Using Clinical Data Warehouse to optimize the vaccination strategy against COVID-19: a use case in France

Julien Grosjean[a,b], Thibaut Pressat-Laffouilhère[a,c], Marie Ndangang[a], Jean-Philippe Leroy[a], Stéfan J. Darmoni[a,b]

[a] *Department of Biomedical Informatics, Rouen University Hospital, Rouen, France*
[b] *Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, U1142, INSERM, Sorbonne Université & Sorbonne Paris Nord, Paris, France*
[c] *Laboratoire LITIS EA4108, Rouen Normandie University, France*

Abstract

Clinical Data Warehouses (CDW) are gold mines and may be useful to manage the COVID-19 outbreak. This article details the use of CDW in order to retrieve patients for vaccination purposes. A list of 34 diseases (or conditions) was published by French Health Authorities to target individuals at a high risk of developing a severe form of COVID. Using a multilevel search engine, 23 queries were built based on structured or unstructured data using natural language processing features. The Diagnosis Related Group coding system was used alone in three queries (13.0%), coupled with unstructured data in four queries (17.4%), and unstructured data were used alone in 16 queries (69.6%). Eleven diseases (conditions) were too broad to be translated into queries. Finally, 6,006 unique re-identified patients were retrieved. This use case demonstrates the usefulness of the Rouen University Hospital CDW in retrieving patients for other purposes than translational research.

Keywords:

Data Warehousing; Natural language processing; COVID-19

Introduction

In recent years, Clinical Data Warehouses (CDW) have been developed in hospitals. Various tools relying on CDW have been created in order to mine information from structured and unstructured data such as EMERSE [1], CREATE [2] or Dr. Warehouse [3]. Their main common feature is the exploration of clinical narratives (CN) based on natural language processing (NLP) algorithms. These search engines are used to retrieve patients for trial eligibility, retrospective and prospective studies [4–7].

Faced with the COVID-19 outbreak, humanity has deployed all the resources at its disposal: lockdown, massive diagnosis, vaccination, etc. Electronic Health Records (EHR), feeding the CDW, have been modified and enhanced with tools to manage the growing influx of patients, to diagnose patients with COVID-19 and to improve follow-up care [8–11]. Specific CDW about COVID have been created for research purposes [12,13]. However, as far as we know, none have proposed solutions or logistical support for vaccination, which is a key weapon against COVID-19. On the other hand, several countries are using a *defacto* EHR at a national level to solve this problem (e.g. Israel) [14].

In France, the vaccination program against COVID-19 began on the 27th of December, 2020. The target population for the vaccine has evolved starting with the oldest and most vulnerable populations.

The objective of this study was to select patients that could be prioritized for vaccination using real world data from Rouen University Hospital's (RUH) CDW.

Methods

Entrepôt de Santé Normand (EDSaN)

The Entrepôt de Santé Normand (EDSaN) is an in-house solution to query RUH's CDW. EDSaN gathers clinical data since the 1990's from about 2 million patients. Several data types have been integrated so far from the EHR and various clinical data bases: structured data from biology, virology, diagnoses, procedures and unstructured data such as CN (discharge summaries, letters, procedure results, prescription letters, etc.). More precisely, diagnoses and procedures codes are collected from the French Diagnosis Related Groups (DRG), known as PMSI in French language.

The French Commission on Informatics and Liberty (CNIL) approved EDSaN in October 2020. The CNIL is an independent regulatory body whose mission is to ensure that data comply with privacy laws. The CNIL has verified the consistency of EDSaN with General Data Protection Regulations (GPDR) (EU 2016/679). RUH researchers have already used EDSaN for 170 studies since that date.

EDSaN consists in: (a) a query tool that allows to search and/or mine data. For example, it can be used to identify patients from different criteria; (b) a selection tool that allows to filter and explore datasets collected from (a).

The data are pseudonymized or de-identified in EDSaN to preserve patient anonymity. However, for specific purposes, it is possible to re-identify the data. For example, it could be important to contact patients for various reasons; COVID-19 vaccination is one of them.

The EDSaN query software consists in a multilevel search engine that is able to query structured data, unstructured data, and both structured and unstructured data at the same time. This is a very important feature as DRG codes can sometimes lack precision and can sometimes be missing. In these cases, the automatic processing of clinical narrative is used. It combines several NLP algorithms to ensure that searched keywords are relevant (i.e., not in a negative sentence for example) or present in text specific segments (such as conclusion for example) even if those documents are natively unstructured [15].

Patient selection

The French vaccination strategy has evolved several times in the first trimester of 2021. The French Health Authority (HAS) has published a list of diseases or conditions with a high risk of developing a severe form of COVID-19: 1) cancers or malignant hematologic diseases currently being treated with chemotherapies; 2) severe chronic kidney disease including patients treated with dialysis; 3) solid organ transplantation, hematopoietic stem cell allogeneic transplantation; 4) multiple chronic pathologies and presenting at least two organ failure; and 5) some rare diseases according to the following list: https://solidarites-sante.gouv.fr/IMG/pdf/liste_maladies_rares_cosv_fmr.pdf (published the 17th of December, 2020). Some examples of rare diseases/conditions are: amyotrophic lateral sclerosis, autoimmune pancreatitis type 2, Moyamoya angiopathy, trisomy 21.

Thus, RUH decided to use real-world data from EDSaN to identify patients based on the HAS list.

The following steps were applied:

1. Identification of target data types that should be used for each disease/condition (unstructured or structured data or both);
2. Translation of each disease/condition into a specific EDSaN query;
3. Processing and export of queries;
4. Re-identification;
5. Exclusion of deceased patients.

A global filter has been added to limit the queries to patients aged between 18 and 75 years (in 2021) because COVID-19 vaccines are indicated for adults only and elder people were already prioritized in the past months. Moreover, only clinical data produced since 2016 were considered for the queries. The purpose of this time restriction was to minimize the number of patients that could not be contacted (e.g. deaths, moved, cured).

Results

Among the 34 diseases/conditions published, a total of 23 queries were built to identify patients (67.6%).

Three queries (13.0% of build queries) required the DRG coding system exclusively because the quality of the codes was assessed high.

Four queries (17.4% of build queries) combined both structured (DRG codes) and unstructured data (CN).

Sixteen queries (69.6% of build queries) were performed only on unstructured data using NLP algorithms to enhance precision/recall.

Finally, eleven diseases/conditions (32.4%) could not be properly translated into an EDSaN query. For example, the patients with a deficit of AIRE, NFBK2 or interferons, or myopathies with forced vital capacities less than 70% are conditions that are very difficult to identify even in CN.

Some examples of queries are displayed in Table 1.

Table 1 – Examples of queries used to identify patients from diseases/conditions

Disease/condition	Data types (DRG or/and CN)	Query	Patients nb.
allogeneic hematopoietic stem cell transplantation	DRG	Z9480*	11
Patients treated with rituximab	CN	(rituximab OR truxima OR RIXATHON OR MABTHERA)	179
chronic pancreatitis and diabetes	DRG+ CN	("pancreatite chronique" AND diabet*) + (K86* AND (E10* OR E11* OR E14*)	265

A total of 7,781 patients were identified in this study. Then, 1,775 patients (22.8%) were excluded because they died (this data is available only in the RUH information system because they died in RUH. Many other patients may have died elsewhere but this information is not available). Among the 6,006 patients in the final list, 734

(12.2%) have several diseases/conditions (i.e., their clinical data matched with at least two diseases/conditions).

Discussion

Patients that could be prioritized for COVID-19 vaccination were selected using real world data from RUH's CDW. To achieve this goal, the EDSaN solution was used.

The use of such tools has advantages and several limits.

Advantages

Even if building specific queries is a time-consuming task, execution and automatic retrieval is much faster than manual screening. It is a huge time saver.

Moreover, because the queries can be processed on structured and unstructured data, the recall is increased especially for patients suffering from rare diseases not treated at RUH but mentioned in CN. Additional clinical data such as drug prescription are not always structured but this information is present in CN. For example, in this study, patients under rituximab should be identified: this data is not available in the RUH drug information system but it is possible to search it in CN.

Furthermore, the use of CN can sometimes increase precision, especially the identification of rare diseases. As DRG codes (ICD-10) can be imprecise, the use of specific keywords in text documents is often more precise. For example, cavernous hemangioma is not well defined in the ICD-10 and will be generally coded as D180 which regroups various types of hemangiomas.

Limits

Processing CN with NLP is never perfect: some false positives or false negatives can occur (e.g., in detection of negations) [15]. Thus, an evaluation should be conducted to measure the precision of each query. Unfortunately, the recall cannot be measured since, by essence, there are no complete lists of patients for these diseases/conditions.

Moreover, when some queries were quite easy to elaborate, some were very complex because recommendations lack precision or completeness (e.g. multiple chronic pathologies and presenting at least two organ failure). In these cases, it is rather obvious that false negatives exist.

More generally, the published list of diseases/conditions did not precise the temporal insight that would be considered (for non-chronic diseases), except for one query (cases of cancer in the last 3 years). In this study, the delay was arbitrarily set at 5 years in order to minimize patients lost to follow up. Nevertheless, once all patients from the list are contacted, it will be possible to reprocess the queries with an extended date limit if needed.

Approximately one third of the diseases/conditions could not be translated into an EDSaN query. The main reason was the lack of precision of the conditions or situations that were too complex to be expressed with a limited number of keywords in CN. For example, the situation "patients with constitutional bleeding diseases who use a drug in a clinical trial" is very broad. Those complex or fuzzy situations were ignored as corresponding queries led to zero patients (too narrow) or led to thousands of patients because of the lack of precision (too broad).

Finally, this study is only focusing on patients who came to the RUH in the last years. Even if this hospital is the major hospital of the region of Normandy, the use of national databases such as the "Système National des Données de Santé" (SNDS) in France could be an opportunity to identify more eligible patients.

Conclusions

This study demonstrates the usefulness of Rouen University Hospital's CDW and its query tool EDSaN in a situation of a worldwide pandemic. These tools have been successfully used to automatically retrieve patients for COVID-19 vaccination purposes.

Structured and unstructured real-world data are key to ensuring the best precision/recall performances: a formal evaluation will be performed soon.

Acknowledgements

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for her help in editing the manuscript.

References

[1] D.A. Hanauer, Q. Mei, J. Law, R. Khanna, K. Zheng, Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inform* **55** (2015), 290–300.

[2] S. Liu, Y. Wang, A. Wen, L. Wang, N. Hong, F. Shen et al., Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation (Preprint). *JMIR Medical Informatics* (2019)

[3] N. Garcelon, A. Neuraz, R. Salomon, H. Faour, V. Benoit, A. Delapalme et al., A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. *J Biomed Inform* **80** (2018), 52–63.

[4] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The Georges Pompidou University Hospital Clinical Data

Warehouse: A 8-years follow-up experience. *Int J Med Inf* **102** (2017), 21–8.
[5] S.M. Meystre, P.M. Heider, Y. Kim, D.B. Aruch, C.D. Britten, Automatic trial eligibility surveillance based on unstructured clinical data. *Int J Med Inf* **129** (2019), 13–9.
[6] J. Kang, J.H. Kim, K.H. Lee, W.S. Lee, H.W. Chang, J.S. Kim et al., Risk Factor Analysis of Extended Opioid Use after Coronary Artery Bypass Grafting: A Clinical Data Warehouse-Based Study. *Health Inform Res* **25(2)** (2019), 124.
[7] L. Grammatico-Guillon, K. Shea, S.R. Jafarzadeh, I. Camelo, Z. Maakaroun-Vermesse, M. Figueira et al., Antibiotic Prescribing in Outpatient Children: A Cohort From a Clinical Data Warehouse. *Clin Pediatr (Phila)* **58(6)** (2019), 681–90.
[8] P. Anaikatti, S.K. Sheth, A.M. Canlas, N.V. Shanbhag, M.L. Goh, H.C. Lim, Electronic medical record platform enhancements during COVID -19 to support IDENTIFY-ISOLATE-INFORM strategy for initial detection and management of patients. *Emerg Med Australas* **33(1)** (2021), 164–7.
[9] S.A. Deeds, S.L. Hagan, J.R. Geyer, C. Vanderwarker, M.W. Grandjean, A. Reddy et al., Leveraging an electronic health record note template to standardize screening and testing for COVID-19. *Healthcare* **8(3)** (2020), 100454.
[10] J.J. Reeves, H.M. Hollandsworth, F.J. Torriani, R. Taplitz, S. Abeles, M. Tai-Seale et al., Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *J Am Med Inform Assoc* **27(6)** (2020), 853–9.
[11] M. Pulia, D. Hekman, J. Glazer, C. Barclay-Buchanan, N. Kuehnel, J. Ross et al., Electronic Health Record-Based Surveillance for Community Transmitted COVID-19 in the Emergency Department. *West J Emerg Med* **21(4)** (2020), 748-751.
[12] The Dutch ICU Data Sharing Collaborators, L.M. Fleuren, D.P. de Bruin, M. Tonutti, R.C.A. Lalisang, P.W.G. Elbers. Large-scale ICU data sharing for global collaboration: the first 1633 critically ill COVID-19 patients in the Dutch Data Warehouse. *Intensive Care Med* **47(4)** (2021), 478-481.
[13] L'EDS mobilisé face à la COVID-19. 2020; Available from: https://eds.aphp.fr/covid-19
[14] B. Rosen, R. Waitzberg, A. Israeli, Israel's rapid rollout of vaccinations for COVID-19. *Isr J Health Policy Res* **10(6)** (2021).
[15] T. Pressat-Laffouilhère, P. Balayé, B. Dahamna, R. Lelong, K. Billey, S. J. Darmoni, J. Grosjean. Evaluation of Doc'EDS: A French Semantic Search Tool to Query Health Documents from A Clinical Data Warehouse. *Submitted and under second evaluation to BMC Medical Informatics and Decision Making* (2020).

Address for correspondence

Prof Stéfan J. Darmoni: stefan.darmoni@chu-rouen.fr