System	Completed
Integrated Primary Care Information (IPCI)	Netherlands	$2.8\ 10^6$	Primary care database	Completed
Pedianet	Italy	$0.4\ 10^6$	Pediatric database	In Progress
Pharmo	Netherlands	$8.4\ 10^6$	Primary care database	Completed for cohort
Information System of Parc de Salut Mar (IMASIS)	Spain	$1.4\ 10^6$	Hospital database	In Progress
The Information System for the Development of Research in Primary Care (SIDIAP)	Spain / Cataluna	$6.4\ 10^6$	Primary care database	In Progress
The Health Informatics Network (THIN)	United Kingdom	$12\ 10^6$	Primary care database	Completed
Estonian Genome Center at the University of Tartu	Estonia	$52\ 10^3$	Biobank	Completed

ETL requires multi-disciplinary team

Medical knowledge

OHDSI

CDM Knowledge

(Local) Data knowledge

Depending on preferences & available skills, EMIF can take on different roles

Project Coordination

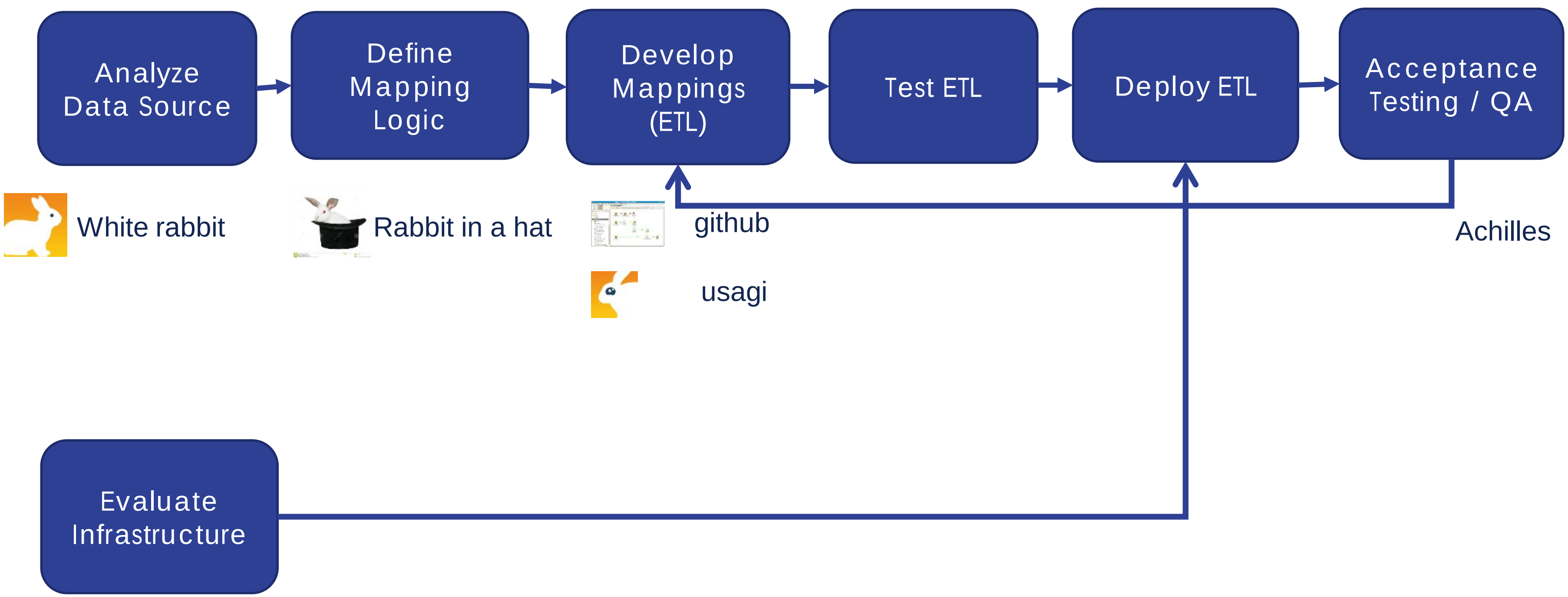
ETL Development

Database / Infrastructure management

Local

EMIF

Tools supporting the process



Assessment of the conversion of ten European Databases to the OMOP CDM and evaluation of the use of OHDSI tools

Michel van Speybroeck[1], 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 [1], 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; 11Genomedics Srl, Florence, Italy; [12]University of Tartu, Estonia;[13] GSK, 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.

Database	Country / Region	Population Size	Type	Status
Agenzia regionale di sanita della Toscana (ARS	Italy / Tuscany	$5\ 10^{6}$	Administrative database of Tuscan population	Completed
Aarhus University Hospital Database	Denmark / Northern Region	$2.3\ 10^{6}$	Administrative database of Central and North Jutland	Completed
Health Search IMS Health LPD	Italy	$1.6\ 10^{6}$	Primary care data of GP's using the Health Search System	Completed
Integrated Primary Care Information (IPCI)	Netherlands	$2.8\ 10^{6}$	Primary care database	Completed
Pedianet	Italy	$0.4\ 10^{6}$	Pediatric database	In Progress
Pharmo	Netherlands	$8.4\ 10^{6}$	Primary care database	Completed for cohort
Information System of Parc de Salut Mar (IMASIS)	Spain	$1.4\ 10^{6}$	Hospital database	In Progress
The Information System for the Development of Research in Primary Care (SIDIAP)	Spain / Cataluna	$6.4\ 10^{6}$	Primary care database	In Progress
The Health Informatics Network (THIN)	United Kingdom	$12\ 10^{6}$	Primary care database	Completed
Estonian Genome Center at the University of Tartu	Estonia	$52\ 10^{3}$	Biobank	Completed

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.

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 1: Proposed Mapping Assessment Flow for the 10 European Data Sources

Evaluation of Achilles

The standalone version of Achilles (version 1.3) was reviewed by 26 users, covering researchers as well as database owners through a structured assessment.

Results

Mapping to the OMOP CDM

Based on the experience in working with 10 data sources, the following factors were found to be most impactful on overall speed and quality of the mapping:

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)

Assessment of the mapping

Table 2, shows an overview of the drug level mappings. All data sources have a link of their drug coding system to ATC. Where available, a more granular mapping to clinical drug or form was performed. Unmapped category indicates records where no standard drug code could be found. Aside from missing concepts, this can also be attributed to the fact that the source tables can have a broader scope e.g. different OTC products.

Table 2: Drug level mapping. % based on record count

Data Source	Ingredient	Clinical Drug Comp	Clinical Drug Form	Quant Clinical Drug	Clinical Drug	Unmapped
ARS	80.5%	0.0%	0.0%	0.0%	0.0%	19.5%
AUH	5.5%	10.6%	12.0%	0.0%	72.0%	0.0%
IPCI	34.9%	3.5%	0.9%	0.0%	56.3%	4.4%
GENOMEDICS	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PHARMO	14.0%	7.6%	3.6%	0.0%	74.7%	0.0%
PEDIANET	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
THIN	7.8%	0.0%	0.0%	1.5%	69.7%	20.9%
EGCUT	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Evaluation of Achilles

User experience was generally very positive with 66% qualifying it as good or excellent and 31% as OK and 4% as poor. Additional features of interest included the possibility to see the frequency distribution per person of a particular entity and the ability to search using local vocabularies. This has now been implemented in Atlas. The full report is available at http://forums.ohdsi.org/t/emif-evaluation-of-achilles/1964

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

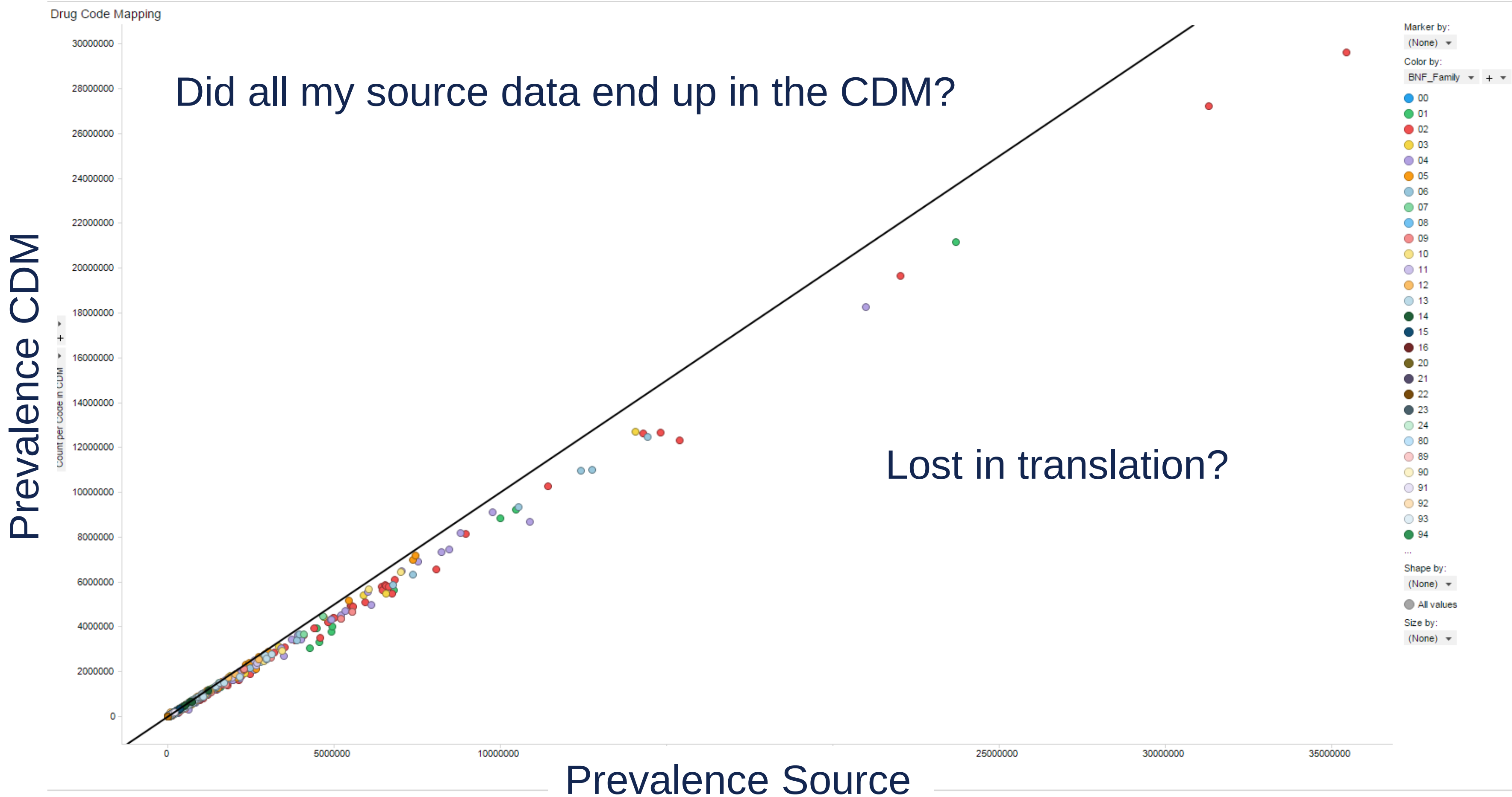
The following factors were found to be most impactful on overall speed and quality of the ETL:

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).

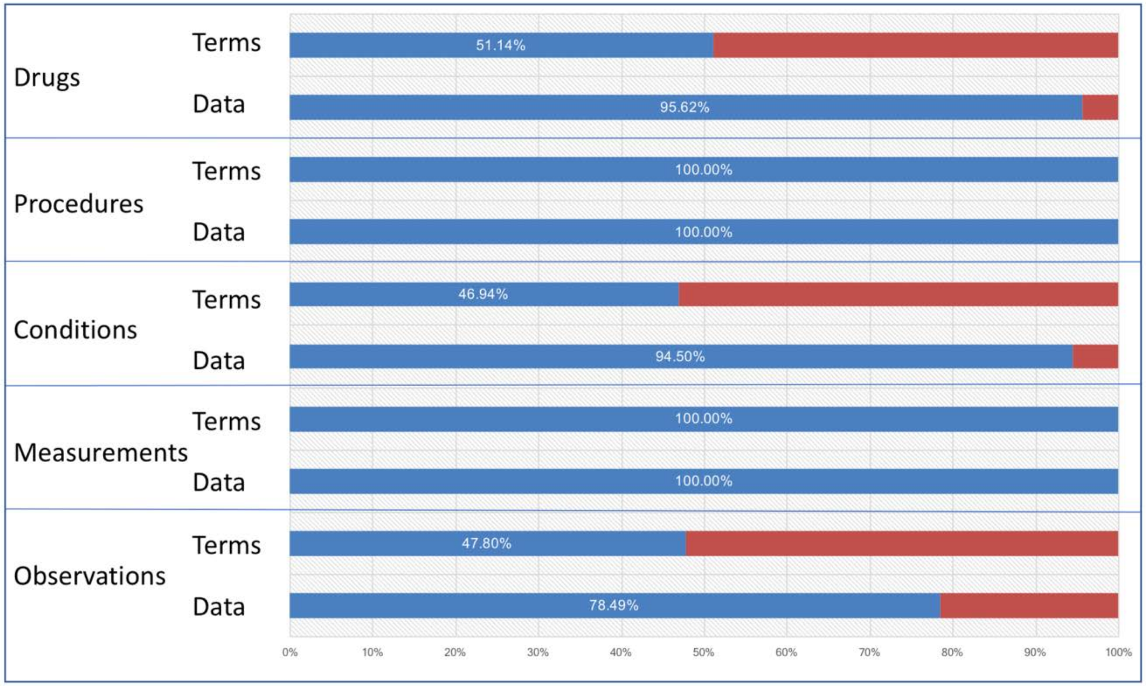
Evaluation of translation: Structural Mapping



- 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

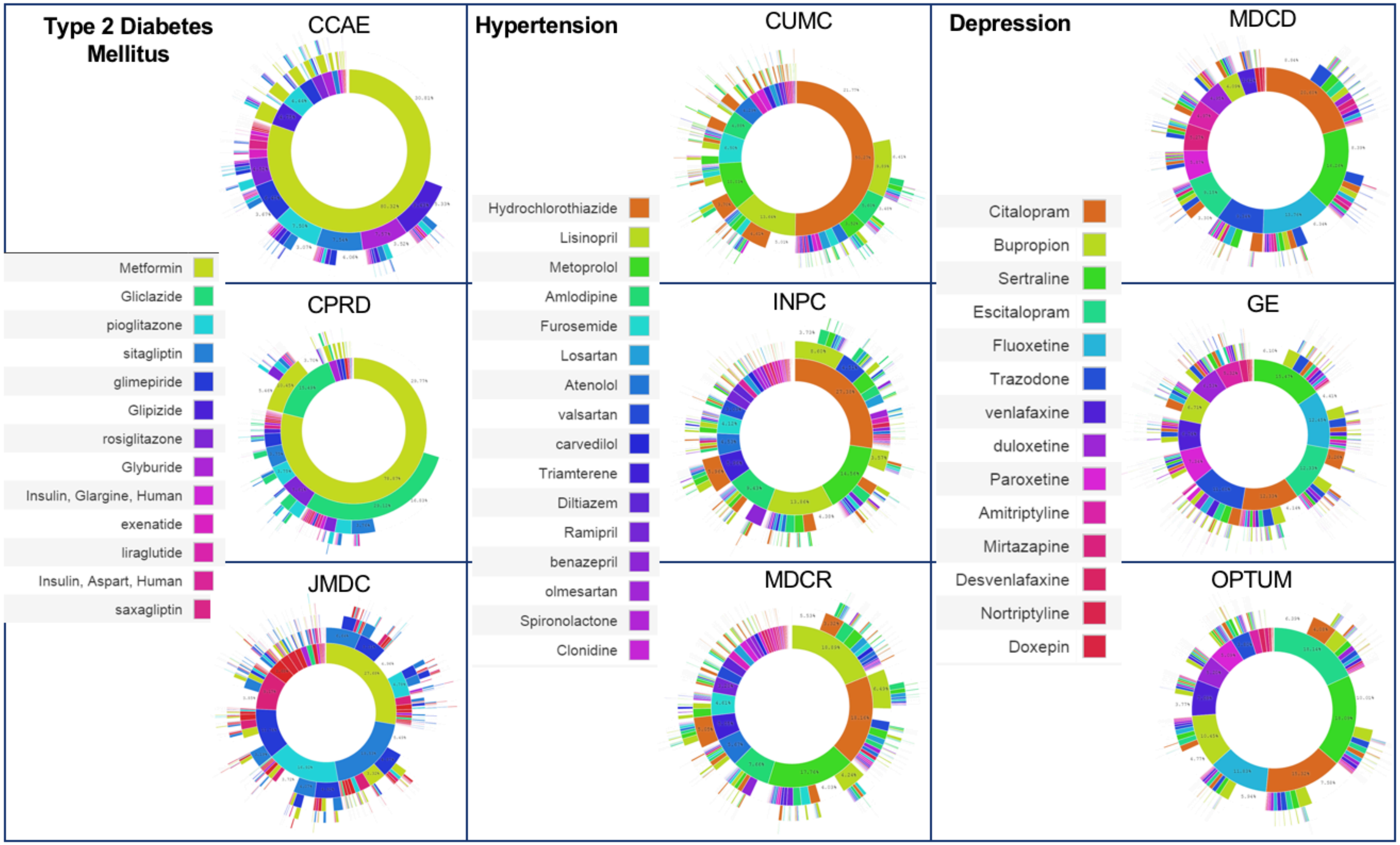


- High data coverage.
- Term coverage is further improved by extending the Standard Vocabularies, e.g. RxNorm-Extension to accommodate European Drug market

Current EMIF CDM Activities

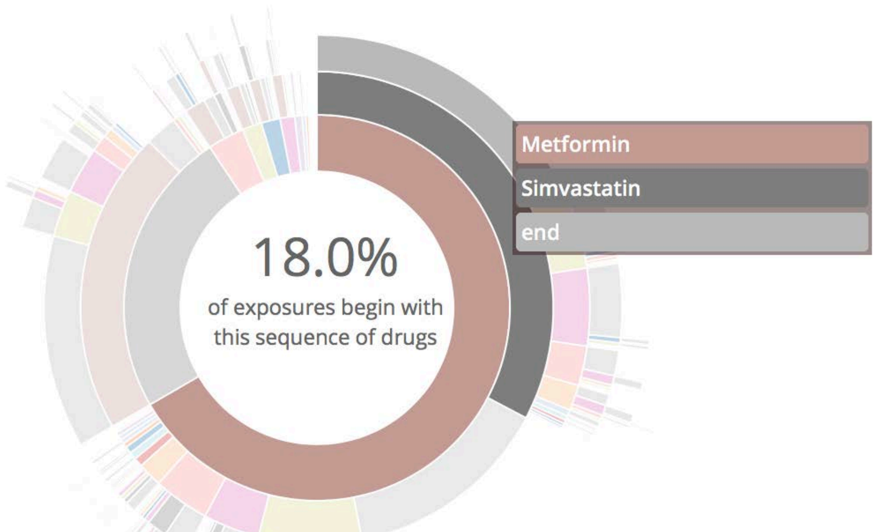
- Replication of existing EMIF Use Case(s) on CDM
- Contribute to vocabulary extension
- Contribute to tool development
- Training of stakeholders in using the CDM and OHDSI Tools
- Initiate and participate in OHDSI Network studies

Example:
Treatment Pathway Study



Hripcsak G. et al.
Characterizing treatment pathways at scale using the OHDSI network. PNAS 2016 113 (27) 7329-7336;

IPCI: Type 2 Diabetes



EMIF will run this study on CDM databases in Europe

Final Remarks

- EMIF will intensify its participation in the OHDSI network by supporting the European OHDSI initiative (www.ohdsi-europe.org) coordinated by Erasmus MC

- EMIF is in the process of a sustainability assessment to support and use the data network post EMIF

We believe that the adoption of the OMOP-CDM and the active OHDSI community will enable transparent and reproducible research at an unprecedented scale in Europe