Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-194

RAVEL: Retrieval And Visualization in ELectronic health records

Frantz THIESSARD[a,b,1], Fleur MOUGIN[a], Gayo DIALLO[a], Vianney JOUHET[a,b], Sébastien COSSIN[a], Nicolas GARCELON[c], Boris CAMPILLO[c], Wassim JOUINI[c], Julien GROSJEAN[d], Philippe MASSARI[d], Nicolas GRIFFON[d], Marie DUPUCH[e], Fayssal TAYALATI[e], Edwige DUGAS[e], Antonio BALVET[e], Natalia GRABAR[e], Suzanne PEREIRA[f], Bruno FRANDJI[g], Stefan DARMONI[d], Marc CUGGIA[c]
[a]*CHU de Bordeaux, Pole de sante publique, SIM, F-33000 Bordeaux, FRANCE*
[b]*Univ. Bordeaux, ISPED, LESIM, F-33000 Bordeaux, FRANCE*
[c]*Inserm U936 – Univ. Rennes 1, France*
[d]*CISMeF, Rouen Univ. Hospital, LITIS EA 4108, Rouen, FRANCE*
[e]*STL, CNRS UMR 8163, Univ. Lille 3, FRANCE*
[f]*VIDAL, Issy les Moulineaux, FRANCE*
[g]*MEDASYS, FRANCE*

Abstract. Because of the ever-increasing amount of information in patients' EHRs, healthcare professionals may face difficulties for making diagnoses and/or therapeutic decisions. Moreover, patients may misunderstand their health status. These medical practitioners need effective tools to locate in real time relevant elements within the patients' EHR and visualize them according to synthetic and intuitive presentation models. The RAVEL project aims at achieving this goal by performing a high profile industrial research and development program on the EHR considering the following areas: (i) semantic indexing, (ii) information retrieval, and (iii) data visualization. The RAVEL project is expected to implement a generic, loosely coupled to data sources prototype so that it can be transposed into different university hospitals information systems.

Keywords. Electronic health record, datawarehouse, semantic indexing, information retrieval, data visualization.

Introduction

The medical domain is experiencing a paradigm change in the way healthcare professionals interact with patient data: clinical data become more and more often defined as "the commodity used to help make patient care decisions" [1]. With the development of Electronic Health Records (EHRs), clinical data are constantly growing and becoming eventually accessible to patients as well.

The EHRs contain a great variety of heterogeneous clinical data which present different informational and decisional values for health care professionals: they are used in different ways and in different situations. Moreover, a large amount of data are still recorded as unstructured and narrative documents although these data may contain

[1]Corresponding Author. Frantz Thiessard, E-mail: Frantz.thiessard@isped.u-bordeaux2.fr

crucial information for a more efficient health care process of patients such as clinical notes, observations, discharge summary reports to name a few.

The sheer accumulation of clinical data within the EHRs, as well as their diversity, leads to the risk of information overflow for both health care practitioners and patients. Moreover, it is considered that up to 50% of the clinical information needed to describe patients' hospital stay is only available within narrative unstructured documents within the patient records [2]. Thus, the evolution of patient records towards the EHR leads to paradoxical consequences: the data are exhaustive yet scantly exploited; they lack a structured and hierarchical presentation, which burdens the decision making process by the medical practitioners.

It is thus essential to provide robust and efficient retrieval and visualization tools for EHRs data. In this frame, the RAVEL project will investigate and implement: (i) the most recent indexing methods in order to enrich semantically structured and unstructured EHRs data. This enrichment will be useful for (ii) information retrieval optimization, and (iii) patients' data visualization.

1. Methods

The strong point of the consortium of the RAVEL project is that the partnership is complementary from the implementation point of view (it contains two industrial companies, two hospitals and four academic teams) and from the scientific point of view (the involved partners bring different and complementary skills). This situation gives us interesting and unique possibility to work on real clinical challenges and to provide scientifically relevant and ambitious solutions. These solutions will thus be coupled with the DxCare® suite and implemented and tested in two French university hospitals' Information System (Rennes and Bordeaux).

Other partners of the RAVEL project bring with them their experience and skills in various areas and provide a contribution for a better access to the semantic contents of EHRs and of their automatic processing. In the following of this section, we describe the challenges we have to face and the methods which we will develop, implement and evaluate at the bedside of patients with the healthcare professionals.

The method of the RAVEL project is organized into seven steps. The first step aims at defining relevant use cases with a strong participation of the clinicians. For illustrating the features required by clinicians in the course of information collection on a given patient, the use cases will identify the functionalities that each work package has to address and they will provide the basis for evaluating the implemented prototype, making their definition extremely important. Then, the following step is related to the data Extraction, Transformation and Loading into a data repository by reusing existing open source tools, such as TALEND [4]. For a normalization and standardization purposes, an alignment of some biomedical terminologies (in particular, MedDRA, ICD10 and SNOMED International) will be performed so that data coded with distinct terminologies can be related to each other. The next step consists in establishing the state of the art of existing solutions to represent EHR (e.g., I2B2 [5]) and the development of a generic database model to handle heterogeneous data (medical data, terminologies, patients, cities, and so on). This generic database model will not only have to store heterogeneous data but also to wrap other models in order to keep the original structure of the data. The objective of the following step is to perform document content processing (see for example [6]) in order to detect relevant

elements from raw text documents and to associate them with the appropriate terminology and semantic tags, as well as to detect the certainty status of these elements. Namely, this process relies on Natural Language Processing (NLP) methods. It will enrich the EHRs and enable information retrieval and visualization tasks. Semantic indexing is used either to find or summarize clinical information [7]. Beyond the simple indexing of the healthcare documents, this step provides also the meta-information of the indexing terms. For instance, it indicates whether a given patient problem is present or absent, or even if its existence is questionable. Indeed, these different situations do have not the same impact on the indexing status and on the medical decisions. Another meta-information is provided according to the temporality. It allows knowing whether a given clinical event is positioned in the past, present and whether it is terminated or not. On the basis of the indexing data, a specific task is in charge of deploying a generic search engine, which will be coupled with the generic database. This engine should be able to execute logical queries on every type of data stored. To do so, existing engines will be listed (among them, MorphoSaurus [8]) and best parts of each one will be reused and adapted to this project. Since this project brings new challenges in information retrieval, a special effort will be made to enhance the capabilities and performance of the engine, including special operators (e.g. =, >, < to manipulate numerical data, in particular biological tests) and special keywords that will be able to handle time aspects (before, after …). Finally, another task which relies on the indexing and document search is related to the visualization of the clinical events. More particularly, it aims at developing and integrating the different views of patient data in the EHR. Previous works on visual representation of medical information in EHRs will be first explored [9]. Then, the best types of representations for the use cases will be designed. Interactivity functionalities with the user (zooming, grouping, filtering and so on) can also be planned in the display so that the general and dynamic aspect of the visualization is preserved. In order to guarantee the quality and acceptability of the automatic processing, a specific task is dedicated to the evaluation. A first phase will focus on evaluating the quality assurance and monitoring of each step by assessing the performance achieved by the different sub-components developed within the RAVEL project. In a second phase, which will be held at the end of the project, an overall evaluation of the prototype will be conducted focusing on its adequacy to the needs expressed in the use-cases with the cooperation of the clinicians. On the whole, the proposed methods aim at satisfying the challenges tackled in this project: the most recent indexing methods in order to enrich semantically structured and unstructured EHRs data, efficient and optimized information retrieval and patients' data visualization within an ergonomic interface.

2. Results

This project received the support of the French Agence Nationale de la Recherche (ANR). It began in January 2011 and will last three years. The prototype will be constructed according to a small set of use cases concerning complex patients (This complexity can result in a very long follow-up or in polypathological patients). For instance, one EHR could concern a 60 years old patient hospitalized for chemotherapy for an esophagus cancer diagnosed 3 years ago initially treated with sessions of radiotherapy. This patient has an implantable drug delivery system and esophagus prosthesis. He suffers besides a well-treated insulin-dependent diabetes. Clinicians

would like to get a synthesis for each medical problem (cancer and diabetes) and always have in the visualization template durations since the diagnosis of cancer, the remission and the recurrence if occurred. It should also make a synthesis of the cures of chemotherapy as well as the cures of radiotherapy and the intercurrent events. Healthcare professionals should be able to query in natural language, the number of infections of the implantable drug delivery system, or the results of all previous cardiac echography and optic fundus useful for the systematic follow-up of the diabetes. The first step will be to extract, transform and load the whole data of the patients corresponding to the use cases into a data repository. Some of the required information can be located in the EHR within narrative unstructured documents or as formal structured data/ Different documents may be involved: the laboratory results, the codes used to calculate the diagnosis-related group, the imaging reports, the prescriptions, the letters addressed or received from other practitioners, etc. Then document content processing will be performed, relying on NLP methods to find the relevant elements and to associate them with the terminology and semantic tags (as well as to detect the certainty status of these elements, the associated date …). On the basis of these available and NLP-generated data, a lifeline will be automatically drawn with the key information concerning cancer and other pathologies. The search engine should be able to execute logical queries (including Booleran queries) on every type of data stored. This search engine should (a) have semantic properties, allowing management of several terminologies and ontologies; (b) manage temporal and numerical data, particularly biological data. For instance, the query "find the number of infections of the implantable drug delivery system" would have to find information from hemocultures, implantable drug delivery system's cultures, antibiograms, notion of fever or quantitative temperature, free text description of the infection, codes corresponding to this problem or the replacement of the implantable drug delivery system.

At the medical level, the project will allow the synthesis of complex heterogeneous information in an understandable way and will get a better readability of clinical information by professionals. This will result in a wider-spread endorsement of the EHR by health professionals (especially by physicians), and lead to an improvement of the quality of healthcare, a faster access to the patient information and the support of medical decisions, epidemiological studies and teaching activities for medical or nurse students.

The exploitable results of the RAVEL project will be integrated into the DxCare product portfolio, in compliance with the guidelines that will be set up by the Consortium Exploitation Agreement and Business Plan. They will be potentially available to other hospitals. The expected benefits include in particular the improvement of the EHR readability and usability and the enrichment of patient data with several reference terminologies that are used in interoperability standards.

3. Discussion

The RAVEL project aims at achieving effective and efficient tools to locate in real time relevant elements of the patients' EHR and visualize them according to synthetic and intuitive presentation models. It has to perform a high profile industrial research and development program on the EHR considering the following areas: (i) Semantic indexing, (ii) Information retrieval (iii) Data visualization. The scientific research

program of the project focuses on the improvement of the state of the art on the three mentioned areas. The RAVEL project is expected to implement a generic, loosely coupled to data sources prototype so that it can be transposed into different University Hospital Information System. This prototype will be evaluated in situations of actual use. Many benefits are expected from the RAVEL project. From the medical point of view, one of the key outcomes will be improving the decision-making process and therefore enhancing patients care. From the societal point of view, the methods and tools developed in this project will allow patients to be more proactive in their approach to healthcare by allowing to feel more empowered about their own medical data. From the scientific point of view, the RAVEL project is a unique opportunity to develop advanced semantic indexation and NLP methods, approaches for the semantic integration of heterogeneous data, retrieval and multimodal representation methods applied to medical data. Finally, from the industrial point of view, the innovations resulting from the project will allow a significant functional and technological leap in an increasingly competitive market.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) and the DGA, under grant Tecsan (ANR-11-TECS-012).

References

[1] Wyatt J. Medical informatics, artefacts or science? Methods Inf Med. 1996;35(3):197–200.
[2] Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc. 2006;13(6):691–5.
[3] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, Happe A, Duvauferrier R. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform 2011;169:584-8.
[4] http://www.talend.com (last accessed January 27, 2012).
[5] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.
[6] Krauthammer M, Nenadic G. Term identification in the biomedical literature. J. of Biomedical Informatics. 2004;37(6):512–26.
[7] Shang Y, Li Y, Lin H, Yang Z. Enhancing Biomedical Text Summarization Using Semantic Relation Extraction. PLoS One. 2011; 6(8): e23862
[8] Schulz S, Daumke P, Fischer P, Müller M, Müller ML. Evaluation of a document search engine in a clinical department system. AMIA Annu Symp Proc. 2008; 647–51.
[9] Roque FS, Slaughter L, Tkatšenko A. A comparison of several key information visualization systems for secondary use of electronic health record content. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents. Stroudsburg, PA, USA: Association for Computational Linguistics 2010;76–83.