Impact of translation on biomedical information extraction: an experiment on real-life clinical notes

Christel Gérardin, Yuhan Xiong, Perceval Wajsbürt, Fabrice Carrat, Xavier Tannier

Submitted to: JMIR Medical Informatics
on: June 03, 2023

Table of Contents

Impact of translation on biomedical information extraction: an experiment on real-life clinical notes

Christel Gérardin[1] MD, MSc; Yuhan Xiong[2]; Perceval Wajsbürt[3] PhD; Fabrice Carrat[1, 4] PhD, MD; Xavier Tannier[5] PhD

[1]IPLESP, Sorbonne Université, INSERM PARIS FR
[2]Shanghai Jiaotong University Shanghai CN
[3]Innovation and Data unit, Assistance Publique Hôpitaux de Paris Paris FR
[4]Public Health Department, Hôpital Saint-Antoine, Assistance Publique Hôpitaux de Paris PARIS FR
[5]Sorbonne Université, Inserm, Laboratoire d'informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS PARIS FR

Corresponding Author:
Christel Gérardin MD, MSc
IPLESP, Sorbonne Université, INSERM
27 rue de Chaligny
PARIS
FR

Abstract

Background: Biomedical natural language processing tasks are best performed with English models, and translation tools have undergone major improvements. On the other hand, building annotated biomedical datasets remains a challenge.

Objective: The objective of our study is to determine whether using English tools to extract and normalize French medical concepts on translations provides comparable performance to French models trained on a set of annotated French clinical notes.

Methods: We compare two methods: a method involving French language models and a method involving English language models. For the native French method, the Named Entity Recognition (NER) and normalization steps are performed separately. For the translated English method, after the first translation step, we compare a two-step method and a terminology- oriented method that performs extraction and normalization at the same time. We used French, English and bilingual annotated datasets to evaluate all steps (NER, normalization and translation) of our algorithms.

Results: The native French method performs better than the translated English one with a global f1 score of 0.51 [0.47;0.55] against 0.39 [0.34;0.44] and 0.38 [0.36;0.40] for the two English methods tested.

Conclusions: Despite the recent improvement of the translation models, there is a significant performance difference between the two approaches in favor of the native French method which is more efficient on French medical texts, even with few annotated documents.

(JMIR Preprints 03/06/2023:49607)
DOI: https://doi.org/10.2196/preprints.49607

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Impact of translation on biomedical information extraction: an experiment on real-life clinical notes

Christel Gérardin[a,*], Yuhan Xiong[a,b], Perceval Wajsbürt[c], Fabrice Carrat[a,d], Xavier Tannier[e]

[a]*IPLESP, 27 rue de Chaligny, Paris, 75012, France*

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

[c]*Innovation and Data unit, IT Department, Assistance Publique- hôpitaux de Paris, 33 bd Picpus, Paris, 75012, France*

[d]*Public Health Department, Hôpital Saint-Antoine, Assistance Publique- hôpitaux de Paris, 184 Rue du Faubourg Saint-Antoine, Paris, 75012, France*

[e]*Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie et des Connaissances en e-Santé, LIMICS, 15 rue de l'école de médecine, Paris, F-75006, France*

**Corresponding author : christel.ducroz-gerardin@iplesp.upmc.fr*

Abstract

Background: Biomedical natural language processing tasks are best performed with English models, and translation tools have undergone major improvements. On the other hand, building annotated

biomedical datasets remains a challenge.

Objective: The aim of our study is to determine whether the use of English tools to extract and normalize French medical concepts on translations provides comparable performance to that of French models trained on a set of annotated French clinical notes.

Methods: We compare two methods: one involving French-language models and one involving English-language models. For the native French method, the Named Entity Recognition (NER) and normalization steps are performed separately. For the translated English method, after the first translation step, we compare a two-step method and a terminology-oriented method that performs extraction and normalization at the same time. We used French, English and bilingual annotated datasets to evaluate all stages (NER, normalization and translation) of our algorithms.

Results: The native French method outperformed the translated English method, with an overall f1 score of 0.51 [0.47;0.55], compared with 0.39 [0.34;0.44] and 0.38 [0.36;0.40] for the two English methods tested.

Conclusion: Despite recent improvements in translation models, there is a significant difference in performance between the two approaches in favor of the native French method, which is more effective on French medical texts, even with few annotated documents.

Keywords: Concept normalization, Named entity recognition, Natural language processing, Translation

1. Introduction

Named entity recognition (NER) and term normalization are important steps in biomedical natural language processing (NLP). Named entity recognition is used to extract key information from textual medical reports, and normalization consists in matching a specific term to its formal reference in a shared terminology such as the UMLS metathesaurus [1]. Major improvements have been made recently in these areas, particularly for English, as a huge amount of data is available in the literature and resources. Modern automatic language processing relies heavily on pre-trained language models, which enable efficient semantic representation of texts. The development of algorithms such as transformers [2, 3] has led to significant progress in this field.

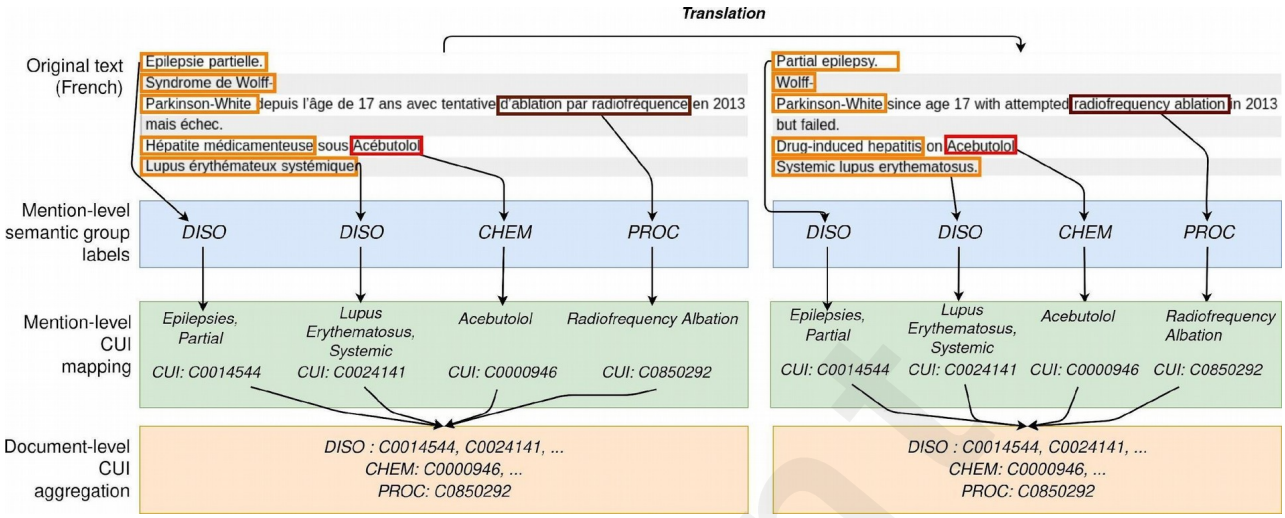


Figure 1: Overall objective of the method: from plain text to the CUI codes of the UMLS metathesaurus, document by document. The term "mention level" indicates that the analysis is carried out at the level of a word or small group of words: first, the NER stage (in blue), then normalization (in green), finally all mentions with normalized CUIs are aggregated at “document level” (orange part). The sets of aggregated CUIs per document predicted by the native French and translated English approaches are then compared to the manually annotated gold standard.

In many languages other than English, efforts remain to be made to obtain such interesting results, notably due to a much smaller quantity of accessible data [4].

In this context, our work explores the question of the relevance of a translation step for the recognition and normalization of medical concepts in French biomedical documents. We compare two methods: 1) a native French approach where only annotated documents and resources in French are used, and 2) a translation-based approach where documents are translated into English, in order to take advantage of existing tools and resources for this language that would allow the extraction of concepts mentioned in unpublished French texts without new training data (zero-shot) as proposed in Van Mulligen et al. [5].

We evaluate and discuss the results on several French biomedical corpora, including a new

set of 42 annotated hospitalization reports with 4 entity groups. We evaluate the normalization task at document level, in order to avoid a cross-language alignment step at evaluation time, which would add a potential level of error and thus make the results more difficult to interpret (see word alignment in [6, 7]). This normalization is carried out by mapping all terms to their Concept Unique Identifier (CUI) in the UMLS metathesaurus [1]. Figure 1 summarizes these various stages, from the raw French text and the translated English text to the aggregation and comparison of CUIs at document level. All our code is available on github [8].

2. Background

The various stages of our algorithms rely heavily on Transformers language models [2]. These models currently represent the state of the art for many natural language processing (NLP) tasks, such as machine translation, named entity recognition, classification and text normalization (also known as entity binding). Once trained, these models can represent any specific language, such as biomedical or legal. The power of these models comes from their neural architecture, but also largely depends on the amount of data they are trained on. In the biomedical field, two main types of data are available: public articles (e.g. PubMed) and clinical electronic medical record databases (e.g. MIMIC III [14]), and the most powerful models are, for example, BioBERT [15], which has been trained on the whole of PubMed in English, and ClinicalBERT [16], which has been trained on PubMed and MIMIC III. In French, the variety of models is less extensive, with camemBERT [17] and FlauBERT [18] for the general domain, and no specific model available for the biomedical domain.

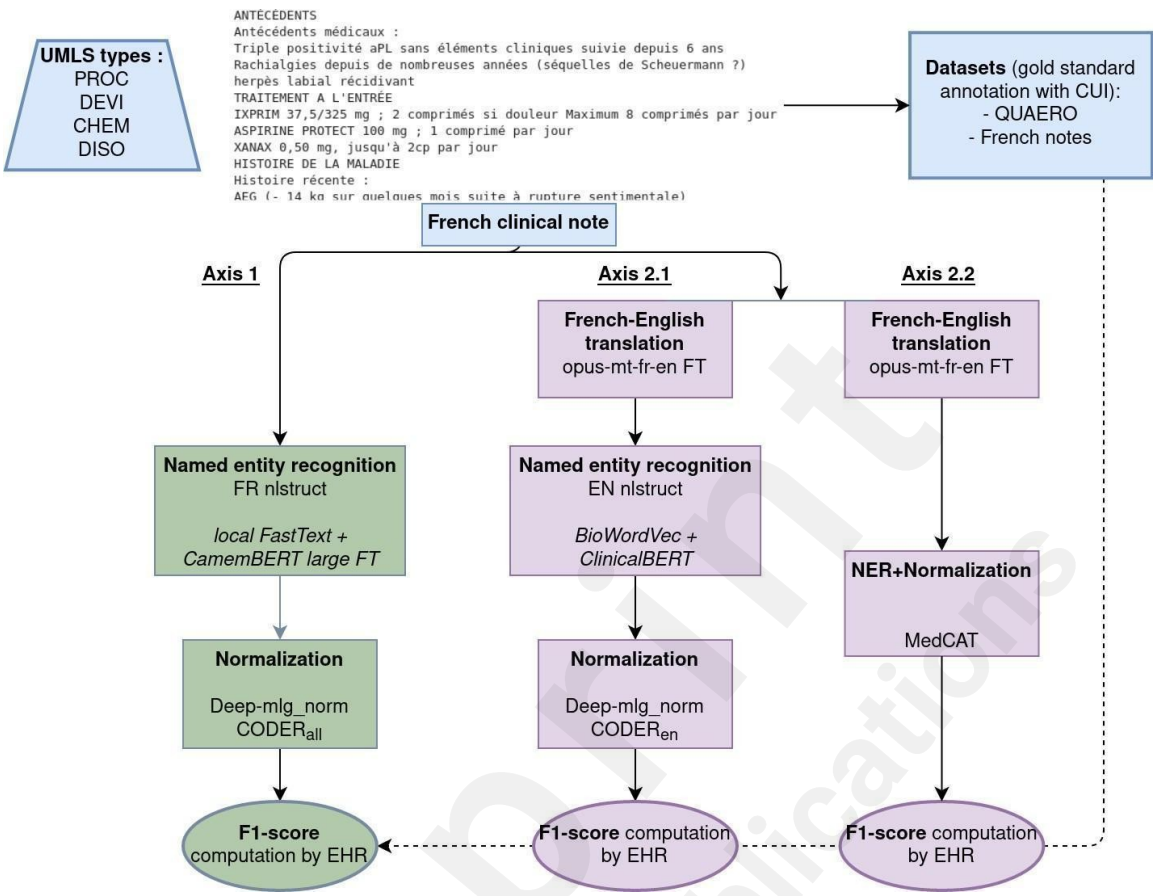


Figure 2: Diagram of different experiments comparing French and English language models without and with intermediate translation steps. Axis 1 (green axis on the left) corresponds to the native French branch with a NER step based on a FastText model trained from scratch on French clinical notes and a CamemBERT model. A multilingual BERT model is then used for the normalization step, with two models tested: a deep multilingual normalization model [9] and CODER[10] with the full version. Axes 2.1 and 2.2 (two purple axes on the right) correspond to the translated English branches, with a first translation step performed by the opus-mt-fr-en model [11] for both. Axis 2.1 (left) with decoupled NER and normalization steps, based on FastText trained from PubMed and Mimic III [12] for the NER part, and deep multilingual normalization [9] or CODER [10] with the English version, for normalization. Axis 2.2 (right) uses a single system for the NER and normalization stages: MedCAT [13].

In addition to particularly powerful English-language pre-trained models, universal biomedical terminologies (i.e. the UMLS metathesaurus) also contain many more English terms than other languages. For example, the UMLS metathesaurus [1] contains at least ten times more English terms than French terms, which may enable rule-based models to perform better in English. As mentioned above, each reference concept in the UMLS metathesaurus [1] is assigned a unique concept identifier (CUI), associated with a set of synonyms possibly in several languages, and a semantic group, such as Disorders, Chemicals and drugs, Procedure, Anatomy, etc.

In parallel, machine translation has also gained in performance thanks to the same type of transformer-based language models, and recent years have seen the emergence of high-quality machine translations, such as opus-mt developed by Tiedemann et al [11], Google Translate® and others. The latter two observations have led several research teams to add a translation step in order to analyze medical texts, for example to extract relevant mentions in ultrasound reports [19, 20] or in the case of standardization of medical concepts [9, 10, 21]. Work in the general (non-medical) domain has also focused on alignment between named entities in parallel bilingual texts [22, 23].

3. Materials and Methods

3.1. Overview

Figure 2 shows the main approaches and models used in our study. We explored one "native French approach axis" (axis 1 in Figure 2) based on French linguistic models learned and applied to French annotated data; and two "translated English approach axes" (axes 2.1 and 2.2) based on a translation step and concept extraction tools in English. We compare the performance of all

axes with an average of document-level CUI prediction precisions for all documents.

3.1.1. Native French approach

Axis 1 consists of two stages: a NER stage and a normalization stage. For the NER stage, we used the nested named entity extraction algorithm presented in section 3.4 below. Next, a normalization step is performed by two different algorithms: a deep multilingual normalization model [9] and CODER [10] with the *all* version (detailed in sections 3.5.1 and 3.5.2 respectively).

3.1.2. Translated-English approach

Axes 2.1 and 2.2 consist firstly of a translation step, presented in section 3.3 below, performed by the state-of-the-art opus-mt-fr-en [11] or Google Translate® algorithm. Secondly, like Axis 1, Axis 2.1 is based on a NER and normalization step. The NER step is performed by the same algorithm but trained on the n2c2 2019 dataset [24] without manual annotation realignment, for the normalization step we used the same deep multilingual algorithm [9] (section 3.5.1) and the English version of CODER [10] based on a BioBERT model [15]. This axis allows us to compare two methods whose difference lies solely in the translation step.

Axis 2.2 is based on the MedCAT[13] algorithm presented in section 3.5.3, which performs NER and normalization simultaneously. In this case, we compare the native French method with a state-of-the-art, ready-to-use English system, which is not available in French.

3.2. Datasets

3.2.1. Overview

For all our experiments, we chose to focus on four semantic groups of the UMLS metathesaurus [1]: *Chemical & Drugs* (CHEM) and *Devices* (DEVI) corresponding to medical devices such as pacemakers, catheters, etc., *Disorders* (DISO) corresponding to all signs,

symptoms, results (e.g. positive or negative results of biological tests) and diseases, *Procedures* (PROC) corresponding to all diagnostic and therapeutic procedures such as imaging, biological tests, operative procedures, etc., as well as the corresponding number of documents.

Table 1 shows the datasets used for all our experiments, and the corresponding numbers of documents. Firstly, two French datasets were used for the final evaluation, as well as for training the Axis 1 models. QUAERO is a freely available corpus [25] based on pharmacological notes with two sub-corpora: Medline (short sentences from PubMed abstracts) and EMEA (drug package inserts). We also annotated a new dataset of real-life clinical notes from the Assistance-Publique Hôpitaux de Paris data warehouse, described in the supplementary materials in section 1.1.1.

Secondly, we used the n2c2 2019 corpus [24] with annotated CUIs, on which we automatically added semantic group information from the UMLS metathesaurus [1], to train the axis-2.1 system and evaluate the NER and English normalization algorithms. We also used the Mantra dataset [26], a multilingual reference corpus for biomedical concept recognition.

Finally, we refined and tested the translation algorithms on the WMT biomedical corpora of 2016 [27] and 2019 [28]. A detailed description of the number of respective entities in the datasets can be found in Supplementary Table 1.

Table 1: Datasets. Overview of all datasets used. When a dataset is used for both training and testing, 80% of the dataset is used for training and 20% for testing. Thus, for the EMEA dataset, 30 documents were used for training and 8 for testing, 34 French notes were used for training and 8 for testing, and so on.

Language	French			English		English-French	
Datasets	Quaero		French notes	n2c2_2019	Mantra (English)	WMT 2016	WMT 2019
	EMEA	Medline					
Type	Drug notices	Medline titles	French notes	English notes	Drug not. & Medline titles	Pubmed abstracts	Pubmed abstracts
Size (docs)	38	2514	42	100	200	> 600k sent.	6542
Used for :							
Train NER	x	x	x	x			
Test NER	x	x	x	x			
Normalization	x	x	x	x			
Test MedCAT				x	x		
Translation (Fine Tuning)						x	x
Translation (test)						x	

The annotation methods for the French corpus are detailed in section 1.1.1 of the supplementary materials with supplementary figure 1. The distribution of entities for this annotation is detailed in Supplementary Table 1.

3.3. Translation

We used and compared two main algorithms for the translation step: the opus-mt-fr-en model [11], which we tested without and with *fine-tuning* on the two biomedical translation corpora of 2016 and 2019 [27, 28], and Google Translate® as a comparison model.

3.4. Named Entity Recognition

For this step, we used the algorithm of Wajsburt et al[29] described in [30]. This model is based on the representation of a BERT transformer [3] and calculates the scores of all possible concepts to be predicted in the text. The extracted concepts are delimited by three values: (start, end, label). More precisely, the encoding of the text corresponds to the last 4 layers of BERT, FastText integration and a max-pool Char-CNN [31] representation of the word. The decoding step is then performed by a 3-layer LSTM [32] with learning weights [33], similar to the method in [34]. A sigmoid function is added to the vertex. Values (start, end, label) with a score greater than 0.5 are retained for prediction. The loss function is a binary cross-entropy, and we used the Adam optimizer [35].

In our experiments, for the native French axis (Axis 1 in Fig. 2), the pre-trained embeddings used to train the model were based on a FastText [36] trained from scratch on 5 gigabytes of clinical text and a camemBERT-large [17] *fine-tuned* on this same dataset. For English Axis 2.1, the pre-trained models were BioWordVec [12] and clinicalBERT [16].

3.5. Normalization algorithms

This stage of our experiments is essential for comparing a method in native French and one

translated into English, and consists in matching each mention extracted from the text to its associated CUI in the UMLS metathesaurus [1]. We compare three models for this step, described below: the deep multilingual normalization algorithm developed by [9], CODER[10] and the MedCAT[13] model, which performs both NER and normalization.

These three models require no training dataset other than the UMLS metathesaurus.

3.5.1. Deep multilingual normalization

This algorithm by Wajsbürt et al [9] considers the normalization task as a highly multiclass classification problem with cosine similarity and a softmax function as the last layer. The model is based on contextual integration, using the pre-trained multilingual BERT model [3] and works in two steps: in the first step, the BERT model is fine-tuned and the French UMLS terms and their corresponding English synonyms are learned. Then, in the second step, the BERT model is frozen and the representation of all English-only terms (i.e. those present only in English in the UMLS metathesaurus [1]) is learned. The same training is used for the native French and translated English approaches. This model was trained with version 2021 of the UMLS metathesaurus [1], corresponding to the version used for annotating the French corpus. The model was thus trained on over 4 million concepts corresponding to 2 million CUIs.

3.5.2. CODER

The CODER algorithm [10] is developed by contrastive learning on the basis of the medical knowledge graph of the UMLS metathesaurus [1], concept similarities being calculated from the representation of terms and relations in this knowledge graph. Contrastive learning is used to learn embeddings through multisimilarity loss [37]. The authors have developed two versions: a multilingual version based on the multilingual BERT [3] and an English version based on the pre-trained BioBERT model [15]. We used the multilingual version for Axis 1 (native French

approach), and the English version for Axis 2.1. Both types of this model (CODER all and CODER en) were trained with the 2020 version of UMLS (publicly available models). CODER all [10] was trained on over 4 million concepts corresponding to 2 million CUIs, and CODER en was trained on over 3 million terms and 2 million CUIs.

For the deep multilingual model and the CODER model, in order to improve performance in terms of accuracy, we have chosen to add semantic group information (i.e. CHEM, DEVI, DISO, PROC) to the model output: i.e., from the first *k* CUIs chosen from a mention, we select the first from the right group.

The MedCAT algorithm is described in detail in section 1.1.1 (supplementary material).

Table 2: NER performance for each model. For all experiments, we used the same NER algorithm described in section 3.4, but with different pre-trained models. FastText* corresponds to a FastText [36] trained from scratch on our local clinical dataset.

Data set		EMEA test			French notes			n2c2 2019 test		
Models		FastText* & camemBERT-FT			FastText* & camemBERT-FT			BioWordVec [12] & ClinicalBERT [16]		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
Groups	CHEM	0.80	0.83	0.82	0.84	0.88	0.86	0.87	0.85	0.86
	DEVI	0.42	0.81	0.55	0.00	0.00	0.00	0.58	0.51	0.54
	DISO	0.54	0.63	0.59	0.67	0.65	0.66	0.74	0.72	0.73
	PROC	0.73	0.78	0.74	0.78	0.72	0.75	0.80	0.78	0.79
	Overall	**0.71**	**0.77**	**0.74**	**0.73**	**0.71**	**0.72**	**0.78**	**0.76**	**0.77**

4. Results

The sections below present the performance results for each stage. The n2c2 2019 challenge corpus [24] enabled us to evaluate the performance of our English models on clinical data and the Biomedical Translation 2016 shared task [27] to evaluate our translation performance on

biomedical data with a BLEU score [38].

4.1. NER performances

To be able to compare our approaches in native French and translated English, we used the same NER model (section 3.4), trained and tested on each of the datasets described above (section 3.2). Table 2 shows the corresponding results. Overall F1 scores are similar across datasets: from 0.72 to 0.77.

4.2. Normalization performance

This section presents only the normalization performance based on the gold standard's entity mentions, without the intermediate steps. The results are summarized in Table 3. The deep multilingual algorithm performs better for all corpora tested, with an improvement in F1 score from +0.6 to +0.11. By way of comparison, the winning team of the 2019 n2c2 challenge had achieved an accuracy of 0.85 using the n2c2 dataset directly to train their algorithm [24]. In our context of comparing algorithms between two languages, normalization algorithms are not trained on data other than the UMLS metathesaurus. MedCAT's performance (shown in supplementary table 2) cannot be directly compared with that of other models, as this method performs both NER and normalization in a single step. However, we note that this algorithm performs as well as Axis 2.1 in terms of overall performance, as shown in Table 5.

Table 3: Performance of the normalization step. Model results are calculated from the annotated datasets, focusing on the four semantic groups of interest: CHEM, DEVI, DISO, PROC.

Dataset Models	EMEA test	French notes	n2c2 2019 test
deep mlg norm	**0.65**	**0.57**	**0.74**

CODER all CODER en	0.58 –	0.51 –	– 0.63

4.3. Translation performances

For both translation models, the respective BLEU scores [38] are calculated on the shared 2016 Biomedical Translation task [27]. The chosen BLEU algorithm is the weighted geometric mean of the n-gram precisions per sentence.

A fine-tuned version of opus-mt-fr-en [11] on the 2016 and 2019 Biomedical Translation shared tasks was also tested. For fine-tuning, we used the following hyperparameters: a maximum sequence length of 128 (mainly for computational memory reasons), a learning rate of 2e-5, with a weight decay of 0.01, we varied the number of epochs up to 15 epochs (the error function curve stops decaying after 10 epochs). The Google translate model could not be used for our clinical score experiments for reasons of confidentiality.

Table 4 presents the BLEU score results for the three models, showing that fine-tuning the opus-mt-fr-en model [11] on biomedical datasets gave the best results, with a BLEU score[38] of 0.51. This is the model used to calculate the overall performance of axes 2.1 and 2.2.

4.4. Overall performances from raw text to CUI predictions

This section presents the overall performance of the 3 axes, in an end-to-end pipeline. For axis 2, the results are those obtained with the best normalization algorithm (presented in Table 3). The model used for translation is the opus-mt-fr-en [11] fine-tuned model. The results are presented in Table 5, with the best results obtained by the native French approach on the EMEA corpus [25] and French clinical notes. The 95% confidence intervals were calculated using the

empirical bootstrap method [39].

Table 4: **Translation performances.** BLEU scores of Translation models. *opus-mt-fr-en* FT corresponds to the *opus-mt-fr-en* model [11] *fine-tuned* on biomedical translated corpus from [27] and [28].

Data set		wmt biomed 2016 test
Models	Google Translate	0.42
	opus-mt-fr-en	0.31
	opus-mt-fr-en FT	**0.51**

Table 5: **Overall performances.** The normalization step is performed by the deep multilingual model and the translation by the opus-mt-fr-en FT model.

		EMEA test			French notes		
		precision	recall	f1-score	precision	recall	f1-sc
Methods	Axis 1 (French NER+normalization)	0.63	0.60	**0.61** [0.53;0.65]	0.49	0.53	**0.51** [0.47;0.5
	Axis 2.1 (Translation+NER+normalization)	0.53	0.40	0.45 [0.38;0.51]	0.41	0.38	0.39 [0.34;0.
	Axis 2.2 (Translation+MedCAT[13])	0.53	0.46	0.49 [0.38;0.54]	0.38	0.38	0.38 [0.36;0.

5. Discussion

In this article, we compared two approaches for extracting medical concepts from clinical notes. A French approach based on a French language model and a translated English approach where we compare two state-of-the-art English biomedical language models, after a translation step. The main advantages of our experiment are that it is reproducible and that we were able to analyze the performance of each step of the algorithm: NER, normalization and translation, and to test several models for each step.

5.1.1. The quality of the translation is not sufficient

We show that the native French approach outperforms the two translated English approaches, even with a small French training dataset. This analysis confirms that, where possible, an annotated dataset improves feature extraction. The evaluation of each intermediate step shows that the performance of each module is similar in French and English. We can therefore conclude that it is rather the translation phase itself that is of insufficient quality to allow the use of English as a proxy without loss of performance. This is confirmed by the translation performance calculations, where the calculated BLEU scores are relatively low, although improved by a fine-tuning step.

In conclusion, although translation is commonly used for entity extraction or term normalization in languages other than English [20, 40, 41, 42, 5], due to the availability of turnkey models that do not require additional annotation by a clinician, we show that this induces a significant performance loss.

Commercial API-based translation services could not be used for our task due to data

confidentiality issues. However, the opus-mt model is considered state-of-the-art, it is adjustable to domain-specific data, and the translation results presented in Table 4 confirm the absence of performance difference between this model and the google translate model.

Although our experiments were carried out on a single language, the French-English pair is one of the best performers in recent translation benchmarks[43]. Other languages are unlikely to produce significantly better results.

5.1.2. Error Analysis

In these experiments, the overall results may appear low, but the task is still complex, especially because the UMLS Metathesaurus [1] contains many synonyms with different CUIs. To better understand, we performed an error analysis on the normalization task only, as shown in Supplementary Table 3, with a physician's evaluation, on a sample of 100 errors for both models. We calculated that 24% and 39% of the terms found by the deep normalization algorithm [9] and CODER [10] respectively were in fact synonyms but with two different UMLS CUIs. This highlights the difficulty of achieving normalization on the UMLS metathesaurus. The UMLS metathesaurus indeed groups together numerous terminologies whose mapping between terms is often imperfect, implying that certain synonyms, as shown here, do not have the same CUI, as pointed out by [45] and [46]. For example, cardiac ultrasound has CUI C1655737 while echocardiography has another CUI C0013516, similarly H/O: thromboembolism has a CUI of C0455533 while history of thromboembolism has a CUI of C1997787 and so on.

Moreover, to be more precise, each axis has its own errors: overall, the errors in axis 2 are essentially due to loss of information in translation. One notable error is literal translation: for example, "dispersed lupus erythematous" instead of "systemic lupus erythematosus", or "crepitant" instead of "crackles". This loss of translation leads to more errors in the extraction of named entities.

In addition to the loss of translation information, axis 2.1. is also penalized by the NER step, due to the difference between the training set (n2c2 notes) and the test set: the translated French notes (the aim being to compare the performance of English-language turnkey models with the performance of French-language models from an annotated set). Axis 2.1, for example, omits the names of certain drugs more often.

Finally, both axes are penalized by abbreviations. These are often badly translated (for example, "MFIU": "mort foetale in utero" meaning "intra-uterine fetal death" is not translated), which penalizes axis 2. Nevertheless, if they are indeed extracted by NER steps in Axis 1, they are not correctly normalized due to the absence of a corresponding CUI in the UMLS metathesaurus.

5.1.3. Limitations

This work has several limitations. Firstly, the actual French clinical notes contained very few terms in the "Devices" semantic group, which prevented the NER algorithm from finding them in the test dataset. However, this drawback, which penalizes the native French approach, still allows us to conclude on the results. Furthermore, in this study, we did not take into account attributes of the extracted terms such as negation, hypothetical attribute or belonging to a person other than the patient, this for comparison purposes, as the QUAERO [25] and n2c2 2019 [24] datasets did not have this labeled information.

Ethics

The study and its experimental protocol were approved by the AP-HP Scientific and Ethical Committee (IRB00011591, decision number CSE 20-0093). Patients were informed that their

EHR information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were applied in accordance with the relevant guidelines (CNIL reference methodology MR-004: Commission nationale de l'informatique et des libertés [44]).

Data availability

The datasets analyzed as part of this study are not accessible to the public due to the confidentiality of data from patient files, even after de-identification. However, access to raw data from the AP-HP data warehouse can be granted by following the procedure described on its website: www.eds.aphp.fr, by contacting the ethical and scientific committee at secretariat.cse@aphp.fr. Prior validation of access by the local institutional review committee is required. In the case of non-APHP researchers, a collaboration contract must also be signed.

Acknowledgments

The authors would 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 wish to thank all the medical colleges, including internal medicine, rheumatology, dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency and intensive care units, that gave their agreements for the use of the clinical data.

Competing interest

Authors declare no competing interest

Consent for publication

Not applicable

Funding

Not applicable

Authors contribution

Christel Gérardin: worked on the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, original drafting, writing, revising, and editing the manuscript.

Yuhan Xiong: worked on investigation, methodology, software, validation.

Perceval Wajsbürt: worked on the investigation, the software, the revision of the manuscript.

Fabrice Carrat: worked on conceptualization, methodology, project administration, supervision, writing - original version, writing - revision and editing of the manuscript.

Xavier Tannier: worked on conceptualization, formal analysis, methodology, writing - original version, writing - revision and editing of the manuscript.

References

[1] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[3] 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[4] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (1) (2018) 1–13.

[5] E. M. van Mulligen, Z. Afzal, S. A. Akhondi, D. Vo, J. A. Kors, Erasmus MC at CLEF ehealth 2016: Concept recognition and coding in french texts, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), Working Notes of CLEF 2016 Conference and Labs of the Evaluation forum, E´vora, Portugal, 5-8 September, 2016, Vol. 1609 of CEUR Workshop Proceedings, CEUR-WS.org, 2016, pp. 171–178.

[6] Q. Gao, S. Vogel, Parallel implementations of word alignment tool, in: Software engineering, testing, and quality assurance for natural language processing, 2008, pp. 49–57.

[7] S. Vogel, H. Ney, C. Tillmann, Hmm-based word alignment in statistical translation, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996.

[8] Github link, https://github.com/ChristelDG/biomed translation.

[9] P. Wajsbürt, A. Sarfati, X. Tannier, Medical concept normalization in French using multilingual terminologies and contextual embeddings, Journal of Biomedical Informatics 114 (2021) 103684.

[10] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, S. Yu, Coder: Knowledge-infused cross- lingual medical term embedding for term normalization, Journal of biomedical informatics (2022) 103983.

[11] J. Tiedemann, S. Thottingal, OPUS-MT Building open translation services for the World, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[12] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, Biowordvec, improving biomedical word embeddings with subword information and mesh, Scientific data 6 (1) (2019) 1–9.

[13] Z. Kraljevic, D. Bean, A. Mascio, L. Roguski, A. Folarin, A. Roberts, R. Bendayan, R. Dobson, Medcat–medical concept annotation tool, arXiv preprint arXiv:1912.10166 (2019).

[14] A. Johnson, T. Pollard, L. Shen, L.w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035. doi:10.1038/sdata.2016.35.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36

(4) (2020) 1234–1240.

[16] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and pre- dicting hospital readmission, arXiv preprint arXiv:1904.05342 (2019).

[17] L. Martin, B. Muller, P. J. Ortiz Suarez, Y. Dupont, L. Romary, de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7203–7219.

[18] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabb´e, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490.

[19] L. Campos, V. Pedro, F. Couto, Impact of translation on named-entity recognition in radiology texts, Database 2017 (2017).

[20] V. Suarez-Paniagua, H. Dong, A. Casey, A multi-bert hybrid system for named entity recognition in spanish radiology reports, CLEF eHealth (2021).

[21] Naiara Perez-Miguel, Montse Cuadros, and German Rigau. 2018. Biomedical term normalization of EHRs with UMLS. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

[22] Y. Chen, C. Zong, K.Y. Su, On jointly recognizing and aligning bilingual named enti- ties, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10, Association for Computational Linguistics, USA, 2010, p. 631–639.

[23] Y. Chen, C. Zong, K.Y. Su, A joint model to identify and align bilingual named entities, Computational linguistics 39 (2) (2013) 229–266.

[24] S. Henry, Y. Wang, F. Shen, O. Uzuner, The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (OHNLP) shared task on clinical concept normalization for clinical records, Journal of the American Medical Informatics Association (2020).

[25] A. Névéol, C. Grouin, J. Leixa, S. Rosset, P. Zweigenbaum, The QUAERO French medical corpus: A resource for medical entity recognition and normalization, in: Proc of BioTextMining Work, 2014, pp. 24–30.

[26] J. A. Kors, S. Clematide, S. A. Akhondi, E. M. Van Mulligen, D. Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: the mantra GCS, Journal of the American Medical Informatics Association 22 (5) (2015) 948–956.

[27] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, M. Zampieri, Findings of the 2016 conference on machine translation, in: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 131–198. doi:10.18653/v1/W16-2301.

[28] R. Bawden, K. Bretonnel Cohen, C. Grozea, A. Jimeno Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, A. Siu, K. Verspoor, M. Vicente Navarro, Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies, in: Proceedings of the Fourth Conference on Machine Translation

(Volume 3: Shared Task Papers, Day 2), Association for Computational Linguistics, Florence, Italy, 2019, pp. 29–53. doi:10.18653/v1/W19-5403.

[29] P. Wajsbürt, Extraction and normalization of simple and structured entities in medical documents, Theses, Sorbonne Universit´e (Dec. 2021).

[30] C. Gérardin, P. Wajsbürt, P. Vaillant, A. Bellamine, F. Carrat, X. Tannier, Multil- abel classification of medical concepts for patient clinical profile identification, Artificial Intelligence in Medicine 128 (2022) 102311.

[31] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[32] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[33] J. Kim, M. El-Khamy, J. Lee, Residual LSTM: Design of a deep recurrent architecture for distant speech recognition, arXiv preprint arXiv:1701.03360 (2017).

[34] J. Yu, B. Bohnet, M. Poesio, Named Entity Recognition as Dependency Parsing (Jun. 2020). arXiv:2005.07150.

[35] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[36] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[37] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair

weighting for deep metric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.

[38] Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[39] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, L. E. Meester, A Modern Introduction to Probability and Statistics: Understanding Why and How, SPRINGER NATURE, 2007.

[40] V. Cotik, H. Rodrıguez, J. Vivaldi, Spanish named entity recognition in the biomedical domain, in: Annual International Symposium on Information Management and Big Data, Springer, 2018, pp. 233–248.

[41] J. Hellrich, U. Hahn, Enhancing multilingual biomedical terminologies via machine translation from parallel corpora, in: International Conference on Applications of Natural Language to Databases/Information Systems, Springer, 2014, pp. 9–20.

[42] G. Attardi, A. Buzzelli, D. Sartiano, Machine translation for entity recognition across languages in biomedical documents., in: CLEF (Working Notes), Citeseer, 2013.

[43] Tiedemann, Train Opus-MT models, Language Technology at the University of Helsinki (Jun. 2022).

[44] Homepage — CNIL, https://www.cnil.fr/en/home.

[45] Cimino, James J. "Auditing the unified medical language system with semantic methods." Journal of the American Medical Informatics Association 5.1 (1998): 41-51.

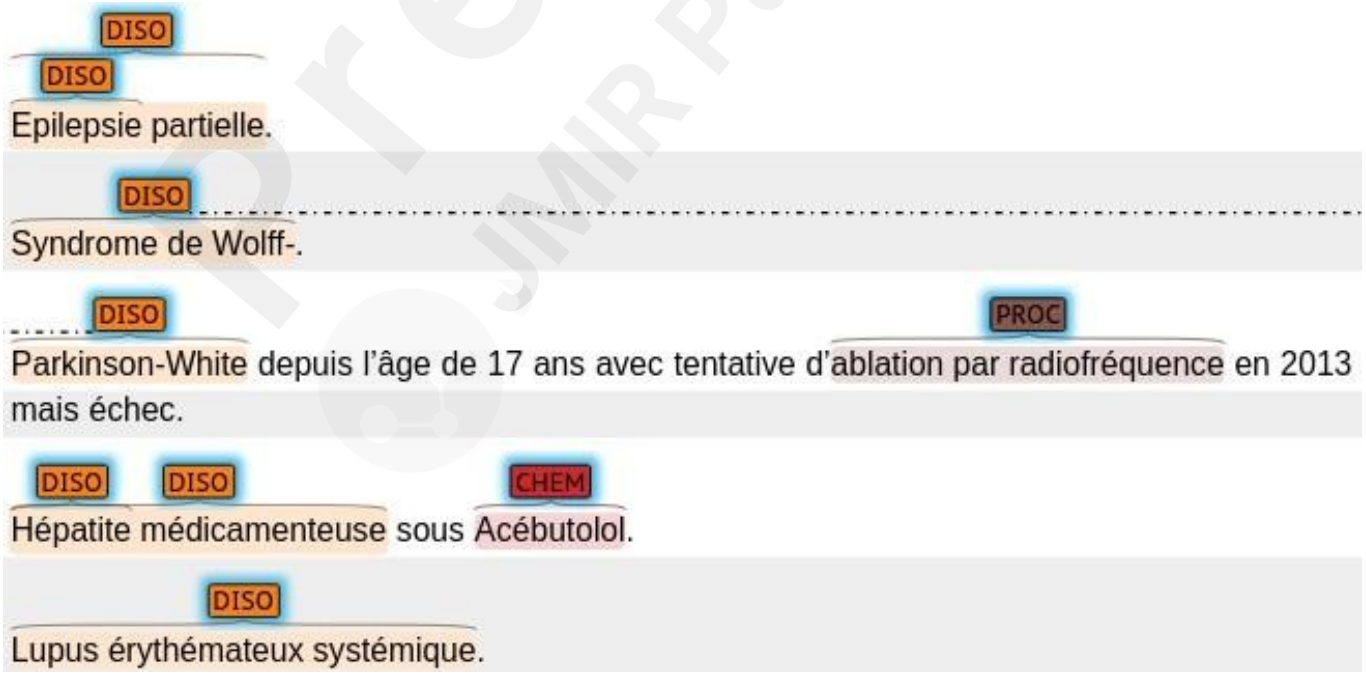
[46] Jiménez-Ruiz, Ernesto, et al. "Logic-based assessment of the compatibility of UMLS ontology

sources." Journal of biomedical semantics 2.1 (2011): 1-16.

Supplementary Materials

Supplementary Table 1: Datasets. Detailed description of the size of the datasets with the number of annotated entities (with UMLS semantic groups and CUIs).

Datasets Types	QUAERO		French notes	n2c2 2019	Mantra English
	EMEA	MEDLINE			
CHEM	2481	1053	869	1556	2600
DEVI	170	128	57	193	113
DISO	1509	2837	2617	4463	4115
PROC	952	1749	2310	3834	1150
Overall	5112	5767	5853	10046	7662



Supplementary Figure 1: Example of clinical notes annotation with the four groups of interest (CHEM, DEVI, DISO, PROC)

1.1.1. French corpus annotation

Semantic group and CUI annotation on clinical notes was performed by a physician. A total of 42 hospitalization reports from different departments: internal medicine, gynecology, rheumatology, etc. were annotated corresponding to a total of 5853 entities, as shown in the Supplementary Table 1. All corresponding CHEM, DEVI, DISO and PROC entities were annotated, including the case where the concept was not present in the UMLS® (neither in French nor in English). Figure 3 gives an example of this annotation with the four different groups. Following QUAERO[25] annotation rules, nested entities were also annotated. The section on data availability below details the conditions for sharing this corpus.

Supplementary Table 2 : MedCAT performances. Presentation of the overall results of MedCAT (NER + normalization), focusing on the four semantic groups of interest : CHEM, DEVI, DISO, PROC.

	n2c2 2019 test			Mantra English		
MedCAT	precision	recall	f1-score	precision	recall	f1-score
	0.41	0.54	0.46	0.52	0.47	0.48

Supplementary Table 3 : Error analysis for the normalization step. This table presents a list of normalization errors made by the deep multilingual model described in section 3.5.1, terms with very similar meaning or synonyms are highlighted in orange.

true text mention	**CUI pred**	**UMLS term**	**CUI stand**	**UMLS term**
fcs	C0587248	FCS-304	C0000786	Spontaneous abortion
hb	C0019046	Hemoglobin	C0518015	Hemoglobin measurement
transfusion	C1879316	Transfusion (procedure)	C0005841	Blood Transfusion
fibrinogène	C0337428	Fibrinogen assay	C2065007	fibrinogen transfusion
bilan biologique	C2717898	Biostatistics	C1511148	Biological Testing
ionogramme	C0288925	Ionogran	C0853360	Blood electrolytes
ecg	C1623258	Electrocardiography	C0013798	Electrocardiogram
hospitalisée	C0701159	Patient in hospital (finding)	C0019993	Hospitalization
rai	C3178806	Right Atrial Isomerism	C0430262	Indirect Coombs Test
sleeve	C0878989	Conductive Sleeves	C4758637	Gastric sleeve
péridurale	C1283259	Epidural device	C0002913	Epidural Anesthesia
curetage	C0180236	Curette	C0010468	Dilatation and Curettage
remplissage	C0035139	Surgical Replantation	C4761258	Blood volume expansion
pfc	C0070493	PFL protocol	C0016709	Fresh frozen plasma
noradrénaline	C0202145	Norepinephrine measurement	C0028351	norepinephrine
artério-embolisation	C3163695	Transarterial embolization	C0397760	Embolization of artery
nacl	C0037494	sodium chloride	C0036082	Saline Solution
kta	C0313203	Blood group antibody EnˆaˆKT	C0581447	Arterial catheter
morphine	C2698261	Morphine Measurement	C0026549	morphine
rl	C2935745	yellow RL	C0073385	Lactated Ringer's Solution
rx	C0809849	Radiation Rx	C1306645	Plain x-ray

normotendue	C0606030	Normoten	C2712122	Normal blood pressure
calcémie corrigée normale	C3468265	CALCR normal variant	C0853771	Blood calcium normal
troponine en diminution	C0041199	Troponin	C5442263	Troponin decreased
examen clinique	C1456356	examination; clinical	C0031809	Physical Examination
imagerie	C0441633	Scanning	C0011923	Diagnostic Imaging
thromboemboliques	C0016018	Fibrinolytic Agents	C0040038	Thromboembolism
sigmoïdites diverticulaires	C1395719	sigmoid; diverticulum	C0012814	Colonic Diverticulitis
histologique	C0344441	Histology Procedure	C0679557	histological diagnosis