Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-121

Querying EHRs with a Semantic and Entity-Oriented Query Language

Romain LELONG[a], Lina SOUALMIA[a,b], Badisse DAHAMNA[a], Nicolas GRIFFON[a,b] and Stéfan J. DARMONI[a,b,1]
[a] *Department of Biomedical Informatics, Rouen University Hospital, France*
[b] *French National Institute for Health, INSERM, LIMICS UMR-U1142, Paris, France*

Abstract. While the digitization of medical documents has greatly expanded during the past decade, health information retrieval has become a great challenge to address many issues in medical research. Information retrieval in electronic health records (EHR) should also reduce the difficult tasks of manual information retrieval from records in paper format or computer. The aim of this article was to present the features of a semantic search engine implemented in EHRs. A flexible, scalable and entity-oriented query language tool is proposed. The program is designed to retrieve and visualize data which can support any Conceptual Data Model. The search engine deals with structured and unstructured data, for a sole patient from a caregiver perspective, and for a number of patients (e.g. epidemiology). Several types of queries on a test database containing 2,000 anonymized patients EHRs (i.e. approximately 200,000 records) were tested. These queries were able to accurately treat symbolic, textual, numerical and chronological data.

Keywords. Electronic Health Records, Information Storage and Retrieval, Search Engine, Controlled vocabulary

1. Introduction

Electronic Health Records (EHR) play a central role since they include a long-term record of care and a record of events from different types of care, including instructions, prospective information such as plans, orders and evaluations. In this context, the goal of an Information Retrieval (IR) System on EHR is to provide physicians with the correct information at the right place for the right person. Several tools and frameworks for searching in EHRs for one patient have been proposed. These tools are adapted according to each data format: structured, not structured or mixed. The main system is Informatics for Integrating Biology and the Bedside (I2B2), an open source platform developed in the USA and dedicated to translational research. The I2B2 center focuses on developing a scalable informatics framework to bridge clinical research data with basic sciences research data. The framework uses coded data, biological data and other genomic data. The scope of the search concerns clinical search and statistical data analysis. Data semantics is particularly important as it derives from the concrete healthcare providing process in hospitals. EHR data is mainly composed of several key

[1] Pr. Stéfan J. DARMONI, MD, PhD, Department of Biomedical Informatics, 1 rue de Germont 76031 Rouen Cedex, France, Cours Leschevin, Porte 21, 3^{éme} étage, Email : Stefan.Darmoni@chu-rouen.fr

entities semantically related to one another: (a) patient, (b) hospital, (c) stay and then (d) the "classical" and more basic levels (procedures, diagnosis related group (DRG) coding, lab tests, reports, metadata from reports etc.). As a consequence, IR from EHR is more difficult and different when compared to the "classical" IR. In this context, the aim of this study was twofold. First, describe a conceptual data model (CDM) which represents the conceptual and intuitive representation that non-IT medical provider users can have of EHR data. Secondly, describe a query language (QL) used to query those data and providing users the possibility to build queries accessing the entire set of EHR entities by taking advantage of the semantic network of entities. This study has been carried out within the context of the Retrieval and Visualization In Electronic Health records (RAVEL) project.

2. Materials

EHR Data Sources: A corpus of 2,000 anonymized patients and 200,000 reports from Rouen University Hospital (RUH) was used in this study, approved by the French National Commission on Computers and Liberty. Almost any clinical information available in the EHR is integrated in the RAVEL model, e.g. DRG codes (ICD10), patient data (age, gender), lab tests and all medical reports.

EHR Conceptual schema and data model: The underlying database of the system is based on a generic Entity-Attribute-Value (EAV) physical data model [1]. This data model is able to integrate all types of data in only a few tables without structural changes to the data model (e.g. columns or tables addings). This helps to optimize IR, maintain the database and manage heterogeneous data types. A dedicated CDM was designed to abstract the EHR data contained in the physical database data model. The query language syntax is patterned on that CDM instead of the physical database schema which provides the Search Engine (SE) with semantic features and capabilities.

3. Materials

3.1. Query Language Description

The specific QL syntax is based on the CDM. Hence, building a query only requires real-life knowledge of existing entities in the database, their properties and their relationships with each other. This QL has three main characteristics:

- *Semantic IR capabilities:* The QL is built with an entity-oriented vision. It enables semantic information retrieval since it provides the ability to display and query EHRs semantically related entities on any level (patient, stay, procedure, biology etc.). It can also deal with multiple terminologies and hierarchical relationships.
- *Scalability & flexibility:* The QL automatically handles modifications on the CDM (i.e. new conceptual entities, attributes and relationships between entities) without any SE modification. This enabled an easy and rapid extension to omics data [2].
- *Comprehensive querying:* The full scope of entities can be queried using constraints built upon several types of data: Textual and symbolic data (e.g.

patient(gender="M")), Numerical data (e.g. *medicalTest(6<numericResult<= 6.25))* and Chronological (eg. *stay(entryDate>2010-03-10))*. All comparators and operators available are specified in Table 1.

Table 1. Types of data handler by the search engine

Data type	Available operators	Available comparators
Character string data	None	= (equal), != (not equal), * (wildcard)
Numerical data	+ (add), - (substract), * (multiply), / (divide)	=, !=, < (lower), <= (lower or equal), > (greater), >= (greater or equal)
Chronological data	+,-	=,!=,<,<=,>,>=

3.2. Query Language Description

Basic querying: The query language is basically composed of nested syntactical units with the following syntax *ENTITY(CONSTRAINTS_CLAUSE)*. *ENTITY* can correspond to any kind of entity of the CDM (e.g. *patient*, *stay*, *medicalUnit* etc.) and specify the type of object that the SE should return (or target when nested). For instance, the queries *patient()* and *medicalUnit()* would respectively return all the patients and all the medical units of the database. The *CONSTRAINTS_CLAUSE* is a boolean expresion enabling to apply constraints to the targeted *ENTITY*. For instance, the query *patient(birthDate=1937-01-01 AND gender="M")* uses the two attributes birthDate and gender of the patient entity to return all male patients born on *1937-01-01*. *stay(leavingDate–entryDate>=10)* will return stays with a duration of 10 days or more.

Semantic querying: The strength of the query language originates from its ability to deal with nested syntactical units. For instance the query *stay(patient(id= "DM_PAT_42"))* targets stays link to at least one relationship to the patient 42. More complex queries can be performed by using the relationships between these entities (Table 2). This nesting functionality allows the exploitation of the relationships between entities and thereby enables to build queries based on the full semantic network. The QL has other querying capabilities: full text search, minimum and maximum on numerical data, hierarchical expansion, chronological and temporal queries.

4. Result

Several use cases were successfully answered in the RAVEL project:

- *Use case 1:* Visualize over time the neutrophil rate of a patient with rheumatoid arthritis,
- *Use case 2:* Produce all the medical reports containing the concept of metastasis,
- *Use case 3:* Retrieve all stays where *"REMICADE"* (infliximab) was used.

The use cases resolution required to use: Automatic Indexing in medical records, full text search, and multiple terminological resources. Some of the queries used to answer these three use cases are shown in Table 2.

4.1. Comparison to I2B2 workbench

The I2B2 workbench and the QL described in this study are both tools designed for searching in EHRs. However, the two tools have differences which are summarized in Table 3. The I2B2 workbench provides numerous default features which cover a lot of use cases. It notably enables to detect the number of occurrences of an event contrary to the QL described in this study. The database on which the QL operates integrates currently 69 English and French terminologies which represent 2,340,655 concepts partially translated into French. I2B2 workbench includes 14 terminologies (cf. Table 3) English for the major part. Other terminologies can be added. In contrast to I2B2 workbench, reports are automatically indexed and can be queried using the terminology terms with the QL. As regards cohort patient selection, I2B2 and the QL share most of their functionalities such as: numerical, chronological and textual constraints, full-text search on reports, search using concept subsumption, use of clinical data as constraints (stay, medical unit, patient, etc.) and omic variant data management.

Table 2. Query examples

Semantic query examples	*stay(patient(id_"DM PAT 1736") AND medicalUnit(label="Cardiology"))* All the patient 1736 stays which occur in the Cardiology medical unit.
	stay(icd10SC(label="Burns involving less than 10% of body surface")) stays with a diagnosis of *Burns involving less than 10% of body surface* (T31.0 sub category of ICD10).
	medicalTest(medicalTest(label="Sodium") AND numericResult<lowerBound AND patient(id="DM PAT 125")) For a given patient (number 125), display all hyponatremia test results.
	patient(stay(icd10SC(id = "CIM SC T31.0") AND medicalTest(exe(label="Sodium") AND numericResult>upperBound))) patients coded with the T31.0 sub category of ICD10 DRG code showing hypernatremia in that stay.
RAVEL queries	*stay(patient(id="DM PAT 21") AND procedure(label="BLOOD SAMPLE"))* Patient 21 stays in which a blood sample procedure was performed.
	medicalUnit(stay(patient(id="DM PAT 21") AND procedure(label="BLOOD SAMPLE"))) Medical units of the patient 21 stays in which a blood sample was taken.
	*biologicalTest(patient(id="DM PAT 1078") AND exe(label="Platelets") AND 10*numericResult<lowerbound)* Patient 1078 platelet tests with a result more than 10 times lower than normal level.
	procedure(ccamMP(id="CCA AM EQQM006") AND procedureDate="MAX") The last procedure coded with EQQM006.

Table 3. QL vs I2B2 Functionalities

	QL	I2B2
Querying scope	1 or n entity	n patients
Querying	Textual query	Graphical query
Detection of number of event occurrences	NO	YES
Lab test unit choice	NO	YES
Defaultly supported terminologies	69	14
Record Automatic Indexing	YES	NO
Omic data expression analyses (genes, proteins, micro-RNA, exons)	YES	PARTIALLY

5. Discussion

As described by Terry et al.[3], there are five basic options for searching specific data in EHR: (i) pre-determined queries: users select a query option from the software menu; (ii) simple customizable queries: users have some input into the queries to generate reports; (iii) advanced customizable queries: allow a greater amount of user input than the second level, often using Boolean logic; (iv) structured query language interface: using a special interface to enter Structured Query Language (SQL) commands; (v) data extraction and analysis with database tools. To date, the query language described in this paper is able to deal with levels 1 to 4 of Terry et al [3]. The global architecture of the underlying EHR system and the data querying strategy is closer to level 5 than to level 4 since, as reported by Terry et al [3] regarding level 5, the query language is based on the EHR's conceptual model. However, more advanced data analysis querying possibilities would probably be necessary to be considered as a full level 5 search options. Despite the fact the query language is quite complex to use, the public health professionals to whom it has been presented in fact stated that they would be able to use it after basic training. This training should also enable medical librarians, information scientists and IT specialists to use it. However, in contrast, several graphical user interfaces will be needed for health care professionals. These interfaces should provide access to more customizable queries than simple search. The I2B2 graphical interface could be a source of inspiration. To address this difficulty, an information extraction method was also designed in [4] to allow physicians to query EHRs using natural language instead of the dedicated QL. The SE has been tested outside the Rouen University Hospital, Normandy: at Bordeaux University Hospital, Aquitaine, France. However, the current model still does not operate on the establishment level but should become operational in the near future. Furthermore, the comparative evaluation of this query language with I2B2 should be improved. A parser enabling to share data between I2B2 data model and the RAVEL data model could be implemented to accurately assess precision as well as querying scope of the query languages. A scaling up study is underway at Rouen University Hospital with all the patients with at least one stay (in or outpatient) in the dermatology department since 1992 (n=65,000). This study aims at querying EHR data in a multi-patient context in order to create a patient cohort.

6. Acknowledgements

This search engine was partially funded by the French National Agency (TecSan program) in the RAVEL project.

References

[1] Prakash M Nadkarni. Qav: querying entity-attribute-value metadata in a biomedical database. Computer methods and programs in biomedicine, 53(2):93–103, 1997.

[2] Chloé Cabot, Julien Grosjean, Romain Lelong, Arnaud Lefebvre, Thierry Lecroq, Lina F. Soualmia, and Stéfan J. Darmoni. Omic data modelling for information on retrieval. In 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, pages 415–424, 2014.

[3] Amanda L Terry, Vijaya Chevendra, Amardeep Thind, Moira Stewart, J Neil Marshall, and Sonny Cejic. Using your electronic medical record for research: a primer for avoiding pitfalls. Family Practice, 27(1):121–126, 2010.

[4] Lina F Soualmia, Romain Lelong, Badisse Dahamna, and St´efan J Darmoni. Rewriting natural language queries using patterns. In Multimodal Retrieval in the Medical Domain, pages 40–53. Springer, 2015.