From doc2vec to advanced keyword queries: searching for phenotypes in large clinical document databases

Christel Gérardin[a], Yuhan Xiong[a,b], Fabrice Carrat[a,c], Xavier Tannier[d]

[a]Sorbonne Université, Inserm, Institut Pierre-Louis d'Epidémiologie et de Santé Publique, Paris, France F75012.

[b]Shanghai Jiaotong University, 800 Dongchuan RD. Minhang District, Shanghai, China

[c]Département de Santé Publique, Hôpital Saint-Antoine, Paris, France F75012.

[d]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, Paris, 75006, France

*Corresponding Author :
Dr. Christel Gérardin,
Sorbonne Université, Inserm,
Institut Pierre-Louis d'Epidémiologie et de Santé Publique,
Paris, France F75012.
christel.ducroz-gerardin@iplesp.upmc.fr

Abstract

Background: The latest deep learning algorithms have significantly improved natural language processing tasks, including in the medical domain, by directly extracting patient information from clinical notes. However, these algorithms come with a high computational cost and are often not applicable at the scale of very large databases in the temporality of clinical practice.

Objective: The objective of our study is the automatic detection of clinical documents of interest for a specific clinical question, with low computational cost, to be applied on a database of millions of documents. These sets of documents of interest constitute a pre-screening to allow the development of more complex algorithms.

Method: The task was considered as an information retrieval task in French clinical texts. Two different methods were compared. For the first method, we used several state-of-the-art vector representations: TF-IDF, doc2vec, docBERT and tested if the closest documents are relevant. The second method consists in building a powerful query expansion from an entered key term, its French synonyms from UMLS and synonyms found by similarity with the embeddings of the CODER algorithm. These methods are developed and evaluated on a set of 8 and 20 phenotypes respectively. Our database corresponds to 2 million documents from a cohort of patients with four autoimmune diseases: systemic lupus erythematosus, scleroderma, antiphospholipid syndrome and Takayasu disease, from the AP-HP data warehouse.

Results: Our experience does not support the vector representation model of clinical notes for searching similar patients. However, searching with an advanced synonym search method can lead to very good results without additional burden for the clinician: we achieved a precision of 0.93 [0.90; 0.96] and a recall of 0.78 [0.71; 0.85] evaluated on the basis of the ICD-10 codes of the retrieved patients, in a very reasonable time.

keywords: document retrieval, medical informatics, clinical phenotypes

INTRODUCTION

Background

The emergence of large health databases has provided access to a large volume of medical information. Nevertheless, natural language processing algorithms for the analysis of unstructured text, used in particular for prognostic evaluation [1, 2], diagnostic prediction [3], automated selection of patient cohorts [4, 5], drug analysis [6] or therapeutic decision support [7], are now regularly performed by deep neural models, such as transformers [8], which offer high performance but come with significant computational and environmental costs. Reliable pre-selection of specific clinical documents is therefore essential, in order to reduce the total number of documents to be processed from several million to a few thousand, or even less.

Automatic retrieval of documents of interest is a major current concern, particularly in the biomedical field, from searching for articles on Pubmed [9, 10] to searching for clinical documents [11, 12], and remains a challenge in this area due to the extensive use of acronyms, abbreviations and complex, ambiguous terms. The use of vector representation models such as doc2vec [13] and docBERT [14] for document retrieval has been used in the legal and commercial [15, 16, 17] and biomedical fields [9, 18].

In this study, we used the UMLSⓇ (for Unified Medical Language System) whose Metathesaurus encompasses over 30 vocabularies from many different languages, including French. All UMLS medical concepts have a unique concept identifier, CUI, shared with synonyms and sometimes abbreviations.

Finally, biomedical normalization algorithms, which link any medical concept to a standard concept in a thesaurus such as the UMLSⓇ for example, have made great strides in providing an efficient representation of concepts, enabling them to be used directly for information retrieval[19, 20, 21].

Goal of this study

The task addressed in this work is the automatic identification of patients with a predefined phenotype. This clinical task is approached by two different information retrieval methods: on the one hand, a document similarity search based on an exemplary case (or index case) and, on the other hand, a keyword query. The first, document-oriented method is based on the vector representation of documents: all documents whose vector representation is close to that of an initial index document (i.e. documents mentioning the same diseases or symptoms) are selected for the search. Three models were tested: TF-IDF, doc2vec [13] and docBERT [14]. We compared document-based methods with another approach, namely a keyword-based approach based on an advanced keyword query: the clinician enters a diagnosis of interest and a text search is performed using these keywords, their synonyms in UMLS and related terms found by similarity in CODER embeddings [21]. Figure 1 gives an overview of the two methods compared.

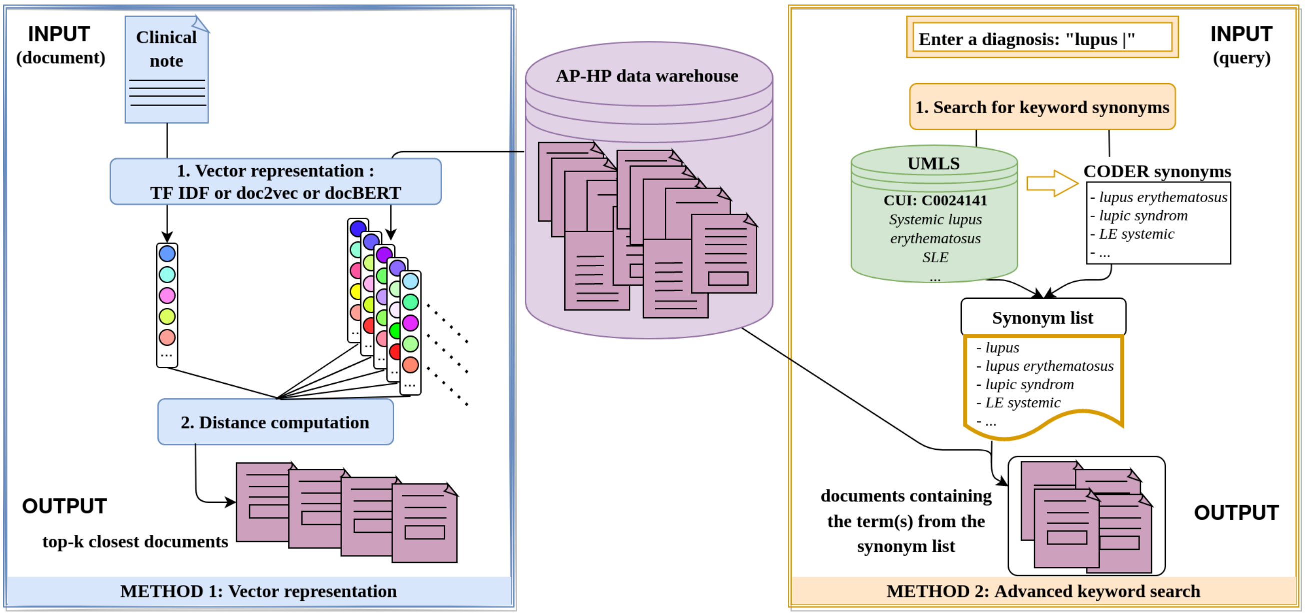

Figure 1: Overview of the two patient pre-selection methods in our data warehouse. The clinical task is the same, but method 1 is document-oriented, while method 2 is keyword-oriented. The step-by-step description of each method using the example of the "Systemic Lupus Erythematosus" phenotype as illustrated in the figure is as follows. In method 1, the clinician provides a typical clinical observation of a lupus patient (describing symptoms, organ damage, typical lupus treatment). A vector representation of this clinical note is then created using 3 different methods: TF-IDF, doc2vec[13] or docBERT[14]. In parallel, a vector representation of all other clinical notes is produced. Finally, a cosine distance similarity calculation is performed between the vector representing the initial lupus patient's clinical score and all the vectors representing the other clinical scores. The closest vectors should correspond to the documents of patients with the same pathology. In the second method, the clinician enters not one observation but one (or two) keywords for his phenotype of interest, and a query expansion method automatically calculates synonyms for this term in the UMLS by direct matching and using the CODER[21] model.

This pre-screening or pre-selection of patients can be useful in many applications, notably in epidemiology to carry out a feasibility study before setting up a study, but also in pharmaceutical studies to target all patients potentially eligible for a treatment. Similarly, pre-selection can be used to design a preventive campaign (screening, for example), offering a complete medico-economic approach to anticipate costs.

MATERIAL AND METHODS

1. Dataset:

For this experiment, we used an extraction from the AP-HP datawarehouse (after IRB agreement, decision number 20-93) concerning all patients aged 15 or over with systemic lupus erythematosus (SLE), antiphospholipid syndrome, systemic scleroderma and Takayasu disease - based on ICD-10 codes - and who were seen at least once between 01/07/2017 and 31/12/2020 in one of the 39 hospitals of the Assistance Publique Hôpitaux de Paris (AP-HP). This cohort corresponds to around 38,000 patients and 2 million documents, including around 300,000 discharge summaries, with an average word count of 1,500 words per summary. The detailed number of patients per pathology and the ICD-10 codes used as inclusion criteria are provided in Supplementary Table 1.

In France, ICD-10 codes (corresponding to the 10th International Classification of Diseases) are used for the financial valuation of hospital stays, and are decoupled into main diagnosis (disease associated with the reason for hospitalization), related diagnosis and associated diagnosis for each stay. Only main and active pathologies are associated with the stay.

For the document vector representation assessment, a physician annotated 256 hospital reports with four phenotypes "systemic lupus erythematosus (SLE) nephritis", "osteoporosis," "pulmonary infection," and "scleroderma interstitial lung disease." This annotation is performed at the document level: if one of these four phenotypes is positively mentioned in the document, the document is annotated with the corresponding phenotype. Among 256 randomly selected documents, 23 were annotated with "osteoporosis," 48 with "nephritis in SLE," 20 with "interstitial lung disease in scleroderma," and 33 with "pulmonary infection."

Given the large volume of documents on which the query is performed (around 2,000,000 reports in total, of which 20% are hospital reports), it is not humanly reasonable to read and classify all the documents according to the phenotypes of interest, so obtaining an overall gold standard (manual annotation with a high degree of precision) is not an option. Two silver standards were therefore used: automatic annotation with a lower degree of precision, and ICD-10 coding. The automatic annotation corresponds to the presence of a list of keywords in the texts, with a list of manually preset terms for all possible synonyms attached to the phenotypes of interest (for example, for the phenotype "lupus nephropathy", the following terms were proposed: "glomerulonephritis", "LGM"-respecting the capital letter, "GEM", "nephritis" etc., associated with the term "lupus", "lupus" or "LES" in the text). This list of pre-established terms was drawn up by an internist clinician from outside the study.

2. Vector document representation

Three main algorithms were tested for representing clinical notes and finding nearby documents.

A TF-IDF (Term Frequency-Inverse Document Frequency) method that assigns numerical scores to terms in a document based on their frequency in the document and their relative importance in the set of documents. It particularly captures relevant words and is of major interest for our autoimmune disease database. We have used sklearn's TfIdfVectorizer [22].

A doc2vec model [13] was also trained on our global dataset. This document representation technique captures the semantic relationships between documents. It is based on the word2vec model [23], which transforms words into numerical vectors in such a way that similar words have similar vectors. In the doc2vec model, a document vector is trained in parallel with the training of word vectors. Basically, this representation should reflect the "subject" of the document, and two documents with the same subject should be close together, while two documents with different subjects should be far apart. We used the doc2vec model from the gensim library [24] with a dimension of 300. Two versions of the model were used: DBOW (Distributed Bag of Words) where the model considers the whole document as a "bag of words" and DM (Distributed Memory) which takes into account the context and order of words in the document.

Similarly, a docBERT model [14] derived from a camemBERT-large model [25, 26] was also used as a document representation. This model is an extension of the BERT model, based on the Transformers architecture [8], which has been trained to learn contextual representations of words and phrases. The docBERT Model [14] combines BERT representations of text segments to represent the document.

3. Advanced keywords search:

Another proposed method was to directly enter the desired diagnosis(es) by keyword (e.g. "rheumatoid arthritis" or "acute renal failure", etc.). The clinician enters the term or list of terms directly and, for each item in the list, an automatic search for synonyms is performed. First, all terms are mapped to their unique UMLS concept identifier, which allows finding certain abbreviations, acronyms and synonyms of terms. Then, on this last set of terms, the CODER algorithm [21] is used to find other synonyms, i.e. terms with close CODER embeddings.

CODER (Cross-lingual knowledge-infused medical term embedding) is an algorithm trained by contrastive learning on UMLS, based on a BERT model allowing term normalization (i.e. linking a term to its specific UMLS concept). The underlying principle is that UMLS synonyms have a close cosine similarity between their respective embeddings. We used the CODER_all version specifically on the French UMLS vocabularies and obtained the synonyms of the terms. An example of a query and its synonyms is presented in Table 1, which allows us to obtain a significantly larger list and especially to access the acronyms of the terms. Once the list of synonyms was found, we performed a direct search of the original terms and synonyms on the documents.

Several cosine similarity thresholds on our training dataset were tested, performing a grid search to find the threshold that maximizes recall while preserving precision.

Query	“Purpura thrombopénique idiopathique”	“Idiopathic thrombocytopenic purpura”
Same UMLS CUI terms	PTI, purpura thrombocytopénique auto immun, maladie de werlhof, purpura thrombopénique immunologique, purpura thrombopénique idiopathique, purpura thrombocytopénique autoimmun,	ITP, autoimmune thrombocytopenic purpura, werlhof's disease, immunological thrombocytopenic purpura, idiopathic thrombocytopenic purpura, autoimmune thrombocytopenic purpura, autoimmune thrombocytopenic purpura, idiopathic thrombocytopenic purpura of

	purpura thrombopénique autoimmun, purpura thrombopénique idiopathique de werlhof, PTAI, syndrome de werlhof, purpura thrombopénique auto immun, purpura idiopathique, purpura thrombocytopénique idiopathique	werlhof', AITP, werlhof syndrome, autoimmune thrombocytopenic purpura, idiopathic purpura, idiopathic thrombocytopenic purpura
CODER Synonyms (limit 0.75)	purpura thrombocytopénique idiopathique, purpura thrombopénique idiopathique, purpura thrombopénique idiopathique de werlhof, maladie de werlhof, purpura thrombopénique à médiation immunitaire, purpura thrombocytopénique autoimmun, thrombocytopénie idiopathique, purpura hyperglobulinémique, purpura thrombopénique auto immun , purpura thrombocytopénique auto immun, purpura idiopathique, syndrome de werlhof, purpura thrombopénique immunologique thrombocytose idiopathique, purpura thrombopénique autoimmun	idiopathic thrombocytopenic purpura, idiopathic thrombocytopenic purpura of werlhof, disease of werlhof, immune-mediated thrombocytopenic purpura, autoimmune thrombocytopenic purpura, idiopathic thrombocytopenia, hyperglobulinemic purpura, autoimmune thrombocytopenic purpura , autoimmune thrombocytopenic purpura, idiopathic purpura, werlhof syndrome, immunological thrombocytopenic purpura idiopathic thrombocytosis,

Table 1 : Example of advanced keyword search with UMLS and CODER synonyms (in French and corresponding terms in English).

4. Evaluation

4.1. Vector document representation

Two evaluations were performed. The first evaluation concerned all 256 annotated documents with a phenotype directly present in the text. For these documents, we computed their vector representations by the three methods TF-IDF, doc2vec and docBERT. In parallel, we computed the vector representation of all the documents in the database (restricted to 10,000 randomly selected documents for computation time reasons). Finally, for the three methods, we computed the cosine distance of the 256 documents with the 10,000 other documents to find for each document which are the k closest.

On a un ensemble de "documents-requetes", un ensemble de "documents-recherche". Pour une requête, on calcule les représentation vectorielles de la requête et des documents, et on fait la similarité cosinus et on récupère les top K.

Chaque requête a un phénotype et une liste de synonymes, et les "bons" documents, les ground truth, seront les documents contenant au moins un des synonymes (phenotype inclu). On calcule ensuite l'accuracy.

For each phenotype of the 256 documents (Nephritis in SLE, etc...) we had a list of terms manually defined by a clinician as exhaustive as possible of synonyms of this particular phenotype, the list of documents containing these terms corresponding to our gold standard. Supplementary Table 1 summarizes the list of manual terms entered. We then calculate the accuracy of the top-k documents by checking whether we find any of these terms in the text documents. Among the 10,000 documents, the manual list method extracted 664 patients with nephritis in SLE, 731 with osteoporosis, 943 with lung infection, and 886 with ILD in scleroderma.

For the second assessment, we took a random sample of 100 documents from the overall database and considered ICD-10 coding as a silver standard. For these 100 random documents, we calculated whether the k closest documents had the same ICD-10 coding for the principal diagnosis (main disease), for at least one associated diagnosis (comorbidities), and for at least one principal or associated diagnosis with the sampled document. We calculated the average accuracy with a 95% confidence interval on the 100 documents in the sample. It is important to note that ICD-10 coding is not done at the document level but at the patient level (i.e. the aggregation of the coding of all the documents of the same patient). For example, if a patient comes for an endoscopy, or a specialist consultation, the corresponding report will not mention verbatim all the patient's previous diseases (e.g. "osteoporosis") but the coding will still contain the corresponding code. We therefore also considered a "patient-level" precision, by computing how many patients in the top k candidates had at least one ICD-10 coding in common with the index patient. In this experiment, we considered "coarse-grained" ICD-10 codes, i.e., one letter, two digits.

4.2. Advanced keywords search evaluation:

To evaluate the performances of our advanced keyword search with few terms entered by a clinician enhanced with an automated list of synonyms, we used a development set separate from the evaluation set to choose the best configuration (for synonym expansion and similarity threshold) *via* grid search.

On this development dataset, accuracy of the extracted documents with respect to the query was computed against a goal standard consisting of a manual list of terms, defined in advance by a clinician for 8 arbitrary phenotypes. Those phenotypes were *Nephritis in SLE, osteoporosis, lung infection, ILD in scleroderma, pulmonary hypertension, gastroesophageal reflux, gougerot sjogren's syndrome, pulmonary embolism*. The list of terms was the same as for the vector representation evaluation for the 4 first phenotypes. In our entire database, we had over 250,000 discharge summaries. Table 2 shows the number of documents extracted for each phenotype by the manual lists of terms.

Phenotype	**Number of documents with automatic annotation silver standard**
Nephritis in SLE	19248
Osteoporosis	20897
Lung infection	18355
ILD in scleroderma	23511
Pulmonary hypertension	28514
Gastroesophageal reflux	20648
Gougerot Sjogren syndrome	16420
Pulmonary embolism	39523

Table 2: Number of documents with automatic annotation silver standard for the 8 training phenotypes.

For these training phenotypes, a list of corresponding ICD-10 codes was also prepared in advance. For instance, pulmonary embolism is “I26”. The recall was defined as the number of patients found by our advanced keyword document search method versus the number of patients to be found with the corresponding ICD-10 of the phenotype.

We also had a set of 20 queries as a test set to evaluate our method. These arbitrarily chosen queries were “*rheumatoid arthritis”, “Takayasu disease”, “pericarditis in Lupus”, “acute myocardial infarction”, “antiphospholipid syndrome”, “kidney transplantation”, “prostate cancer”, “lung tuberculosis”, “autoimmune hepatitis”, “dermatomyositis”, “idiopathic thrombocytopenic purpura”, “acute kidney injury”, “Raynaud syndrome”, “pyelonephritis”, “HIV”, “scleroderma”, “diabetes” (type 1 or 2), “stroke”, “stroke in lupus”, “spontaneous miscarriage”.* These queries are purposely not very specific, in order to reflect the clinical situation, where a clinician does not always know all the synonyms of interest.

For those test queries, a clinician assessed a random set of 50 documents, verifying if the document mentioned the disease or not. This additional verification allowed us to evaluate directly the accuracy at document scale. The recall is computed as previously with the ICD-10 coding of the patients. For more complex queries such as “*stroke in lupus*” or “*pericarditis in lupus*”, we considered the patients with both “stroke” and “lupus” encodings (i.e. the intersection of the two).

RESULTS

1. Vectorial document representation:

1.1. In comparison with the automated annotation from a manual list (silver standard):

Table 3 shows the mean accuracy results for each phenotype. Since docBERT only considers 512 tokens, we chose to test both the first 512 ones and 512 tokens starting randomly in the text.

model	*nephritis in SLE*	*osteoporosis*	*Lung infection*	*ILD in Scc*
TF_IDF	0.21 [0.19; 0.23]	0.21 [0.20; 0.22]	0.29 [0.28; 0.31]	0.58 [0.54; 0.61]
doc2vec	0.41 [0.36; 0.45]	0.37 [0.35; 0.40]	0.32 [0.29; 0.36]	0.58 [0.53; 0.63]
docBERT 512 first tokens	0.02 [0.01;0.02]	0.02 [0.01;0.02]	0.02 [0.02;0.02]	0.04 [0.03; 0.04]
docBERT 512 tokens random start in the text	0.14 [0.13; 0.14]	0.11 [0.10; 0.11]	0.08 [0.08; 0.08]	0.10 [0.10; 0.10]

Table 3: Comparison of precisions@100, the precision at 100 corresponds to the number of positive documents for the phenotype out of the 100 closest by cosine similarity. Cosine similarity is calculated from the gold standard of 256 documents. A document is defined as positive if it contains at least one silver standard key term. For the query expansion method, there is no distance, and the results presented here are calculated on 100 random candidates. This analysis was carried out for all four methods on the same restricted base of 10,000 documents (corresponding to the silver standard containing a list of keywords pre-established by a clinician outside the study), including 664 candidates to be found for lupus nephropathy, 731 for osteoporosis, 943 for pulmonary infection and 886 with interstitial lung involvement in scleroderma. As an indication, the advanced keyword search method gives the following results for this experiment: 0.99 [0.98; 1.0] for lupus nephropathy, 0.98 [0.97; 0.99] for osteoporosis, 0.99 [0.98; 1.0] for lung infection and 0.99 [0.98; 1.0] for ILD in Scc.

An analysis with similar results was carried out directly on the gold standard dataset of 256 documents: for each phenotype, documents were taken in turn as document-indexes and precision@10 was defined as the number of positive documents (in the annotation sense) in the 10 closest. We obtained an average precision@10 of 0.26 [0.18; 0.37] for the TF-IDF method, 0.29 [0.27; 0.32] for the doc2vec method, 0.12 [0.09; 0.15] for the docBERT (random sequence of 512 tokens) and 0.98 [0.96;1.0] for the advanced query expansion model.

1.2. In comparison with IDC-10 encodings:

As described in the “Method” section, we also evaluated the accuracy of document retrieval for the vector representation of documents with a set of 100 sample documents by comparing the common ICD-10 encodings of the k (here 100) closest documents of the entire sample set, as shown in table 4. We checked whether the documents had the same ICD-10 encoding for the principal diagnosis, or at least one associated diagnosis, or at least one principal or associated diagnosis. As in the previous experiment, we obtained unsatisfactory results for clinical practice, especially for the docBERT model.

Evaluation **Model**	*At least one ICD-10 diagnostic in common*	*Same principal ICD-10 diagnostic*	*At least one associated ICD-10*	*Patient level (at least one common ICD-10)*
TF_IDF	0.55 [0.49; 0.60]	0.22 [0.16;0.27]	0.39 [0.34;0.45]	0.67 [0.61; 0.72]
doc2vec	0.58 [0.52; 0.64]	0.25 [0.20; 0.31]	0.39 [0.33; 0.45]	0.61 [0.55; 0.67]
docBERT	0.20 [0.18; 0.22]	0.07 [0.05; 0.09]	0.08 [0.08; 0.12]	0.27 [0.24; 0.30]

Table 4: ICD-10 precision results for the top 100 candidates based on a sample of 100 random documents from a restricted database of 10,000 documents. Here, precision is calculated by counting the number of the 100 closest documents having at least one ICD-10 coding in common with the document tested.

2. Advanced keywords search on the training set

We first conducted an experiment on training phenotypes to determine the best method for advanced keyword search. A summary of the precision and recall results is presented in supplementary Table 3. We used the silver standard of automatic annotated documents from a manual list for precision calculation, and recall is calculated at the patient level with the set of ICD-10s expected for the corresponding phenotype. Given the results, in order to maximize the number of synonyms and documents without decreasing accuracy, we chose a threshold of 0.75 for the cosine similarity of CODER embeddings and a preprocessing with stemming.

3. Advanced keywords search on the test set

As shown by Table 5, regarding the results on the 20 arbitrary queries, we obtained a precision evaluated on 50 random extracted documents for each query, with a mean precision of 0.93, and a confidence interval at 95% of [0.90; 0.96]. Overall recall had a mean of 0.78 and a 95% confidence interval of [0.71; 0.85].

	Query	**Precision** (on 50 manually annotated document)	**Recall** (with respective ICD-10) - 2 words query	**Number of documents**
1	“Rheumatoid Arthritis”	0.98	0.73	15189
2	“Takayasu”	1	0.94	2459
3	“Pericarditis in lupus”	0.92	0.93	7490
4	“Acute myocardial infarction”	0.94	0.58 - 0.90	10583 - 24017
5	“APS”	1.0	0.48 - 0.90	4406 - 23181
6	“Kidney transplantation”	0.92	0.98	10716
7	“Prostate cancer”	1.0	0.83	2971
8	“Lung tuberculosis”	1.0	0.55	3586
9	“Autoimmune hepatitis”	0.8	0.85	2797
10	“Dermatomyositis”	1.0	0.77	3510
11	“Idiopathic thrombocytopenic purpura”	0.98	0.81	3749
12	“Acute kidney injury”	0.86	0.81	15775
13	“Raynaud syndrome”	0.98	0.98	31900
14	“Pyelonephritis”	1.0	0.77	7475
15	“HIV”	0.90	0.98	43582
16	“Scleroderma”	1.0	0.92	24199
17	“Diabetes”	0.96	0.96	51224

18	“Stroke”	0.64	0.63	28162
19	“Stroke in lupus”	0.72	0.52	5650
20	“spontaneous miscarriage”	0.98	0.66 - 0.82	9071 - 14329
	Overall	**0.93 [0.90; 0.96]**	**0.78 [0.71; 0.85] - 0.83 [0.77; 0.89]**	

Table 5: Results from the 20 arbitrary queries, precision is measured from a manual evaluation by a clinician on a sample of 50 documents for each query. Recall (or sensitivity) is approximated using the standard ICD10 coding silver (the ICD10 codes of the corresponding stay in the hospital report) for each query. The "two-word query" type corresponds to the case where the clinician enters two different query terms for a given phenotype, instead of just one.

Some queries, however, had poorer recall results due to the absence of frequent French acronyms in the UMLS. This is the case of *“antiphospholipid syndrome”* which is very frequently mentioned in French only with its acronym “APS” (“SAPL” in French) which is not present as a distinct entity in the UMLS. Similarly, “spontaneous miscarriage” is frequently referred to as “FCS” in French (stands for “Fausse Couche Spontanée”), but the latter is not the corresponding entity in the UMLS. Concerning “acute myocardial infarction” it is now frequently mentioned as “ACS” (“SCA” in French for “Syndrome Coronarien Aigu”) with the same corresponding ICD-10 encoding but is not present as the right synonym in the UMLS (since “SCA” stands for “Sudden Cardiac Arrest” in English). The addition of these acronyms in their respective queries has indeed led to a significant improvement in recall: from 0.48 to 0.9 for “APS”, from 0.58 to 0.9 for “acute myocardial infarction”, and from 0.66 to 0.82 for the “spontaneous miscarriage”.

Furthermore, “pulmonary tuberculosis” may seem to have low recall, but after an error analysis performed on 50 random false negative documents, only 14% of the latter actually had pulmonary tuberculosis. The remaining documents mentioned a wrong site of TB infection such as: “lymph node tuberculosis”, “disseminated tuberculosis”, or “meningeal tuberculosis”. Hence the IDC10 encodings were not as precise as our method.

For the “stroke” results, an error analysis showed that only 18% of 50 random documents not found were ultimately relevant, with “cerebral vasculitis” or “cerebral thrombophlebitis” mentioned. Supplementary Table 4 shows the error analysis results.

Finally, concerning the calculation times of the different methods, for the processing of 10,000 documents, with a CPU of 30G RAM, the TF-IDF method took 10 seconds, the doc2vec method 218 seconds, my docBERT method 455 seconds and the advanced query expansion method 0.9 seconds.

DISCUSSION

In this study, we show that the classical document representations by TF-IDF, doc2vec [13] and docBERT [14] are not efficient in retrieving patients with specific written clinical phenotypes, or the same ICD-10 coding, even at a parent ICD-10 level (one letter, two digits). This can be explained by the fact that these methods extract a vector summarizing the document, but similarity between patients can appear along many different dimensions. Distance metrics such as cosine are applied uniformly over the vectors and do not make the dimensions of interest explicit. The keyword search method makes this dimension of interest explicit, which explains the better performance.

Nevertheless, we propose a new interesting method, hybrid between the symbolic method (directly based on metathesaurus) and the deep learning method, corresponding to an advanced keyword search. We achieved very good precision and recall based on ICD-10 coding comparison. Our method is fully replicable and all our code is available on github at https://github.com/ChristelDG/EHR_2_vec.

Other models propose to search clinical documents of interest, one of the most interesting in the field being the "advanced cohort engines" [27] based on a temporal query language. Their algorithm is scalable and incorporates many variables but the cohort is initially built on ICD9 coding, and still requires adaptation by any user to learn how to write a query. Similarly, in French, Pressat et al [28] propose Doc'EDS, a search tool based on structured and unstructured data. They also retrieve documents from keywords with UMLS synonyms, but do not add CODER synonyms and evaluate the results at the document level.

One of the main advantages of our method is that it takes into account clinical time and enables patients to be searched quickly using a single term or acronym. The use of CODER synonyms makes it possible to search for terms even if the input diagnosis is not directly in the UMLS or is misspelled, and is one of the main contributions of our study. In addition, we have proposed a method evaluated at document and patient level by one gold standard, two silver standards and with clinical control of documents and error analysis.

One of the limitations of our study is that none of the proposed evaluation measures is perfectly satisfactory. Indeed, a manual term list may omit part of the documents containing misspelled terms or other synonyms not taken into account, so that it cannot be taken into account for the recall calculation and can only correspond to a standard silver, which moreover favors the keyword search approach. The analysis carried out on a gold standard of documents annotated by a clinician only concerns 256 documents and therefore cannot evaluate recall on the scale of the entire database.
Furthermore, as the error analysis shows and as has already been mentioned several times in the literature [29, 30, 33], IDC10 encodings are not sufficiently accurate. This is due to the fact that ICD-10 coding is carried out to financially value patient stays (and in particular the coding "pulmonary tuberculosis" corresponds globally to the same valuation as lymph node or meningeal tuberculosis). However, these various metrics, taken as a whole and coupled with error analysis, give a good idea of performance and the problems encountered by users.

Another limitation is that our method is still sensitive to acronyms or abbreviations that are not all present in the UMLS. As an illustration of this problem, we show that adding no more than three usual French acronyms absent from the UMLS enabled to improve the overall recall from 0.78 to 0.83 [0.77; 0.89].
Finally, the method used is a classic query expansion, and in particular not compared with more recent methods such as Splade [34], which enables parsimonious query expansion based on the BERT model [25].

Nevertheless, the use of the CODER[21] model has not been carried out in the past to our knowledge, and is certainly less sophisticated in terms of the model, but seems to us to be richer in terms of the use of medical knowledge. On the other hand, this document pre-selection stage, intended to be applied to a vast database, had to be computationally light. For example, greedy vector-based document representation methods such as Longformer[35] could not be tested.

We propose this algorithm as a first step before more computationally complex NLP algorithms. Thus, documents are extracted including if the diagnoses are denied, suggested or belong to a family member of the patient. These features are expected to be taken into account and filtered in a second step thanks to deep learning models [31, 32, 5].

CONCLUSION

We have tested three state-of-the-art document vector representations for extracting documents of interest from a large database. We show that none of these three methods is sufficiently efficient for this task in the context of hospitalization reports of autoimmune patients. We propose a new advanced keyword search method with automatic synonym search with very good performance in terms of precision and recall.

ETHICS

The results shown in this study derive from the analysis of the AP-HP data warehouse. The study and its experimental protocol was approved by the AP-HP Scientific and Ethical Committee (IRB00011591 decision number CSE 20-0093). All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the CNIL: Commission Nationale de l'Informatique et des Libertés https://www.cnil.fr/en/home). All medical records have been pseudonymized. Patients are informed by the AP-HP data warehouse that the data are pseudonymized and that they can object to their sharing. Their consent was therefore collected prior to our study.

ACKNOWLEDGMENT

First and foremost, we would like to thank Professor Pierre Zweigenbaum for his careful review of this work. The authors would also like to thank the AP-HP data warehouse, which provided the data and the computing power to carry out this study under good conditions. We would like to thank all the medical colleges, including internal medicine, especially Pr Jean-Emmanuel Kahn, Dr Guillaume Bussone, Pr Sébastien Abad, Dr Virginie Zarrouk, Dr Noémie Chanson, Dr Antoine Dossier, Pr Luc Mouthon, Dr Geoffrey Cheminet , rheumatology, especially Dr Augustin Latourte, Dr Florent Eymard, Pr Xavier Mariette, Dr Gaétane Nocturne, Pr Raphaele Serror, Pr Sébastien Ottaviani, Pr Francis Berenbaum, Pr Jérémie Sellam, Pr Yannick Allanore, Pr Jérôme Avouac, Pr Maxime Breban, Dr Félicie Costantino, dermatology, nephrology, pneumology, hepato-gastroenterology, haematology, endocrinology, gynaecology, infectiology, cardiology, oncology, emergency and intensive care units, that gave their agreements for the use of the clinical data.

AUTHOR CONTRIBUTIONS

Christel Gérardin: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing- original draft, Writing - review and editing. **Yuhan Xiong:** Investigation, Methodology, Software, Validation **Fabrice Carrat:** Conceptualization, Methodology, Project administration, Supervision, Writing - original draft, Writing - review and editing. **Xavier Tannier**: Conceptualization, Formal Analysis, Methodology, Writing - original draft, Writing - review and editing.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The datasets analyzed during the current study are not publicly available due the confidentiality of data from patient records, even after de-identification. However, access to the AP-HP data warehouse's raw data can be granted following the process described on its website www.eds.aphp.fr, contacting the Ethical and Scientific Commity at secretariat.cse@aphp.fr. A prior validation of the access by the local institutional review board is required. In the case of non-APHP researchers, the signature of a collaboration contract is moreover mandatory.

ABBREVIATIONS

AP-HP: assistance publique hôpitaux de Paris

APS: Antiphospholipid syndrome

CNIL: Commission Nationale de l’Informatique et des Libertés

DBOW: Distributed bag-of-words

DM: Distributed Memory

EHR: Electronic health records

ILD: Interstitial Lung Disease

NER: named-entity-recognition

NLP: natural language processing

SLE: systemic lupus erythematosus

TF-IDF: Term Frequency - Inverse Document Frequency

UMLS: Unified Medical Language System

REFERENCES

[1] Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer research, 2019; 79(21):5463-5470.

[2] Lieu TA, Herrinton LJ, Buzkov DE, Liu L, Lyons D, Neugebauer R. ... & Baer DM. Developing a Prognostic Information System for Personalized Care in Real Time. eGEMs, 2019;7(1).

[3] Jia Z, Zeng X, Duan H, Lu X, & Li H. A patient-similarity-based model for diagnostic prediction. International Journal of Medical Informatics, 2020;135:104073.

[4] Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, ... & Burgun A. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. Journal of biomedical informatics, 2017;73:51-61.

[5] Gérardin C, Mageau A, Mékinian A, Tannier X, Carrat F, Construction of cohorts of similar patients from automatic extraction of medical concepts, JMIR Preprints

[6] Neuraz Antoine, et al. "Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic." *Journal of medical Internet research* 22.8 (2020): e20773.

[7] Ng K, Kartoun U, Stavropoulos H, Zambrano JA, & Tang, PC (2021). Personalized treatment options for chronic diseases using precision cohort analytics. Scientific reports, 2021;11(1):1-13.

[8] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[9] Dynomant, Emeric, Stéfan J. Darmoni, Émeline Lejeune, Gaétan Kerdelhué, Jean-Philippe Leroy, Vincent Lequertier, Stéphane Canu, and Julien Grosjean. "Doc2Vec on the PubMed corpus: study of a new approach to generate related articles." *arXiv preprint arXiv:1911.11698* (2019).

[10] Sankhavara, Jainisha. "Biomedical document retrieval for clinical decision support system." *Proceedings of ACL 2018, Student Research Workshop*. 2018.

[11] Frasca, Maria, and Genoveffa Tortora. "Visualizing correlations among Parkinson biomedical data through information retrieval and machine learning techniques." *Multimedia Tools and Applications* 81.11 (2022): 14685-14703.

[12] Hanauer, David A., et al. "Electronic medical record search engine (EMERSE): an information retrieval tool for supporting cancer research." *JCO clinical cancer informatics* 4 (2020): 454-463.

[13] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. PMLR, 2014.

[14] Adhikari, Ashutosh, et al. "Docbert: Bert for document classification." *arXiv preprint arXiv:1904.08398* (2019).

[15] Sugathadasa, Keet, et al. "Legal document retrieval using document vector embeddings and deep learning." *Science and information conference*. Springer, Cham, 2018.

[16] Gaulin, Maclean, and Xiaoxia Peng. "Compensation Disclosure: A Study via Semantic Similarity." *Available at SSRN* (2021).

[17] Arora, Jhanvi, et al. "Artificial Intelligence as Legal Research Assistant." *FIRE (Working Notes)*. 2020.

[18] Gutierrez, Bernal Jimenez, et al. "Document classification for covid-19 literature." *arXiv preprint arXiv:2006.13816* (2020).

[19] Wajsbürt, Perceval, Arnaud Sarfati, and Xavier Tannier. "Medical concept normalization in French using multilingual terminologies and contextual embeddings." Journal of Biomedical Informatics 114 (2021): 103684.

[20] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment Pre-training for Biomedical Entity Representations. arXiv, preprint arXiv:2010.11784 (2020).

[21] Yuan, Zheng, et al. "CODER: Knowledge-infused cross-lingual medical term embedding for term normalization." *Journal of biomedical informatics* 126 (2022): 103983.

[22] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

[23] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[24] Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

[26] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de La Clergerie ÉV ... & Sagot B. CamemBERT: a tasty French language model. arXiv preprint arXiv:2019;1911.03894

[27] Alison Callahan, Vladimir Polony, José D Posada, Juan M Banda, Saurabh Gombar, Nigam H Shah, ACE: the Advanced Cohort Engine for searching longitudinal patient records, *Journal of the American Medical Informatics Association*, Volume 28, Issue 7, July 2021, Pages 1468–1479,

[28] Pressat-Laffouilhère T, Balayé P, Dahamna B, et al. Evaluation of Doc'EDS: a French semantic search tool to query health documents from a clinical data warehouse [published correction appears in BMC Med Inform Decis Mak. 2022 Apr 22;22(1):107]. *BMC Med Inform Decis Mak*. 2022;22(1):34. Published 2022 Feb 8. doi:10.1186/s12911-022-01762-4

[29] Ryan, Olivia F., et al. "Factors associated with stroke coding quality: a comparison of registry and administrative data." *Journal of Stroke and Cerebrovascular Diseases* 30.2 (2021): 105469.

[30] Peng, Mingkai, et al. "Coding reliability and agreement of International Classification of disease, 10th revision (ICD-10) codes in emergency department data." *International Journal of Population Data Science* 3.1 (2018).

[31] Zavala, Renzo Rivera, and Paloma Martinez. "The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study." *JMIR Medical Informatics* 8.12 (2020): e18953.

[32] Althari, Ghadeer, and Mohammad Alsulmi. "Exploring Transformer-Based Learning for Negation Detection in Biomedical Texts." *IEEE Access* 10 (2022): 83813-83825.

[33] Rochoy, Michaël, et al. "Vascular dementia encoding in the French nationwide discharge summary database (PMSI): Variability over the 2007–2017 period." *Annales de Cardiologie et d'Angéiologie*. Vol. 68. No. 3. Elsevier Masson, 2019.

[34] Formal T, Piwowarski B, Clinchant S. SPLADE: Sparse lexical and expansion model for first stage ranking. InProceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval 2021 Jul 11 (pp. 2288-2292).

[35] Beltagy, I., M.E. Peters, and A. Cohan, Longformer: The long-document transformer. arXiv. preprint arXiv:2004.05150, 2020.

Pathologies	**ICD-10 Codes**	**Number of patients**
Systemic lupus erythematosus (SLE)	M310, M321, M328, M329, L930, L931	22 252
Takayasu disease	M314	576
Systemic scleroderma	M340, M341, M348, M349	7 711
Antiphospholipid syndrome	D686	16 401

Supplementary Table 1 : Description of the study patient cohort.

Supplementary Table 2: Corresponding queries for comparison with document extraction

Phenotypes	*Manual list*	*Additional list*
Nephritis in SLE	['lupus nephrotopathy', 'lupus glomerulonephritis', 'lupus with renal involvement', 'lupus renal involvement', 'renal failure secondary to lupus', 'lupus glomerulopathy', 'lupus gn', 'class IV renal involvement', 'class V renal involvement', 'class III renal involvement', 'class VI renal involvement', 'extra membranous glomerulonephritis class V']	['glomerulonephritis', 'chronic renal failure', 'chronic renal disease', 'GEM', 'HSF', 'segmental and focal hyalinosis', 'renal damage'] AND ['lupus']
osteoporosis	['osteoporosis', 'osteoporotic']	
Lung infection	['inhalation pneumopathy', 'pneumopathy to', 'legionellosis', 'pulmonary infection', 'infectious pneumopathy', 'mechanically ventilated pneumopathy', 'MVP', 'pneumonia', 'bilateral pneumopathy',	

	'basal pneumopathy', 'bi-basal pneumopathy', 'basal lobar pneumopathy', 'ALFP', 'acute frank lobar pneumopathy', community-acquired lung disease', 'acute lung disease', 'documented lung disease', 'ventilator-associated lung disease', 'pulmonary-acquired sepsis', 'pulmonary-acquired septic shock', 'upper lobar lung disease', 'necrotizing lung disease', 'bronchopneumonia', 'bronchopneumonia']
ILS in Scc	['interstitial lung disease', 'interstitial lung disease', 'interstitial syndrome', 'lung damage', 'IPD', 'PINS', 'pulmonary fibrosis', 'interstitial fibrosis', 'IDF', 'interstitial damage', 'fibrosing lung disease'] AND ['systemic scleroderma', 'Scc', 'diffuse cutaneous scleroderma', 'limited cutaneous scleroderma', 'CREST syndrome','CREST']

Supplementary Table 3: Accuracy and recall results for each phenotype.

Phenotype	**CODER Synonym limit 0.75**	**CODER Synonym limit 0.8**	**CUI and CODER synonyms limit 0.8 without stemming**	**CUI and CODER synonyms limit 0.75 with stemming**	**CUI and CODER synonyms limit 0.8 with stemming**
nephritis in SLE	acc 0.99 rec 0.85	acc 0.99 rec 0.85	acc 0.99 rec 0.84	acc 0.99 rec 0.85	acc 0.99 rec 0.85
osteoporosis	acc 0.99 rec 0.90	acc 0.99 rec 0.90	acc 0.99 rec 0.88	acc 0.99 rec 0.90	acc 0.99 rec 0.90
Lung infection	acc 0.99 rec 0.56	acc 0.99 rec 0.56	acc 0.98 rec 0.52	acc 0.99 rec 0.57	acc 0.99 rec 0.57
ILD in Scc	acc 0.99 rec 0.73	acc 0.99 rec 0.72	acc 0.99 rec 0.67	acc 0.99 rec 0.72	acc 0.99 rec 0.72
pulmonary hypertension	acc 0.99 rec 0.70	acc 0.99 rec 0.68	acc 0.99 rec 0.94	acc 1.0 rec 0.96	acc 1.0 rec 0.96
Reflux	acc 0.98 rec 0.54	acc 0.98 rec 0.53	acc 0.93 rec 0.84	acc 0.94 rec 0.90	acc 0.94 rec 0.90
Gougerot Sjögren	acc 1.0 rec 0.87	acc 1.0 rec 0.87	acc 1.0 rec 0.82	acc 1.0 rec 0.88	acc 1.0 rec 0.88
Pulmonary Embolism	acc 0.99 rec 0.97	acc 0.99 rec 0.97	acc 0.99 rec 0.96	acc 1.0 rec 0.97	acc 1.0 rec 0.97
Overall	**acc 0.99 [0.99; 1.0]** **rec 0.77 [0.70; 0.83]**	**acc 0.99 [0.99; 1.0]** **rec 0.76 [0.69; 0.83]**	**acc 0.98 [0.97; 0.99]** **rec 0.81 [0.75; 0.87]**	**acc 0.99 [0.98; 1.0]** **rec 0.84 [0.79; 0.90]**	**acc 0.99 [0.98; 1.0]** **rec 0.84 [0.79; 0.90]**

Supplementary Table 4 : Error analysis on the three phenotypes with less performances : Analysis is performed by a clinician on a set of 50 random documents for each phenotype

Phenotype	**percentage of pertinent documents not found**	**examples of mention on pertinent document**	**examples of mention on not pertinent document**
Tuberculose pulmonaire	14%	“tuberculose pleurale et pulmonaire” (pleuro-pulmonary tuberculosis)	“tuberculose ganglionnaire” (lymph node tuberculosis) “tuberculose neuroméningée” (neuromeningitis tuberculosis) “tuberculose disséminée” (disseminated tuberculosis)
AVC	18 %	“vascularite cérébrale” (cerebral vasculitis), "thrombophlébite cérébrale" (cerebral thrombophlebitis)	other thrombus mentionned but no cerebral involvment in the document : “Budd chiari syndrom”, thombus intra cardiaque (intra-cardiac thrombus), “méningiome” (meningioma)
AVC Lupus	34 %	“hémorragie sous arachnoïdienne” (subarachnoid hemorrhage) “vascularite cérébrale” (cerebral vasculitis), "thrombophlébite cérébrale" (cerebral thrombophlebitis)	“scleromyosite” (scleromyositis, no lupus mentioned) “myositis” or no stroke mentioned