Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220510

Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients

Mathilde FRUCHART[a,1] Paul QUINDROIT[a], Haris PATEL[a], Jean-Baptiste BEUSCART[a], Matthieu CALAFIORE[a,b] and Antoine LAMER[a]
[a] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*
[b] *Univ. Lille, Département de médecine générale, F-59000 Lille, France*

Abstract. The implementation of clinical data warehouses has advanced in recent years. The standardization of clinical data in these warehouses has made it possible to carry out multicenter studies and to formalize the clinical vocabulary. However, there is limited insight into a patient's overall care pathway in the clinical domain. Regarding primary care data, the implementation of this type of warehouse in a routine way is hindered in particular by the analysis of textual data provided by general practitioners during patient consultations. In our study we collected primary care data for standardization in a data warehouse. The purpose of this analysis was to assess the feasibility of analyzing primary care data, and particularly to study the consultations and prescriptions of the elderly patient contained in our primary care data warehouse.

Keywords. Primary care, Data warehouse, Elderly patient, Data Reuse, Electronic Health Record

1. Introduction

The digitalization of health facilities has enabled the automated collection of electronic health records primarily generated for care or administrative purposes [1]. This offers the possibility of a secondary use for research [2], particularly through the implementation of clinical data warehouses [3].

Data warehouses are mainly developed on the scope of hospitals or national claims. On one hand, in the hospital, they provide a complete view of patient management in the care units with the diagnoses, the procedures, the drugs administrations, the biology results and the discharge letters. However, they do not allow us to explore the patient outcomes after discharge and in other hospitals. On the other hand, national claims contain data collected anonymously and prospectively for all national health insurance beneficiaries [4]. They are dedicated to reimbursed inpatient and outpatient care (e.g., general practitioners [GPs], pharmacies, nurses, etc.) and do not include clinical data. New attempts have emerged for primary care [5,6]. They provide data the same patient

[1]Corresponding Author, Mathilde Fruchart, ULR 2694, Lille University, 2 place de Verdun, F-59000, Lille, France; E-mail: mathilde.fruchart@univ-lille.fr.

over the long term, such as weight, body mass index, results of blood tests and symptoms experienced by the patient during each medical visit [7].

In our study, we have implemented a data warehouse with data collected in a GPs center and we studied clinical and longitudinal data from the general population and the elderly.

2. Methods

We implemented a data warehouse with the GPs center of Wattrelos, in north of France. The center includes 6 general practitioners. The software Weda (WEDA, Montepellier, France) collects the data during the consultations, with the patient history and symptoms, the vital signs and biometric data (e.g., heart rate, arterial pressure, temperature, weight, height, body mass index), the drug prescriptions, the blood tests, and the billing data. Drug prescriptions are documented with the *Code Identifiant de Présentation* (CIP), a French terminology. CIP codes were mapped with Anatomical Therapeutic Chemical (ATC) codes. Patient history and symptoms are described with the International Classification of Diseases in the 10th version (ICD10) and free text. The blood tests results are documented after receiving the analysis reports from the laboratories, and the codes depend of the laboratory terminology. The vital signs and biometric data are specified with non-standardized questions. The mapping to the ATC, blood tests labels and questions has been conducted beforehand and verified by a team of pharmacists, geriatricians and GPs belonging to the research team. We removed all sensitive data (i.e., identity, birth date, patient's address) and we created a unique artificial identifier per patient. The data warehouse is stored on a secure server disconnected from the internet.

An extract-transform-load (ETL) process was implemented to automatically read XML files and integrated all the information listed in a relational database. The process will be relaunched each time we get a new extraction from the editor. In order to enhance reproducibility and to collaborate with other centers and other software, we implemented the Observational medical outcomes partnership Common data model (OMOP-CDM) [8]. The use of this CDM allows to reproduce and share data, methods and results across different centers. The ETL was implemented in Python and Postgresql. The use of OHDSI tools was applied in particular for vocabulary mapping (Atlas, Athena) [8].

We extracted all records of consultations that occurred between 2013 to 2020 with at least one drug prescription. The extraction results in a set of XML files for each patient with a hierarchy of information. An exploratory analysis was done to judge the quality of the data and to visualize the variables that could be extracted. We identified if data was presented in a structured or unstructured form, if there was missing data and if standardized vocabularies were utilized. Some records did not always include contact with the patient, and were related to a physician action in the record, or the receipt of a document (e.g., biology results). Therefore, we filtered on records with at least one mention of a drug prescription.

We described two populations to explore the database: the general population and the elderly aged 75 and more.

3. Results

3.1. Data integration

We implemented the tables PERSON, VISIT_OCCURRENCE, DRUG_EXPOSURE, MEASUREMENT, NOTE, OBSERVATION, LOCATION, PROVIDER in the OMOP-CDM. The following data quality problems were identified: missing birth date, structured data but without terminology, free text, non-standardized questions, missing zip codes and non-standardized city labels.

3.2. Descriptive analysis

Between 2013 and 2020, 16,396 patients were admitted for at least one consultation with a drug prescription. A total of 181,527 consultations were analyzed. The overall population had a mean (standard-deviation) age of 47.6 (24.2) and 54% were male, at consultation. The figure 1 displays the number of patients and consultation per age (at the consultation). There were 472 patients of 20 years old, who have benefited from 1,258 consultations, for a ratio of 2.66 consultations per patient, while there were 428 patients of 70 years old, who have benefited from 1.901 consultations, for a ratio of 4.44 consultations per patient. In average, 3.39 (2.88) drugs were prescribed per consultation for the general population. The regular measurement of biometric data during the consultations makes it possible to follow their evolution over time, as shown in figure 2 with the evolution of weight at different ages at the consultation.

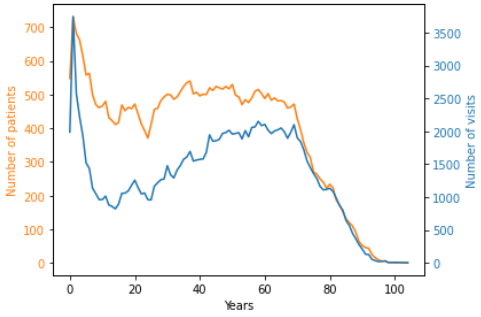

Figure 1. Number of patients and consultations per age at the consultation.

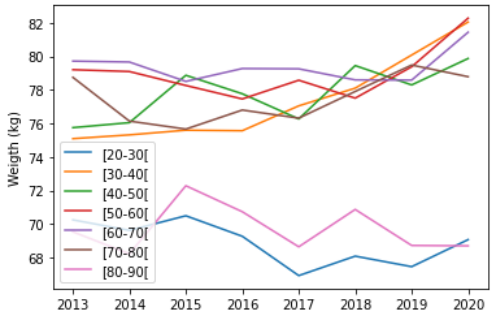

Figure 2. Change in weight from 2013 to 2020 for patients aged 20 (20 to 29), 30 (30 to 39), 40 (40 to 49), 50 (50 to 59), 60 (60 to 69), 70 (70 to 79), 80 (80 to 89) years old at consultation

3.3. Elderly

The elderly population (over 75 years of age) represents 5,5% of the general population or 900 patients for 13,867 consultations (7.6% of the consultations of the general population). This population received an average of 4.5 consultations per year. Elderly population benefited from 5.97 (3.72) drug prescriptions on average per consultation. The drugs prescribed corresponded to 871 unique ATC codes. The figure 3 represents the top 30 most prescribed drugs. The most prescribed drugs treat pain (paracetamol) for 10% of the prescriptions, followed by NSAIDs (acetylsalicylic acid) for 2.5% and esomeprazole for 2.3% of the prescriptions.

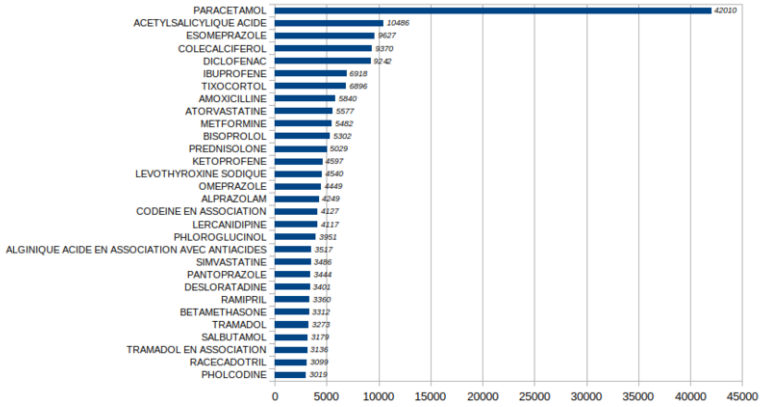

Figure 3. The top 30 most prescribed drugs for elderly patients.

4. Discussion

We integrated primary care data from a practice with 6 practitioners, over 8 years, for a total of 16,396 patients and 181,527 consultations with at least one prescription of drug. This allowed to highlight that the elderly population had more visits than the general population and received more medications. Surprisingly, the most prescribed drugs for elderly patients treat the pain rather than the symptomatology. With data collected in primary care, we can track clinical parameters over several years. The data comes directly from the source software and is more accurate than that collected secondarily for billing. Moreover, it contains blood tests results that are not available in national claim databases. This original nature of this type of data source offers the possibility of comparing the evolution of biology with treatment over the long term.

Despite the use of two standard terminologies (ICD10 and CIP), some data remain complicated to homogenize. In particular, the labels of the blood tests depend on the laboratory, and examinations and questionings are documented in free text during the consultation. Data quality problems in primary care EHR have already been identified in the literature [6,9] and we could apply the method proposed by Lacroix-Hugues et al. To map unstructured data (i.e., symptoms and diagnoses) to ICPC2 classification (International Classification of Reference for Primary Care) [6].

This work will be replicated in other centers equipped with the same software. By using the OMOP-CDM, we can also integrate data from other software editors for further collaborations in a multisite network.

5. Conclusions

We implemented a data warehouse with data collected from a GPs center. Despite many manuals' entries, we were able to produce longitudinal statistics. This work will be completed by integrating data from other GPs centers.

6. Funding

This research was funded by PreciDIAB, which is jointly supported by the French National Agency for Research (ANR- 18- IBHU- 0001), by the European Union (FEDER - Agreement NP0025517), by the Hauts-de-France Regional Council (Agreement 20001891/NP0025517) and by the European Metropolis of Lille (MEL, Agreement 2019_ESR_11).

References

[1] Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, Applebaum S. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood). 2012 Dec;31(12):2805-16. doi: 10.1377/hlthaff.2012.0884. Epub 2012 Nov 15. PMID: 23154997.

[2] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017 Aug;26(1):38-52. doi: 10.15265/IY-2017-007. Epub 2017 Sep 11. PMID: 28480475; PMCID: PMC6239225.

[3] Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform. 2017 Jun;102:21-28. doi: 10.1016/j.ijmedinf.2017.02.006. Epub 2017 Feb 16. PMID: 28495345.

[4] Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc JL, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015 Jun;36(6):411-7. doi: 10.1016/j.revmed.2014.11.009. Epub 2014 Dec 26. PMID: 25547954.

[5] Gentil ML, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, Renault E, Bouzille G, Chapron A. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak. 2017 Sep 25;17(1):139. doi: 10.1186/s12911-017-0538-x. PMID: 28946908; PMCID: PMC5613384.

[6] Lacroix-Hugues V, Darmon D, Pradier C, Staccini P. Creation of the First French Database in Primary Care Using the ICPC2: Feasibility Study. Stud Health Technol Inform. 2017;245:462-466. PMID: 29295137.

[7] de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006 Apr;23(2):253-63. doi: 10.1093/fampra/cmi106. Epub 2005 Dec 20. PMID: 16368704.

[8] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8. PMID: 26262116; PMCID: PMC4815923.

[9] Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, Chevendra V, Maddocks H, Shadd J, Burge F, Thind A. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak. 2019 Feb 12;19(1):30. doi: 10.1186/s12911-019-0740-0. PMID: 30755205; PMCID: PMC6373085.