Name:	Peter Rijnbeek
Affiliation:	Erasmus MC, Rotterdam, The Netherlands
Email:	p.rijnbeek@erasmusmc.nl
Presentation type (s):	Poster

Conversion of a Dutch general practitioner's database to the OMOP-CDM: The IPCI Database

Peter R. Rijnbeek, PhD[1,3], Erica A. Voss, MPH[1,2,3], Marcel de Wilde[1,3], Martijn Schuemie, PhD[2,3], Johan van der Lei, PhD[1,3]

[1] Erasmus University Medical Center, Rotterdam, The Netherlands, [2]Janssen Research and Development, Raritan, NJ, [3] Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA

Abstract

The Integrated Primary Care Information (IPCI) database is a longitudinal observational database containing electronics medical records from a representative sample (n=750) of general practitioners (GPs) in the Netherlands. The database contains up to 10 years of observational data from 2.36 Million persons. The IPCI data is collected from 9 different GP systems that are normalized into a common data structure. Subsequently, many quality control steps are performed before the data is considered research ready. To participate in the OHDSI data network and utilize the analytical tools an extraction, transformation, and loading (ETL) process was developed to translate the IPCI research ready database to the OMOP Common Data Model. In this abstract, we describe this ETL process and assess the completeness of the mappings to the standard vocabularies. Results showed that the IPCI database structure could be adequately mapped to the OMOP-CDM V5. Further improvements can be made with respect to the coverage of local codes in standard vocabularies, especially in the drug domain.

Introduction

The Integrated Primary Care Information (IPCI*)* project is a longitudinal collection of electronic patients records from Dutch general practitioners (GPs) [1]. The IPCI project (www.ipci.nl) was started in 1989 and has been proven suitable for the conduct of descriptive and analytic research, e.g., incidence and prevalence of disease, drug utilization by indication, safety and effectiveness of prescription drugs. Currently, the IPCI data is collected from 750 GPs from 9 different vendors and contains data from 2.36 Million patients.

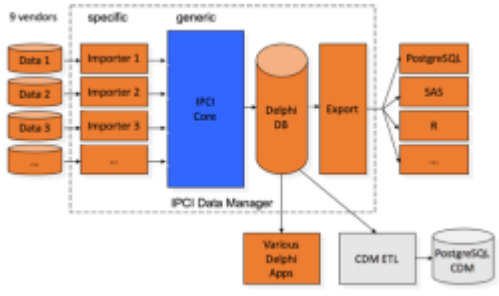

Figure 1. IPCI Data processing pipeline

In Figure 1, the data processing pipeline is shown in more detail. For each individual vendor, a specific import module is created that translates its format to the generic IPCI database format (IPCI Core). Subsequently, data is anonymized and further processed in a custom built object-oriented Delphi database that allows for very fast data retrieval. In the IPCI group several valuable visualization tools are built to perform a quality control assessment of the data. A research ready observation period is defined per patient before the data is exported and loaded in a PostgreSQL database, SAS etc.

The IPCI data contains a rich set of data elements: demographics, conditions (ICPC1), prescriptions (Z-index), procedures (NHG), visits, laboratory (NHG) results and a large corpus of clinical notes. Our objective was to develop an ETL process for the IPCI database to the OMOP-CDM V5 to participate in the OHDSI data network and leverage all great OHDSI tools.

Methods

To start the process, the IPCI database was characterized with the open source Java tool called WhiteRabbit. Subsequently, Rabbit-In-A-Hat served as a primary documentation tool for a two-day session involving the data custodian, data manager, CDM experts. This initial ETL description was used as the basis for an extraction and transformation program written in Delphi against the IPCI-Core database. The most challenging part was the mapping of the local source codes to the standard concepts in the OMOP Vocabulary. The Usagi tool was used by a panel of coders. Drug mapping is implemented using a combination of Java and T-SQL. The input is the local Z-Index tables and counts of Z-Index codes in the IPCI database. The output is a set of tables describing for each code in the Z-Index which concept in the OMOP Vocabulary it should be mapped to. The Java program uses the "Generieke Product Code" GPK that described the drugs at the generic level, including form and strength. After translation to English and several text processing steps, the drugs were mapped to the most specific standard code available in the current drug vocabulary.

Results

The first CDM version of the IPCI database has been realized and all Achilles Heel errors have been resolved. The current observation period conventions did not allow us to store all history data in the CDM. We have to assess the impact of this on our observational studies. Table 1, shows the proportion of mapped terms and database records for the data domains. Note that the 100% coverage for procedures and measurements is too optimistic because unmapped concepts in these domains were not placed in these tables. Table 2, shows the percentage of codes and percentage of data that could be mapped to each drug level.

Table 1. Proportion of mapped term and data

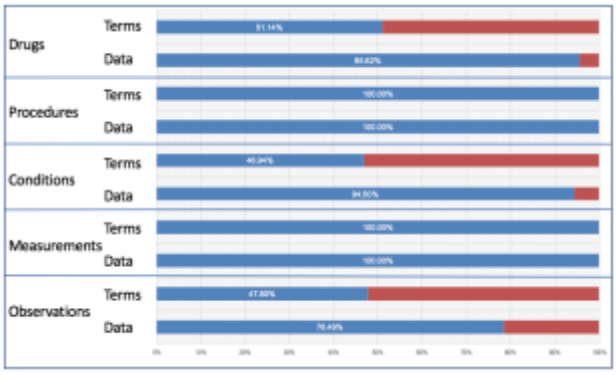

Table 2. Drug level mapping

Category	Code count	(%)	Row count	(%)
Clinical Drug	68065	28.72	42829250	56.26
Clinical Drug Comp	8664	3.66	2683190	3.52
Clinical Drug Form	1948	0.82	707551	0.93
Ingredient	42530	17.94	26577281	34.91
Unmapped	115814	48.86	3334999	4.38

These results show that a high data coverage could be achieved on most domains, however term coverage can be improved considerably for less frequent concepts. Drug mappings should be improved further by extending the vocabularies to capture all drugs on the European market at the clinical drug level.

Conclusion

The IPCI database structure could be adequately mapped to the OMOP-CDM V5. Further improvements are being made to the standard vocabularies to better cover the local coding systems used in the Netherlands.

References

1. Vlug AE, Van der Lei J, Mosseveld BMTh, Van Wijk MAM, Van der Linden PD, Sturkenboom MCJM, Van Bemmel JH Postmarketing Surveillance Based on Electronic Patient Records: The IPCI Project. Methods Inf Med. 1999 Dec;38(4-5):339-44