MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190476

SmartCRF: A Prototype to Visualize, Search and Annotate an Electronic Health Record from an i2b2 Clinical Data Warehouse

Sébastien Cossin[ab], Luc Lebrun[ab], Niamkey Aymeric[ab], Fleur Mougin[ab], Mathieu Lambert[c], Gayo Diallo[ab], Frantz Thiessard[ab], Vianney Jouhet[ab]

[a]Bordeaux Hospital University Center, Pôle de santé publique, Service d'information médicale, Unité Informatique et Archivistique Médicales, F-33000 Bordeaux, France
[b]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France
[c]Bordeaux Hospital University Center, Pôle spécialités médicales, Service de médecine post-urgences

Abstract

Clinical information in electronic health records (EHRs) is mostly unstructured. With the ever-increasing amount of information in patients' EHRs, manual extraction of clinical information for data reuse can be tedious and time-consuming without dedicated tools. In this paper, we present SmartCRF, a prototype to visualize, search and ease the extraction and structuration of information from EHRs stored in an i2b2 data warehouse.

Keywords:

Electronic Health Records, Information Storage and Retrieval, User-Computer Interface

Introduction

A significant part of data daily produced in Electronic Health Records (EHRs) is either in an unstructured (free text) or semi-structured (forms) format [1]. Bordeaux University Hospital is no exception to this observation. Information extraction (IE) for secondary use of clinical data consists in transforming heterogeneous data in an EHR to a structured format of a case report form (CRF). A CRF is a specialized document used in clinical research to collect standardized information about a patient for further statistical analysis [2]. So far, researchers needed to transcribe the data from an EHR to a CRF manually. With the ever-increasing amount of information in patients' EHRs, this task has become tedious and both time and cost-consuming [3]. RAVEL [4], a previous research project carried on in Bordeaux has shown the importance of search engine and data visualization tools to retrieve information in EHRs.

The objective was to develop an interface to speed-up information extraction (IE) task for researchers.

Methods

Bordeaux University Hospital deployed an i2b2 [5] data warehouse for secondary use of medical data. Multiple data sources were integrated such as claims data, lab tests, drug prescription and dispensing, discharges summaries and radiology reports. On the top of i2b2, we've developed two tools for patient centered information retrieval and exploration:

- **i2b2 webclient timeline plugin**: This plugin was developed using D3.js library. The plugin interacts with the i2b2 data warehouse to retrieve data. A simple free text search engine was added in order to find relevant observations.
- **Standalone prototype for data exploration and CRF data capture**. Unstructured data of an electronic health record is first normalized, tokenized and lemmatized with Stanford NLP tools and TreeTagger [6] for French language. Then, noun phrases are extracted with linguistic methods and regular expression [7]. These extracted terms are then indexed in ElasticSearch™ for autocompletion and information retrieval. A web interface was developed with the Shiny package [8] of the R programming language to visualize information in an EHR.

Results

I2b2 webclient timeline plugin

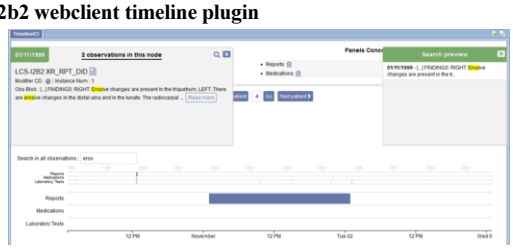

Figure 1 Search in the i2b2 timeline plugin

Figure 1 is a screenshot of the i2b2 timeline plugin. Selected observations were filtered using the search engine with the term "eros". The timeline shows a preview of the records that matched the query. Detail can be obtained by clicking on the preview panel. The term searched is highlighted. The timeline can also manage numerical values. The plugin is available under a GPL-3.0 license[1].

Standalone prototype for data exploration and CRF data capture

An EHR overview is displayed with a timeline (figure 2). The user can zoom in and zoom out the timeline. Each element on the timeline is clickable to show its content. Every element of a same data source has the same background color and different elements within a data source have different colors icons. For example, each ICD-10 category has a specific icon to represent the disease, disorder or symptom.

[1] https://github.com/vianneyJouhet/TimelineD3

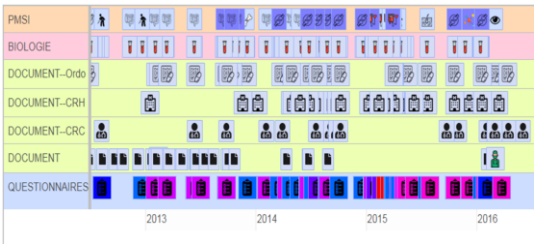

Figure 2 - Timeline of a single patient's EHR. Each element is clickable to display its content. Each background color denotes a data source (claims data, lab tests, discharge summaries and medical forms).

The search engine suggests several noun phrases when a user starts typing (figure 3). Autocomplete finds terms present in the EHR and speeds up human-computer interactions.

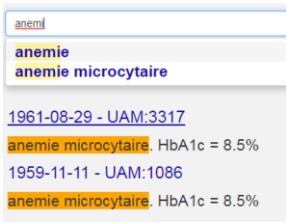

Figure 3 - The search box. Only the sentence containing the word is displayed in the results and the term being queried is highlighted.

A wordcloud of symptoms / diseases is displayed by default (figure 4). Each term weight corresponds to its Term Frequency-Inverse Document Frequency (TF-IDF) value. A click on a term triggers a search query in the search box. Special focuses on structured data are available: a sunburst displays all ICD-10 (International Classification of Diseases, 10th revision) codes and a scatterplot shows numerical values of selected lab tests. Each sub-element of interest, like a lab test result or an ICD-10 code can be added to the timeline. The idea is to have an adaptive interface where user could decide what he wants to display for a specific IE task. For example, if the user must enquire about a past history of anemia for a set of patients he can choose that hemoglobin test and ICD-10 codes of anemia must appear on the timeline by default and the wordcloud must show terms found in unstructured data related to this disease.

To extract information, an annotation module lets the user select a textual content or a code and stores it in a database.

The prototype is open-source[2] and can be tested online[3]. The prototype was tested by medical users that provided valuable feedback to improve the interface and human-computer interactions. A software version based on this prototype is currently developed to be fully integrated in an i2b2 module.

Our next objective will be to connect the interface directly to REDCap (Research Electronic Data Capture) [9], an electronic CRF application, used in our hospital to store data of clinical studies. The REDcap export module allows the user to export data in specific formats for different analysis software.

[2] https://github.com/scossin/SmartCRF

[3] http://www.smartcrf.fr/

Figure 4 - The wordcloud. The bigger the term, the higher the TF-IDF value of that term. A click on a term triggers a search.

Conclusion

In this article, we present a clinical information extraction prototype to speed-up clinical research and facilitate data reuse in EHRs. Our next objective will be to develop an industrialized version fully integrated in our i2b2 data warehouse and to evaluate its performance in terms of usability and precision/recall for data extraction tasks.

References

[1]P. Raghavan, J.L. Chen, E. Fosler-Lussier, and A.M. Lai, How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?, *AMIA Summits Transl. Sci. Proc.***2014** (2014) 218–223.

[2]S. Bellary, B. Krishnankutty, and M.S. Latha, Basics of case report form designing in clinical research, *Perspect. Clin. Res.***5** (2014) 159–166. doi:10.4103/2229-3485.140555.

[3]Y. Matsumura, A. Hattori, S. Manabe, T. Takeda, D. Takahashi, Y. Yamamoto, T. Murata, and N. Mihara, Interconnection of electronic medical record with clinical data management system by CDISC ODM, *Stud. Health Technol. Inform.***205** (2014) 868–872.

[4]F. Thiessard, F. Mougin, G. Diallo, V. Jouhet, S. Cossin, N. Garcelon, B. Campillo, W. Jouini, J. Grosjean, P. Massari, N. Griffon, M. Dupuch, F. Tayalati, E. Dugas, A. Balvet, N. Grabar, S. Pereira, B. Frandji, S. Darmoni, and M. Cuggia, RAVEL: retrieval and visualization in ELectronic health records, *Stud. Health Technol. Inform.***180** (2012) 194–198.

[5]S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, *AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp.* (2006) 1040.

[6]H. Schmid, Improvements In Part-of-Speech Tagging With an Application To German, in: Proc. ACL SIGDAT-Workshop, 1995: pp. 47–50.

[7]S.S. Kathait, S. Tiwari, A. Varshney, and A. Sharma, Unsupervised Key-phrase Extraction using Noun Phrases, *Int. J. Comput. Appl.***162** (2017) 1–5.

[8] Shiny, http://shiny.rstudio.com/ (accessed November 25, 2018).

[9]P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde, Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support, *J. Biomed. Inform.***42** (2009) 377–381. doi:10.1016/j.jbi.2008.08.010.

Address for correspondence

sebastien.cossin@chu-bordeaux.fr