Evaluation of Doc'EDS: A French Semantic Search Tool to Query Health Documents from A Clinical Data Warehouse

Thibaut Pressat-Laffouilhère
Centre Hospitalier Universitaire de Rouen
Pierre Balayé
Centre Hospitalier Universitaire de Rouen
Badisse Dahamna
Centre Hospitalier Universitaire de Rouen
Romain Lelong
Centre Hospitalier Universitaire de Rouen
Kévin Billey
Universite de Rouen
Stéfan Jacques Darmoni
Centre Hospitalier Universitaire de Rouen
Julien Grosjean (julien.grosjean@chu-rouen.fr)
Centre Hospitalier Universitaire de Rouen https://orcid.org/0000-0002-7446-644X

Research article

Keywords: Clinical Data Warehouse, Cohort identification, Electronic Health Record, Information Retrieval, Semantics

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-59497/v1

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on February 8th, 2022. See the published version at https://doi.org/10.1186/s12911-022-01762-4.

Abstract

Background: Unstructured data from electronic health record is a gold mine. Doc'EDS is a pre-screening tool based on textual and semantic analysis. The system provides an easy-to-use interface to search documents in French. The aim of this study is to present the tools and to provide a formal evaluation of its semantic features.

Material & Methods: Doc'EDS is a search tool built on the top of the clinical data warehouse developed in the Rouen University Hospital. This tool is a multilevel search engine combining structured and unstructured data. It also provides basic analytics features and semantic utilities. A formal evaluation has been conducted to measure the implemented Natural Language Processing algorithms.

Results: About 17,3 million of narrative documents are contained in this CDW. The formal evaluation has been conducted over 5,000 clinical concepts that were manually collected. Negation concepts detection F-measure was 0.89, hypothesis concept detection F-measure was 0.57.

Conclusion: We hereby present Doc'EDS, a semantic search tool which deals with language subtleties to enhance an advanced full text search engine dedicated to French health documents. This tool is currently used on a daily basis to help researchers identifying patients thanks to unstructured data.

I. Introduction

In the last 20 years, the hospital data collection and storage has increased massively with the widespread use of clinical information systems (CIS). CISs contain large amounts of data on patients' health and healthcare: billing codes generating diagnosis related groups (DRG), medications, laboratory and imaging results, unstructured data, etc.

Unstructured data embedded in electronic health records (EHR) (mostly narrative reports) are necessary to solve patient inclusion in between 59% and 95% of criteria eligibility of clinical studies [1, 2]. Indeed, a wide range of crucial healthcare data is commonly found within unstructured clinical narratives. Narrative reports allow flexibility of expression such as doubts, negations, or diagnostic hypotheses and complex representation of diseases, clinical examination, patient history, and medical family history [3, 4].

Clinical Data Warehouses (CDWs) enable to register and aggregate different fragmented healthcare data from CISs and secondary data re-use. CDWs can enhance the quality of disease management [5, 6], and support clinical and translational research [7, 8, 9]. There are many CDW solutions such as I2B2 [10], STRIDE [11] and Vanderbilt [12] listed in a recent review [13].

Natural Language Processing (NLP) algorithms and information retrieval (IR) in EHR enable clinicians to identify cohorts [14]. In addition, automatic screening for trial eligibility criteria and extraction is currently being developed based on NLP [2, 15, 16].

Mining clinical narrative in full text is already done by EMERSE [17] and CREATE [18] in English and by Dr.Warehouse [19] and eHOP with Roogle [20] in French. However, there is a need to improve the semantic approach to deal with unstructured data: spelling errors, synonyms, abbreviations, acronyms, temporal notions, including all subtle information cited above. Only evaluations of user satisfaction and use cases are provided in the literature, but formal evaluations are lacking.

We present Doc'EDS (EDS = *Entrepôt de Données de Santé* in French, Health Data Warehouse in English), developed by the Department of BioMedical Informatics, Rouen University Hospital, France: a pre-screening tool to search patient profiles, identify cohorts in a document-oriented database. The aim of the Doc'EDS study was to provide a formal evaluation of search engine able of dealing with the unstructured clinical narratives in the Rouen University Hospital's CDW.

Ii. Methods

A. Database model & ETL

Doc'EDS relies on the CDW developed in the Rouen University Hospital since 2018. It contains patient data (birth date, gender) and its related clinical data (hospital stays, diagnoses, procedures, unstructured data documents and biological lab results) from 1998.

Doc'EDS is based on a document-oriented database where each document corresponds to unstructured data documents from the CDW. Related attributes are: patient ID, patient gender, patient birth date, document date, document type (e.g. discharge summary, procedure report), patient age (calculated from the document date and patient birth date), document production unit (e.g. cardiology, urology), DRG diagnoses & procedure codes (ICD-10 & CCAM classifications) and hospital stay ID. CCAM is a French medical code set that is used to report medical, surgical, and diagnostic procedures and services to entities such as physicians, health insurance companies and accreditation organizations. A specific ETL (Extract Transform Load) program has been developed to select unstructured data documents and their related attributes from the Rouen CDW via SQL queries.

B. Unstructured data analysis

Irrelevant content in documents

Headers and footers contain physician names, medical unit labels but also keywords corresponding to diseases or symptoms managed in medical units (e.g. Alzheimer, Parkinson, headache). This information is ignored in document indexing but remains visible for end users. In order to detect irrelevant content, a global frequency table was created from all documents with line and word position. Frequent words at a given line index are detected using an empiric threshold. Slight differences of line position and useful terms are considered. (e.g. *hypertension* which is a common clinical sign).

Segmentation

Unstructured narrative clinical reports can be segmented in a maximum of 19 different structured segments (based on consensus of four clinicians and based on the Rouen CDW documents): reason for admission (e.g. *dyspnea and fever*), medical history, allergies, usual treatments, anamneses, clinical examinations, laboratory and imaging results, disease evolution during hospitalisation, medical conclusion, treatments received during hospitalisation, prescribed treatments after hospitalisation, treatments, recommendations at discharge, procedure/technical procedure (e.g. *surgical procedure*), post-operative care, geriatric assessment, post-transplantation evolution and monitoring. Segmentation is processed using regular expressions which manages lexical variants.

Using segmentation in queries can be very helpful because it focuses on specific discourses (e.g. searching documents containing *fall* but only in reason of admission, that will exclude medical history of falls).

Special content tagging

Unstructured data from narrative clinical reports are tagged with three different types of “tags”:

- “Negation” corresponds to absence (e.g. *no complication*), negative results from complementary exams (e.g. *Koch research: negative*), reject of a diagnosis (e.g. *endocarditis was excluded*) and also negative sentences (e.g. *patient is not treated by …*).
- “Family medical history” corresponds to all clinical narratives related to one or few family members (or related). (e.g. *her father had a myocardial infarction*, *familial diabetes*)
- “Hypothesis/future” corresponds to events or facts that have not happened yet (e.g. future biology or treatment procedures not yet scheduled), that would happen, doubts, different hypotheses of patient disease, prevention (e.g. *HBPM prevention of thrombosis*).

These three tag types are detected using custom regular expressions based on common nouns/verbs/adverbs forms used in French. “Stop characters” are used to limit greedy expressions (e.g. dots, commas or specific words).

C. Deidentification

The ETL program has access to patient entity during the loading operation, it uses the patient names (last names & fore names) to remove their occurrences in corresponding texts. In addition, regular expressions are used to ensure that every name or identifier information (patient identity but also its related or health professional practitioners) are deidentified. Patient deidentification was evaluated during the ‘formal’ evaluation “tag and segmentation” detailed below.

D. Text search engine

The ETL program loads documents in a Lucene index: for text search, two distinct fields are created corresponding to the patient content and its family/related content. By default, the patient content equals to document text content minus tagged content and comments. Other text fields are generated on the fly by combining segments and “hypothesis” and “negation” sections; for example, it allows searching

negative content in specific segments (e.g. NOT *asthma* in anamneses which correspond to documents where anamneses contain a sentence which states that patient does not have asthma). Doc'EDS is implementing the Collector functionalities of Lucene to group and retrieve results as fast as possible. Each query returns a collection of document IDs and the corresponding number of patients (a single patient can be represented by multiple documents relative to the query).

E. Semantic support

Doc'EDS relies on the HeTOP crosslingual multi-terminology server [21] (URL: www.hetop.eu) which contains terminologies and ontologies in 45 languages (mainly in English and in French). Among the 85 termino-ontologies integrated into HeTOP, only 18 are also integrated in the UMLS [22]. The RUH DBI has partially translated into French several terminologies (such as NCIT, SNOMED CT, ICD-O, OMIM, Radlex, FMA, etc.). In September 2019, UMLS contains around 158,475 concepts sharing the same Concept Unique Identifier (CUI) with at least one translation in French whereas HeTOP contains 444,258 concepts with at least one translation in French. Terminologies exist only in French; e.g. CCAM for procedures or BNPC for chemical substances. HeTOP contains over 540,000 health concepts, with a HeTOP Unique Identifier (HUI, which is similar to the UMLS CUI including other non-UMLS terminologies). From these 540,000 HUI, the RUH DBI was able to create a French health lexicon (COMuF), similar to UMLS Specialist Lexicon.

F. Using Doc'EDS

Building queries

Keywords can be entered and enhanced with the following advanced options: **wildcards** (*) can be used to deal with variations in spelling (e.g. *pso** for *psoriasique, psoriasis…*); **double quotes** ("") can be used to search exact word sequences (e.g. *"allergie au paracetamol"*); **slope** (~) can be used to take in consideration distance between (non-stop) words (e.g. "*abcès anal"~2* can have a variation form as *abcès de la marge anale*); **Boolean** operators can be used to combine terms (e.g. *(paludisme OR "accès palustre" OR palu) AND quinine).* A specific module allows to search in one specific segment or to specify if negative or hypothetical clinical concepts are kept in search results or not. Keyword queries can be combined with structured data (age, sex, document type, medical unit, ICD-10 or CCAM codes).

Query semantic expansion

One of the main features of Doc'EDS is to provide easy and fast access to semantic expansion. The function allows to search synonyms and related terms to leverage the original simple query in order to expand the number of documents (lexical variations, acronyms, etc.). In addition, the number of documents with each term is provided on the fly. This expansion is provided by the HeTOP server that provides a web service to query the CoMUF lexicon. For each concept found, its synonyms, hyponyms and related terms are fetched then automatically used as a subquery to Doc'EDS. When a term matches at least one document, it is proposed in the module to the end user who can choose to modify, add or delete terms dynamically (see Fig. 1, with an example for "osteogenesis imperfecta").

G.Search Results

Document visualization

The end user can navigate through the documents retrieved by Doc'EDS. The document content is displayed on the right side of the screen and it corresponds to the transformed text. Words from queries are highlighted, irrelevant content is shaded, segments and tags are visible. This allows the reader to quickly evaluate relevance and to correct the query in consequences.

Automatic analysis

The number of documents and patients are displayed for each query. An advanced tool using structured data provides statistics as aggregated data (tables and charts) from all the retrieved documents: demographic data (pyramid of ages, male/female ratio), lists of ICD-10/CCAM codes, dates/types of documents, and medical units. This tool has two purposes: 1/ it can help the user to refine the query (e.g. exclude a specific medical unit from the query) and 2/ it can provide direct quantitative information (e.g. what is the median age of patients who had an appendectomy?).

In order to analyze text content related to a given query, Doc'EDS relies on the ECMT tool [23]. ECMT is an automatic semantic annotation program that identifies terminologies and ontology (from the HeTOP server) concepts in unstructured-texts. ECMT relies on the "bag-of-words" algorithm and also on pattern-matching designed for discharge summaries, procedure reports or laboratory results which contain symbolic data (presence or absence), numerical data.

Doc'EDS embeds ECMT to analyze corpora after performing a query as a text mining tool. Therefore, it is possible to identify frequent concepts in a specific corpus (e.g. most related diseases or most prescribed drugs).

H. Access rules and Security

Doc'EDS is only accessible in the Rouen University Hospital by DBMI team experts and developers. Moreover, even if documents are deidentified, each document is linked to a unique patient number in the CDW. Currently, this data warehouse is composed by two distinct databases separated on two different servers. Thus, nominatives elements are stored in one small encrypted database and the rest (clinical data) is stored on the other database. This type of architecture is compliant with GDPR application rules since it explicitly (physically) separates nominative data from deidentified data. Nevertheless, it is possible to re-identify patient numbers thanks to a complex decryption mechanism protected by a password only known by the DBMI experts.

I. Formal evaluation of tags and segmentation

The aim of this first formal evaluation was to compute: precision, recall, F-measure of each tag (negation, hypothesis, and family medical history), true positive (TP) percentage of segmentation functionality, tag occurrence among clinical concepts and documents with their 95% confidence interval. Random documents were drawn from hospitalisation reports, consultation and procedure reports until obtaining sufficient tags according to the estimated ratio of this document type based on structured data. Clinical concepts from documents were manually extracted and analysed by two public health residents as reference (TPL and PB). Tags were categorized in four modalities for each relevant clinical concept collected: Deleterious tags which correspond to an inappropriate tag with an impact on documents found by the query (FP: false positive) and deleterious missing tags (FN: false negative), non-deleterious tags which correspond to: appropriate tags (TP) and appropriate missing tags (TN: true negative). In order to obtain the percentage of well segmented concepts and because segmentation involves 19 different categories based on the “same rules” of detection, segmentation evaluation was common to all segmentation categories. False positive segment “A” could be a false negative segment “B” hence recall and precision are not computable. Segmentation evaluation was based on hospitalisation reports. Expert concordance (Kappa) for each tag and segmentation was computed.

J. Doc’EDS complementarity with PMSI: use cases

In hospital, patient retrieval is usually performed with PMSI (Programme de médicalisation des systèmes d’information) which are the French healthcare claims data (DRG). Patients are identified with ICM-10 and/or CCAM. Queries are limited by the lack of codes (e.g, new practices or rare disease), the use of an inappropriate code [24, 25], absence of code coming from no financial valuation (e.g. medical history), or even code evolution [26]. With a CDW, retrieval is potentially wider. Improvement retrieval against DRG was illustrated by two use cases, demonstrating Doc’EDS added-value.

Iii. Results

A. Doc’EDS

Doc’EDS is a web application, written in Java EE and running on a Tomcat web container. It relies on a Lucene index and includes several additional tools that help to visualize, analyse and export results. The ETL program uses the Rouen CDW SQL database to automatically feed Doc’EDS each week from the Clinical Information System. In July 2020, data volumes are: 2 million patients, 17,682,236 narrative documents for about 15 million consultations/hospital stays. Doc’EDS is been used in routine in the Rouen University Hospital. So far, it helped to respond to 103 various use cases in different specialties. A screenshot of the main query/visualize window is presented in Figure 2. The analysis panel is very useful to refine queries or to collect important information; the Figure 3 is an example of the age distribution for the query *“psychomotor regression” AND epilepsy*.

B. Tag and segmentation evaluation

A total of 5,277 concepts were collected by the two residents (PB and TPL) among 54 hospitalisation reports (3,767 concepts) and 93 procedure reports or consultations. There were 35.9 (mean) (sd = 38.8) concepts per document. Negative concepts represented 11.7% [11%;12.6%], 4.6% [4.1;5.2] were hypotheses, 0.3% [0.2;0.5] were medical family history. A total of 2,000 concepts were evaluated by both TPL and PB (1,737 for segmentation part). Disagreement did not exceed 3.2% see Table 1. Confrontation results in no disagreement, most of them coming from human misunderstood. None deidentification violation was found among the 147 documents (0% [0;2.5]). Concerning negation tag and hypothesis tag F measure was respectively 0.89 and 0.57, (see Table 2). Concerning segmentation (evaluation among 3,767 concepts) 84% IC95% [83%;85%] concepts were well segmented.

Table 1

	Disagreement	**Kappa**
Negation tag	2.8%IC95%[2.1;3.7]	0.88[0.84;0.91]
Hypothesis tag	2.6%IC95%[1.9;3.4]	0.70[0.62;0.77]
Segmentation	3.2%IC95%[2.5;4.2]	0.87[0.83;0.90]
Concordance results between the two residents		

Table 2

Negative Concepts	**Resident +**	**Resident -**	
Doc'EDS +	TP = 551	FP = 60	Precision = 0.90[0.87 ;0.92]
Doc'EDS -	FN = 68	TN = 4,598	NPV = 0.98
	Recall = 0.89[0.86 ;0.91]	Specificity = 0.98	F = 0.89

Hypothesis Concepts	**Resident +**	**Resident -**	
Doc'EDS +	TP = 116	FP = 41	Precision = 0.73[0.66;0.80]
Doc'EDS -	FN = 128	TN = 4,992	NPV = 0.98
	Recall = 0.47[0.41;0.54]	Specificity = 0.98	F = 0.57
Negative and hypothesis tags evaluation Resident versus Doc'EDS			

C. Use cases

Among the 103 use cases processed between January 2019 and July 2020, we chose to focus on two use cases to illustrate the complementarity between Doc'EDS and the French DRG.

Transcatheter Aortic Valve Replacement (TAVR)
TAVR is a relatively new procedure, so the corresponding codes were implemented between 2005 and 2009 in the CCAM (French procedure coding system). Doc'EDS enabled practitioners to retrieve endocarditis associated to TAVR procedure that DRG alone could not. Doc'EDS found 23/53 patients that DRG did not found with 2 false positives (undetected hypotheses) and 2 false negatives (documents did not mention TAVR explicitly).

Patients diagnosed with Fahr disease

Fahr disease or IBGC (Idiopathic Basal Ganglia Calcification) is coded with ICD-10 code G23.8 (Other specified degenerative diseases of basal ganglia). This code regroups different diseases, hence it is not specific. Doc'EDS enabled practitioner to refine research on a restrained sample. The initial ICD-10 query obtained 392 patients with G23.8 and Doc'EDS found 93 patients (3 false positives). This has prevented the researchers to read all 392 patient records to confirm the Fahr disease or not.

Iv. Discussion

Doc'EDS was created in order to obtain an easy to use interface with a multilevel search engine combining structured data, clinical narratives and segmentations. Clinical narratives processing is based on semantic approach to deal with language subtleties and enable query expansion. Doc'EDS provides basic analytics, including descriptive statistics to refine query or estimate study feasibility. The easy to use interface enables to build quick successive queries without any background in information retrieval or computer science.

A. Access Policy

Doc'EDS is not open to the medical community. Unlike most tools based on CDWs, we believe that building queries into such complex databases with subtle algorithms is an expert task. Each use case is a dialog between the researchers and the expert team; it often includes the assistance from a statistician methodologist to ensure that the research question can be met with the expected data extracted from Doc'EDS.

Since the data in Doc'EDS is deidentified the researchers can consult medical narratives to ensure if patients can be included in their studies or not. A selection can be exported and re-identified if necessary,

via a dedicated program, only accessible by the DBMI team.

B. Semantic features

Doc'EDS provides a set of semantic functions that enable query enhancement (synonyms and related terms additions) and concept analysis to get an overview of relevant concepts for a given corpus. As Doc'EDS is based on the HeTOP multi-terminogy server, it contains the largest French medical lexicon. This is very helpful to build queries dealing with complex keywords (e.g. with multiple synonyms).

C. Clinical narratives processing

Contrary to EMERSE or IT solution from Leon Berard center [27], Doc'EDS (as Dr Warehouse or as CREATE) handles negation and family medical history to deal with such subtleties. According to the tools cited above, Doc'EDS is the only one who provides hypothesis tag in French medical documents, which represents 4.6% of concepts in our evaluation. As far as we know precedent tools did not estimate negative (11.7%), or medical family history (< 0.5%) concepts prevalence in French medical documents.

Segmentation provided by Doc'EDS shows that over 80% of clinical concepts are well segmented in hospitalisation reports. Even in hospitalisation reports, segmentation was not possible because lack of key words used by clinicians or line break leading to no segmented concepts. Misplaced or wrong keywords trigger inappropriate segmentation. The 19 different segments were not evaluated or presented in others articles (just barely mentioned).

A very few amounts of family medical history tags were collected; indeed, it depends on medical speciality such as cancer, paediatrics or genetics where it should be more prevalent. Even if there are only 147 documents, external validity was high. Indeed, there are numerous and various types of clinical concepts such as clinical examination, surgical procedure, treatment, lab result, etc. and wide range of expressions from different practitioners. High F-measure was registered for negation tag (0.89) whereas hypothesis tag obtained a low F-measure (0.57). Most of the time, hypothesis tag weakness coming from false negative clinical concepts corresponding to order exit (e.g. patient will have radiologic exam). Thus, it does not impact precision of the document retrieval.

V. Conclusion

Doc'EDS is a search tool used on a daily basis at the Rouen University Hospital in order to create cohorts for research. This tool exploits a massive corpus of more than 17,3 million documents and relies on semantics features to build sophisticated queries. Many other functions are proposed to analyze results or to refine build queries (e.g. basic statistics on documents data distributions, concept automatic extraction).

Subtle semantic processing includes negation detection, hypotheses/future detection and family history detection. This system, based on regular expressions, is continuously updated to enhance performances since the evaluation showed good results for negation and more mixed results for hypotheses/future content.

One another important aspect of Doc'EDS is its performances; most queries are executed in less than two seconds. This is very useful when it could take many tests to get a "good" query.

From January 2019 to July 2020, Doc'EDS has been used in 103 various use cases to identify patient cohorts for clinical research; 28 different medical services (e.g. cardiology, dermatology, pediatrics…) have required the help of Doc'EDS so far.

Vi. Perspectives

Doc'EDS will be re-evaluated on the 5,277 clinical concepts manually extracted to increase negation and hypothesis tag precision after corrections. The amount of unstructured data will grow (nurse transmissions, virology/microbiology reports…). Automatic extraction is being developed around use cases to facilitate data collection for clinical research. A specific module will be developed to help researchers to build cohorts from Doc'EDS results. Finally, this tool could also be used to assist technicians to produce code from ICD-10 in the Rouen University Hospital.

Abbreviations

DBMI: Department of BioMedical Informatics

CDW: Clinical DataWarehouse

DRG: Diagnosis Related Groups

Declarations

Ethics approval and consent to participate

The French Data Protection Authority (CNIL) approved the construction and the usage of the Rouen University Hospital Clinical DataWarehouse, based on a declaration compatible with the General Data Protection Regulation applicable in France. As collecting consents on more than 2 million people is not possible, a global public information has been made and individual information is provided for each new patient in the hospital.

Consent for publication

This publication does not contain personal data and the Rouen University Hospital Clinical DataWarehouse only stores deidentified data.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due the GDPR application in France about individual's data.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially granted by the European FEDER funds and by the Région Normandie funds involved in the PlaIR2.018 research project.

Author's contribution

All authors made substantial contributions to the conception of the work and interpretation of data. BD, RL, KB and JG were responsible for most development work. TPL and PB performed the formal evaluation. TPL and JG drafted the initial manuscript. TPL, SJD, BD and JG drafted the work. All authors revised the manuscript critically. All authors gave final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgments

The authors warmly thank Nikki Sabourin, medical editor, who proofread the manuscript.

References

1. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? :6.
2. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. International Journal of Medical Informatics. 2019 Sep;129:13–9.
3. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data

warehouse. J Am Med Inform Assoc. 2016 Oct 20;ocw144.

4. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association. 2011 Mar 1;18(2):181–6.
5. Karami M, Rahimi A, Shahmirzadi AH. Clinical Data Warehouse: An Effective Tool to Create Intelligence in Disease Management. The Health Care Manager. 2017;36(4):380–4.
6. Plantier M, Havet N, Durand T, Caquot N, Amaz C, Biron P, et al. Does adoption of electronic health records improve the quality of care management in France? Results from the French e-SI (PREPS-SIPS) study. International Journal of Medical Informatics. 2017 Jun;102:156–65.
7. Grammatico-Guillon L, Shea K, Jafarzadeh SR, Camelo I, Maakaroun-Vermesse Z, Figueira M, et al. Antibiotic Prescribing in Outpatient Children: A Cohort From a Clinical Data Warehouse. Clin Pediatr (Phila). 2019 Jun;58(6):681–90.
8. Kang J, Kim JH, Lee KH, Lee WS, Chang HW, Kim JS, et al. Risk Factor Analysis of Extended Opioid Use after Coronary Artery Bypass Grafting: A Clinical Data Warehouse-Based Study. Healthc Inform Res. 2019;25(2):124.
9. Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. International Journal of Medical Informatics. 2017 Jun;102:21–8.
10. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006;1040.
11. Lowe HJ, Ferris TA, Nd PMH, Weber SC. STRIDE - An Integrated Standards-Based Translational Research Informatics Platform.:5.
12. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: The Vanderbilt approach. Journal of Biomedical Informatics. 2014 Dec;52:28–35.
13. Khalaf Hamoud A, Salah Hashim A, Akeel Awadh W. CLINICAL DATA WAREHOUSE A REVIEW. Ijci. 2018 Dec 31 [cited 2019 Jul 2];44(2).
14. Vydiswaran VGV, Strayhorn A, Zhao X, Robinson P, Agarwal M, Bagazinski E, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. Journal of the American Medical Informatics Association. 2019 Jun 14;ocz079.
15. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. International Journal of Medical Informatics. 2019 Sep;129:13–19.
16. Zhou X, Wang Y, Sohn S, Therneau TM, Liu H, Knopman DS. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. International Journal of Medical Informatics. 2019 Oct;130:103943.
17. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. 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). Journal of Biomedical Informatics. 2015 Jun;55:290–300.

18. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. CREATE: Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records using OMOP Common Data Model. :14.
19. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. Journal of Biomedical Informatics. 2018 Apr;80:52–63.
20. 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. Studies in Health Technology and Informatics. 2011;584–588.
21. Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Teaching medicine with a terminology/ontology portal. Stud Health Technol Inform, 2012;180:949-53.
22. A.B. Lindberg, B.L. Humphreys, A.T. McCray. The unified medical language system. Methods Inf Med 32 (1993), 281–91.
23. Cabot C, Soualmia LF, Grosjean J, Griffon N, Darmoni SJ. Evaluation of the Terminology Coverage in the French Corpus LiSSa. Stud Health Technol Inform, 2017;235:126-130.
24. De Léotoing L, Barbier F, Dinh A, Breilh D, Chaize G, Vainchtock A, et al. French hospital discharge database (PMSI) and bacterial resistance: Is coding adapted to hospital epidemiology? Médecine et Maladies Infectieuses. 2018 Oct;48(7):465–73.
25. Perozziello A, Gauss T, Diop A, Frank-Soltysiak M, Rufat P, Raux M, et al. La codification PMSI identifie mal les traumatismes graves. Revue d'Épidémiologie et de Santé Publique. 2018 Feb;66(1):43–52.
26. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors: Medical Care. 2005 May;43(5):480–5.
27. Biron P, Metzger MH, Pezet C, Sebban C, Barthuet E, Durand T. An information retrieval system for computerized patient records in the context of a daily hospital practice: the example of the Léon Bérard Cancer Center (France). Appl Clin Inform. 2014;05(01):191–205.

Figures

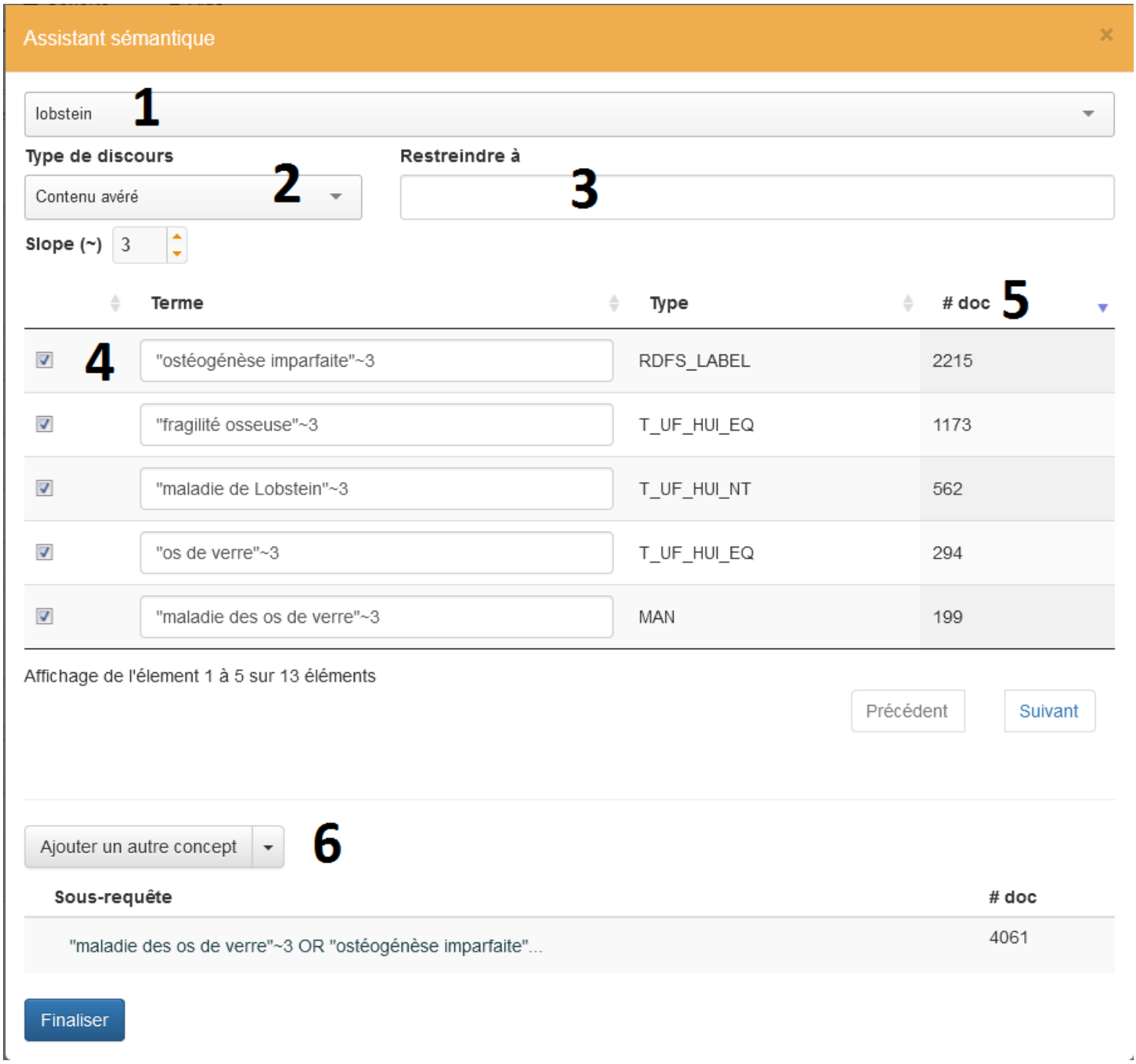

Figure 1

Capture of the semantic assistant that helps the user to enhance its query. 1) Type words to search into HeTOP and select a concept, 2) Select part of speech including negations, hypothetical parts or family history, 3) Select segment(s) to query, 4) Choose relevant terms (which are synonyms or hyponyms extracted from HeTOP) to add to a query, 5) Check at the number of documents found in real time for each proposed term and 6) If needed, repeat the operation in adding a new concept (with OR, AND or NOT operators).

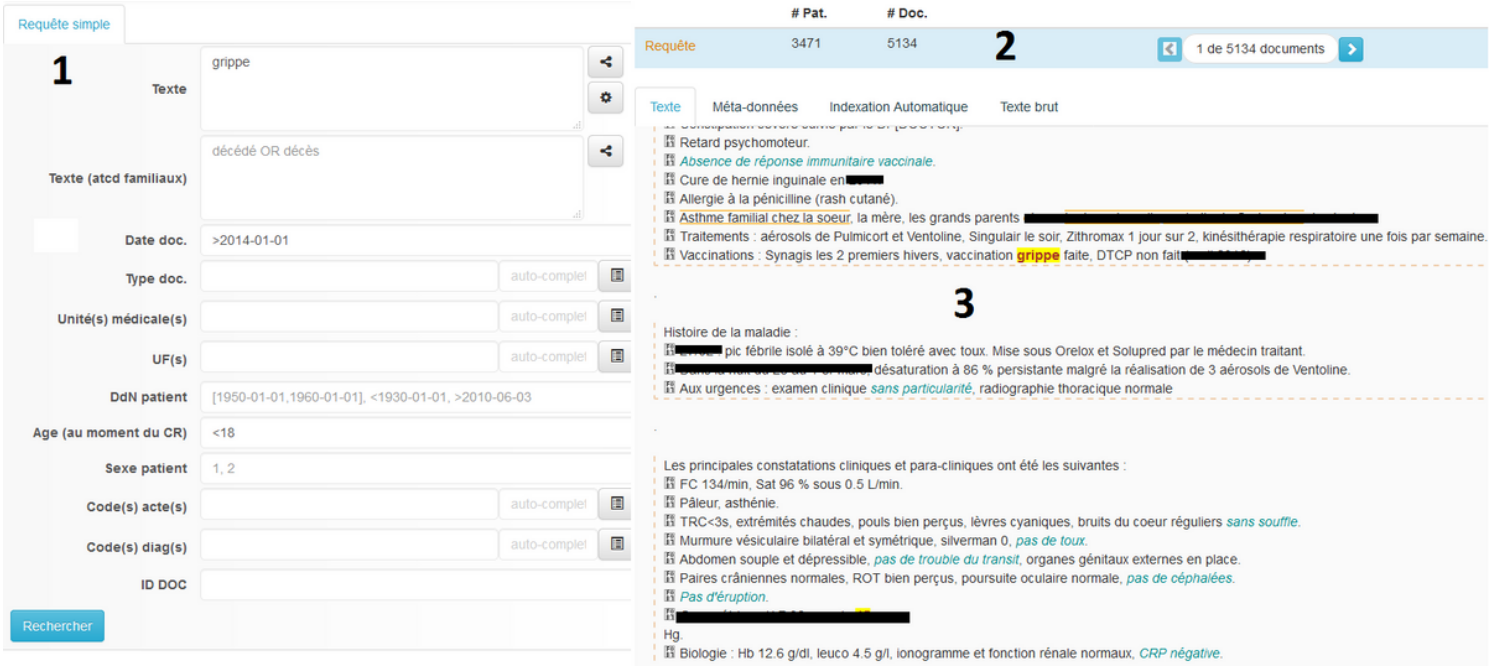

Figure 2

Screenshot of Doc'EDS main page: 1) the query form is on the left side. In addition to keywords, different fields can be used e.g. document date, type, patient age and sex… 2) the number of patients and documents retrieved are displayed, 3) A visualization screen allows users to consult documents (in order to refine queries or collect specific data). In this example, some portions of text (dates) have been blanked to preserve patient identity.

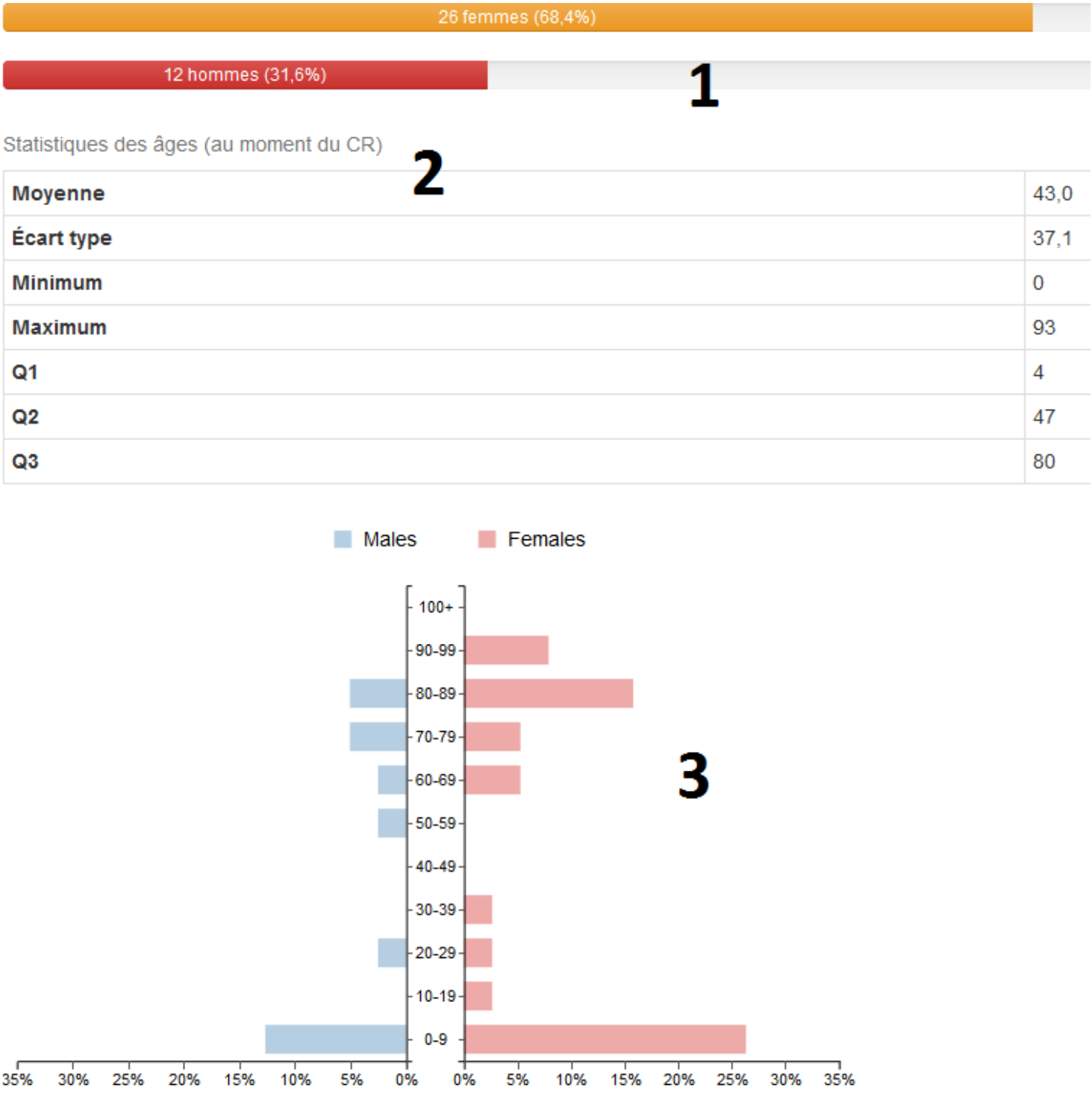

Figure 3

Basic demography statistics for the query (in French) “psychomotor regression” AND epilepsy. 1) Sex distribution, 2) Age basic statistics and 3) Age distribution are calculated