Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200132

Cohort Creation and Visualization Using Graph Model in the PREDIMED Health Data Warehouse

Christophe CANCÉ[a], Pierre-Ephrem MADIOT[b], Christian LENNE[a], Svetlana ARTEMOVA[bc], Brigitte COHARD[b], Marjolaine BODIN[be], Alban CAPOROSSI[abcd], Jean-François BLATIER[b], Jerôme FAUCONNIER[b], Frédérique OLIVE[b], Daniel PAGONIS[b], Dominique LE MAGNY[b], Jean-Luc BOSSON[abcd], Katia CHARRIERE[bc], Ivan PATUREL[b], Bruno LAVAIRE[b], Gabriel SCHUMMER[b], Joseph ETERNO[b], Jean-Noël RAVEY[b], Ivan BRICAULT[abcd], Gilbert FERRETTI[abf], Sébastien CHANOINE[abdf], Pierrick BEDOUCH[abd], Emmanuel BARBIER[e], Julien THEVENON[abf], Pascal MOSSUZ[abf], Alexandre MOREAU-GAUDRY[abcd]

[a] *Univ. Grenoble Alpes (UGA), Grenoble, France*
[b] *CHU Grenoble Alpes (CHUGA), Grenoble, France,*
[c] *Clinical Investigation Centre – Technological Innovation, Inserm, CHUGA, UGA, Grenoble, France*
[d] *TIMC-IMAG Laboratory, UGA, CNRS, Vet Agro, Grenoble, France*
[e] *Grenoble Institut des Neurosciences, Grenoble, France*
[f] *IAB, UGA, Inserm, CNRS, France*

Abstract. Grenoble Alpes University Hospital (CHUGA) is currently deploying a health data warehouse called PREDIMED [1], a platform designed to integrate and analyze for research, education and institutional management the data of patients treated at CHUGA. PREDIMED contains healthcare data, administrative data and, potentially, data from external databases. PREDIMED is hosted by the CHUGA Information Systems Department and benefits from its strict security rules. CHUGA's institutional project PREDIMED aims to collaborate with similar projects in France and worldwide. In this paper, we present how the data model defined to implement PREDIMED at CHUGA is useful for medical experts to interactively build a cohort of patients and to visualize this cohort.

Keywords. Complex and massive data, medical informatics, health data warehouse, cohorts, data visualization

1. Introduction

Selecting a cohort of patients and gathering associated medical data for research usually involves bringing together information from different sources, potentially isolated from each other in a hospital's information system.

It is thus necessary, for all patients that are likely to be members of a research cohort, to combine their clinical notes with the local information of a clinical department of the hospital, then to enrich them with those distributed in the various hospital databases, to gather all the documents containing demographic, clinical, biological, pharmacological information concerning these patients. This phase often

involves the cooperation with the IT department to transform the potentially complex needs expressed by the research team into querying language to extract the expected information from the hospital's multiple databases. Several iterations of this process between researchers and data managers may be necessary to ultimately lead to the expected data set. However, in this way no interactivity on data exploration is offered to researchers, though it would allow them to obtain or approach faster the best possible result. It is then necessary to organize these data in a coherent and easily explorable form for researchers to select the members of the final cohort that will constitute the data set to be used for the study.

The whole process of collecting the data set associated with a cohort of patients is therefore expensive in terms of mobilizing different health and IT specialists for several months. This paper describes methods and tools implemented as a proof of concept in the framework of the health data warehouse called "PREDIMED" of the Grenoble Alps University Hospital, to enable hospital authorized medical specialists to navigate interactively and in a fluid way through the whole data lake or through their own data set, to create and explore their own cohorts of patients.

PREDIMED is a data lake platform designed to integrate and analyze for research, education and institutional management, the data of patients treated at CHUGA since 1998. PREDIMED contains administrative data (patients and visits), healthcare data (diagnoses, prescription, images, etc.) and, potentially, data from external databases. PREDIMED is hosted by the CHUGA Information Systems Department and benefits from its strict security rules. PREDIMED has several views on the data facilitating its exploration and usage according to the specific needs. In this paper, we specifically present the graph model and the associated tools dedicated to its real-time exploration and data visualization of millions of heterogeneous information.

2. Methods and implementation

Here we describe methods and implementation choices that we used to provide healthcare professionals with an environment allowing them to simply and graphically manipulate the concepts and relationships of their expertise domain to navigate through the hospital's data lake without knowledge of any computer language.

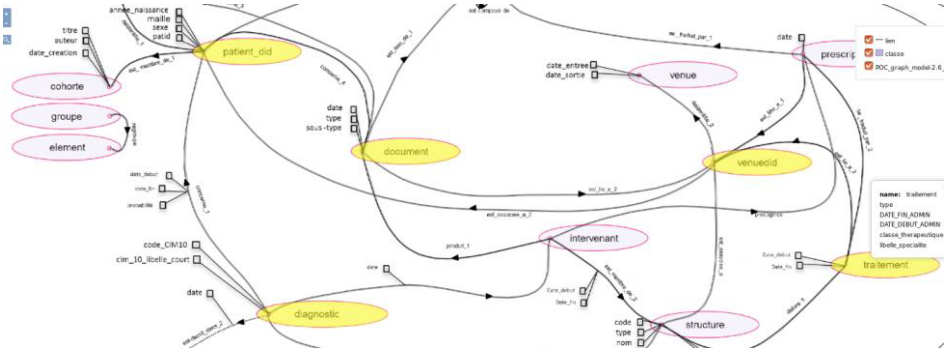

***Figure** 1: Graph data model of the data lake and interactive cohort construction interface.*

The data lake is modeled in an intuitive form that allows each member of the research project, regardless of their profession, to understand its semantics [3]. Figure 1 gives an idea of the graphical visualization of the lake data model. Graph nodes are common medical concepts and graph edges represent the relationships between them.

Patients data from various hospital databases is linked in PREDIMED via an internal key. This technical key is different from the patient number used for care to separate medical information from directly identifying data. The same approach is used for the visit numbers.

Using the graph data model, the team in charge of cohort constitution works with the clinician to build the cohort for the study. The graph model allows starting from any entity since the selection is made by choosing a concept from the model and following an edge that links it to another concept. The traversal can thus start from diagnostic codes (ICD10), periods of hospitalization, administered treatments or any other graph node. The user can then progress in traversing the graph edges by hops, possibly expressing constraints at different levels. Interactively, at each movement through the graph, the system provides the volume of collected data. In Figure 1, one can see that at this intermediate stage, the selected data concerns the entities marked in yellow (here *diagnostics*, *patient_did*, *document*, *venue_did* and *treatment*). The attributes associated with *venue_did* can be filtered (a filter can be set on the hospitalization period *(dt_debut_prestation* for example)).

Behind this graphical selection mechanism is a query language [4] specifically developed for the project that allows the user to query the associated graph database. The query basic hop is built as follows:

"source objects class"[filters]—named relationship[filters]—>"target objects class", ...

The complete query string that traces the graphical navigation path is associated with the current cohort, which allows the user to relaunch it if the lake content gets richer or to trace what has been done during the exploration phase.

Figure 2 illustrates the expression language of a textual query for the first move through the graph. The presented query is based on a proof-of-concept study building a cohort of patients with diabetes (first move in Figure 2: patients over 20 years old, hospitalized in 2016 and whose biological measurements and received treatments corresponded to the chosen characteristics).

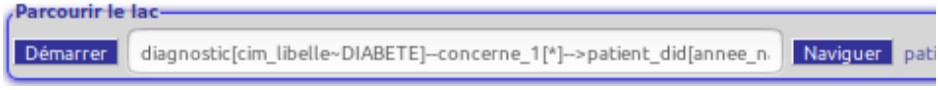

***Figure 2**: Generated query hop during navigation*

3. Results

Through the interactive iterations on this graph, the clinician can visualize the information on the obtained cohort in different ways. Graphical visualization tools associated with the cohort make it possible to highlight the dispersion of patients over the territory or to construct a dashboard with various indicators (see Figure 3).

Based on the same data structure, a visual exploration of the cohort being built provides access to a "360° view" of the cohort elements. The visualization interface in a web browser interacts with the graph database through a software layer (REST API)

that gives an abstraction of the model. This architecture thus limits a strong adherence with the database.

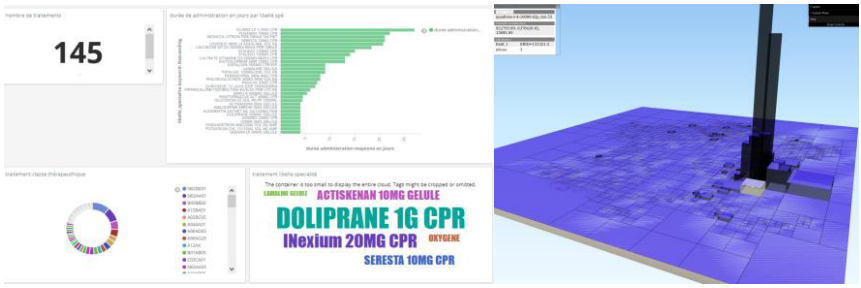

***Figure** 3: Interactive views describing the current selected data-set*

This "360° view" of the patient allows to visualize the different sides of the patient file. The first facet resumes the diagnoses based on the ICD10 codes and the timeline provides the chronology of these diagnoses. Those codes are projected onto a body silhouette highlighting the affected organs or body parts. An oval-shaped representation gives a synthesis of the pathological classes concerned. Other facets provide access to the various documents available in the patient's file. As with diagnoses, a timeline is associated with the list of documents such as drug prescriptions. An equivalent view is available for pharmaceutical prescriptions giving both the list of drugs administered or prescribed (therapeutic class, name of the drug, etc.).

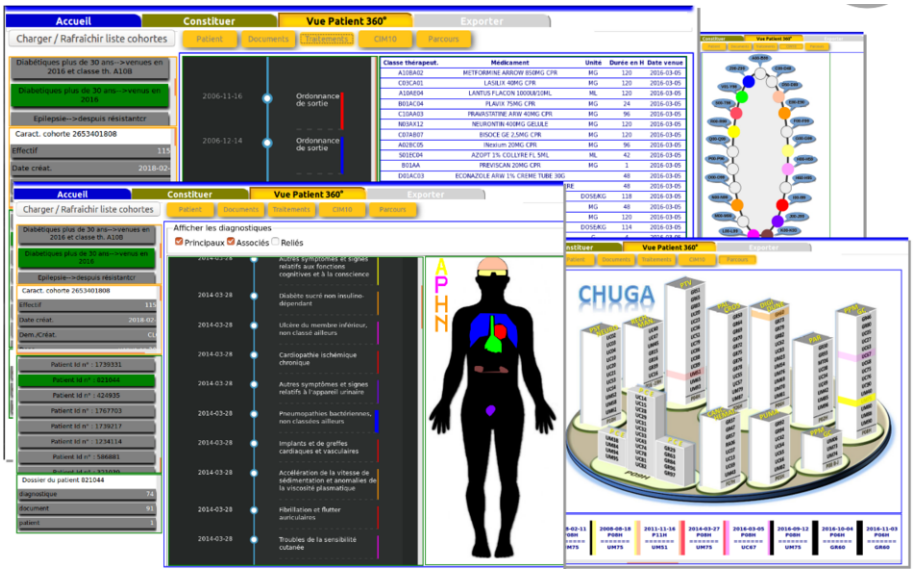

***Figure** 4: Different views on the patient's file.*

4. Discussion

4.1. Flexibility, evolutivity of the model, code genericity.

Graph modeling allows similar concepts to be grouped together and is well-suited to represent flexible and evolving structures. Graph data model implementation with a

graph database is almost straightforward and the high level of manipulated objects is appropriate for the development of generic code, independent of the future data variants.

4.2. Implementation and interoperability

The application environment we have built to operate PREDIMED is based on an open source property graph database, ArangoDB. This in-memory system enables to navigate seamlessly in such complex, highly-connected, massive data organizations. The native JSON-centered data application environment provides both the ability to comply with normative health data standards, to deal with complex objects' attributes (lists, matrices...) and the ability to distribute processing on different and evolutive applications of the PREDIMED architecture. Interoperability is also facilitated by the ability to interconnect with other graphs, using online APIs provided by evolving global data standards for health information such as ICD11.

4.3. Scalability

Unlike a conventional relational database system whose complete request must be written in advance, PREDIMED's alternative request system allows researchers to navigate easily through the data and record the path corresponding to a “built-on-the-fly” request. As the request is executed step by step, even when it is eventually relaunched entirely, there is potentially no limit to the number of steps that constitute the response path to a request. The size of the request and therefore its complexity have no impact on the processing infrastructure.

5. Conclusion

PREDIMED Data Warehouse allows non-IT specialists to use complex and massive data mining tools to effectively and rapidly design patient cohorts for research projects. PREDIMED provides agility and auto-improvement ability in data navigation while making the work traceable, reproducible, and transferable to other data-sets.

This work has been performed as proof-of-concept. The independent supervisory authority in terms of the GDPR (CNIL in France) has authorized the CHUGA to proceed to the operational phase of PREDIMED, on October 10 2019.

References

[1] Artemova S, Madiot P, Caporossi A, PREDIMED group, Mossuz P, Moreau-Gaudry A, PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital, MEDINFO, Lyon, 2019.
[2] Birtwell D, Williams H, Pyeritz R, Damrauer S, D. L. Mowery DL, Carnival: A Graph-Based Data Integration and Query Tool to Support Patient Cohort, MEDINFO, Lyon, 2019.
[3] Bouillé F, The Hypergraph-Based Data Structure: A New Approach to Data Base Modelling and Application, In:Schneider H.J. (eds) GI-7. Jahrestagung. Informatik - Fachberichte, vol 10. Springer, Berlin, Heidelberg, 1977.
[4] Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, Plantikow S, Rydberg M, Selmer P, and Taylor A, Cypher: An Evolving Query Language for Property Graphs, in: Proceedings of the 2018 International Conference on Management of Data, ACM,New York, NY, USA, 2018.