# Multilabel classification of medical concepts for patient clinical profile identification

Christel Gérardin, Perceval Wajsbürt, Pascal Vaillant, Ali Bellamine, Fabrice Carrat, Xavier Tannier

# Multilabel classification of medical concepts for patient clinical profile identification

Christel Gérardin[1,2], Perceval Wajsbürt[3], Pascal Vaillant[4],
Ali Bellamine[2], Fabrice Carrat[1,5], Xavier Tannier[3]

[1] Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Inserm, 27 rue Chaligny, 75012 PARIS

[2] Département de médecine interne, APHP. Sorbonne Université.

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

[4] Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en eSanté (LIMICS), Sorbonne Université, Inserm, F-93000, Bobigny, France.

[5] Public Health Department, Hôpital St-Antoine, APHP. Sorbonne-Université, Paris, France

**Corresponding author :**

Christel Gérardin

**Address:** IPLESP, 27 rue de Chaligny 75012 Paris

**phone:** +33 6 78 14 84 66

**e-mail:** christel.ducroz-gerardin@iplesp.upmc.fr

**ABSTRACT**

Background: The development of electronic health records has provided a large volume of unstructured biomedical information. Extracting patient characteristics from these data has become a major challenge, especially in languages other than English.

Methods: Inspired by the French Text Mining Challenge (DEFT 2021) [1] in which we participated, our study proposes a multilabel classification of clinical narratives, allowing us to automatically extract the main features of a patient report. Our system is an end-to-end pipeline from raw text to labels with two main steps: named entity recognition and multilabel classification. Both steps are based on a neural network architecture based on transformers. To train our final classifier, we extended the dataset with all English and French Unified Medical Language System (UMLS) vocabularies related to human diseases. We focus our study on the multilingualism of training resources and models, with experiments combining French and English in different ways (multilingual embeddings or translation).

Results: We obtained an overall average micro-F1 score of 0.811 for the multilingual version, 0.807 for the French-only version and 0.797 for the translated version.

Conclusion: Our study proposes an original multilabel classification of French clinical notes for patient phenotyping. We show that a multilingual algorithm trained on annotated real clinical notes and UMLS vocabularies leads to the best results.
***Keywords***: biomedical concepts, multilabel classification, NER, transformers, multilingual NLP

## 1. Introduction

The widespread use of electronic health records (EHRs) has provided access to a large amount of health data. In addition to International Classification of Disease (ICD10) coding and biological examination data, a significant amount of patient information comes from narrative records, which are unstructured data. The exploitation of unstructured data has been made possible by significant advances in natural language processing (NLP) algorithms, including new language modeling algorithms [2, 3, 4]. These algorithms have proven to be very efficient in extracting information for various medical applications, including mortality prediction [5], cohort identification [6], and decision support [7, 8], especially in English. However, in French or other languages, efforts are still needed to reach the same level of performance.

We call a patient's *phenotype* the list of observable characteristics; in our case, the main pathological domain of a symptom or a disease, such as "cardiovascular" or "infections". A *disorder* corresponds to a disease, a pathological symptom or function. A *concept* is a generic name for a biomedical term or expression, such as "anuria", "fever", or "Sjögren's syndrome". The MeSH (Medical Subject Headings[1]) terminology was developed by the US National Library of Medicine and is structured like a tree with main categories A (anatomy), B (organisms), C (diseases), *etc.* and subcategories C01 (infections), C04 (neoplasms), *etc.* and finally concepts (i.e. leaves). There is a bilingual French-English MeSH version[2] used in this work.

In our study, we propose an end-to-end approach to automatically extract the main classes of symptoms and diseases from clinical notes. The list of these classes of interest corresponds to the MeSH Category C (diseases) headings, such as *infectious diseases, neoplasms, musculoskeletal diseases, digestive diseases, eye diseases, etc.* These classes are of particular interest since they almost directly represent all medical specializations/organ types (see the complete list of classes in Table 2). These classes are called MeSH-C labels in the rest of the article; MeSH-C is the ensemble of all medical concepts in MeSH category C. The MeSH terminology has several advantages: it exists in English and French, is part of the UMLS vocabulary and contains thousands of medical concepts in a tree structure.

This automatic extraction, allowing the targeting of symptoms and pathologies specific to an organ, can be exploited for several medical applications. In the field of pharmacovigilance, it can help to detect side effects of drugs, especially on large databases, where one can automatically retrieve "ocular" or "digestive" or "infectious" disorders present in the EHR without reading any of the reports in person. In

[1] https://www.nlm.nih.gov/mesh/meshhome.html
[2] http://mesh.inserm.fr/FrenchMesh/

the epidemiological domain, one can also automatically extract patients with similar phenotypes, i.e., with the same type of organic lesions and select them as eligible patients for (e.g., a clinical trial or a case/control or cohort study). In clinical practice, clinicians could also analyze or extract past complications for one or more patients. For example, a rheumatologist might be interested in selecting all patients with ocular, renal or skin complications of lupus and could extract them automatically with our method. Furthermore, it is interesting to note that some diseases have multiple labels in the MeSH-C classification (for instance, Diabetes Mellitus type 1 appears in Nutritional and Metabolic Disorder (C18), Endocrine System (C19) and Immune System Diseases (C20)), making it possible to quickly detect such a disease by cross-referencing all labels.

Such examples of natural language processing for the selection of clinical trial cohorts [9] or pharmacovigilance studies [10] have already been proposed but were task specific.

We see this classification problem as the task of finding concept mentions in the texts. If a MeSH-C concept is found in the textual report and if this concept is not negated, hypothetical, or related to someone other than the patient, then we consider that the patient can be labeled by that concept and, thus, by the associated class.

The MeSH terminology category C contains thousands of concepts. It is not possible to find a corpus containing all these concepts. A fully supervised learning strategy is therefore impossible. For this reason, it is necessary to use the terminology itself and the lists of terms associated with the classes to guide the system.

In this article, we focus on French texts. Healthcare reports related to patient care are and will always be written in the local languages of each country; therefore, it is crucial to ensure that advances in artificial intelligence are not limited to English documents. However, this raises additional challenges due to the much more limited resources existing in languages other than English [11], whether in terms of available corpora, thesaurus coverage or availability of pretrained language models.

For this reason, we experimented with different approaches to take advantage of English terminologies and the latest multilingual embedding models.

Our work on this end-to-end classification system for French clinical documents leads to several contributions:

- We trained a named entity recognition system to produce candidate terms for MeSH-C classification; this system is able to discard negated or hypothetical occurrences of concepts, as well as those not related to the patient.

- We used available terminology resources in English and French to reduce the need for annotated data while maintaining good generalizability. The system does not depend on the nature of the documents or on the objective of the final task (e.g., cohort extraction, pharmacovigilance study).
- In the recent dataset DEFT 2021, the first annotated corpus for French MeSH classification [1], we show that our approach leads to good results even without any labeled data for the classification step. This leads to similar results to those obtained with manually optimized handcrafted rules for the DEFT dataset [12].
- We also compare the contribution of multilingual versus monolingual models and resources.[3]

In the next section, we detail the different sets of documents and terms used to train our model and then describe the different steps of the pipeline: model overview, named entity recognition algorithm, gender classification and multilabel classification.

## 2. Material

### *2.1. DEFT 2021 dataset*

The DEFT 2021 dataset [1] consists of 275 clinical cases annotated, among others, with:

- the mention of the sign or symptom and disease type entities
- the characteristics associated with these mentions (e.g., negation, hypothesis, link with someone other than the patient).
- for some of these mentions, the MeSH-C labels were annotated in association with the symptom and disease annotation. Table 1 shows the entire list of possible labels.
- at the document level, an aggregation of these MeSH-C labels (list of all labels occurring at least once in the document).

Figure 1 provides a concrete understanding of all these annotations.

The objective of the task is to perform phenotyping for each case, i.e., to determine the clinical profile of the case by extracting the pathological features described by the MeSH C chapter headings. Table 2 shows the number of documents and words in the dataset, with the split between training and test datasets provided by the challenge organizers.

[3] The code for all experiments described in this paper is available at the following URL: https://github.com/xtannier/MeSH-C_classification

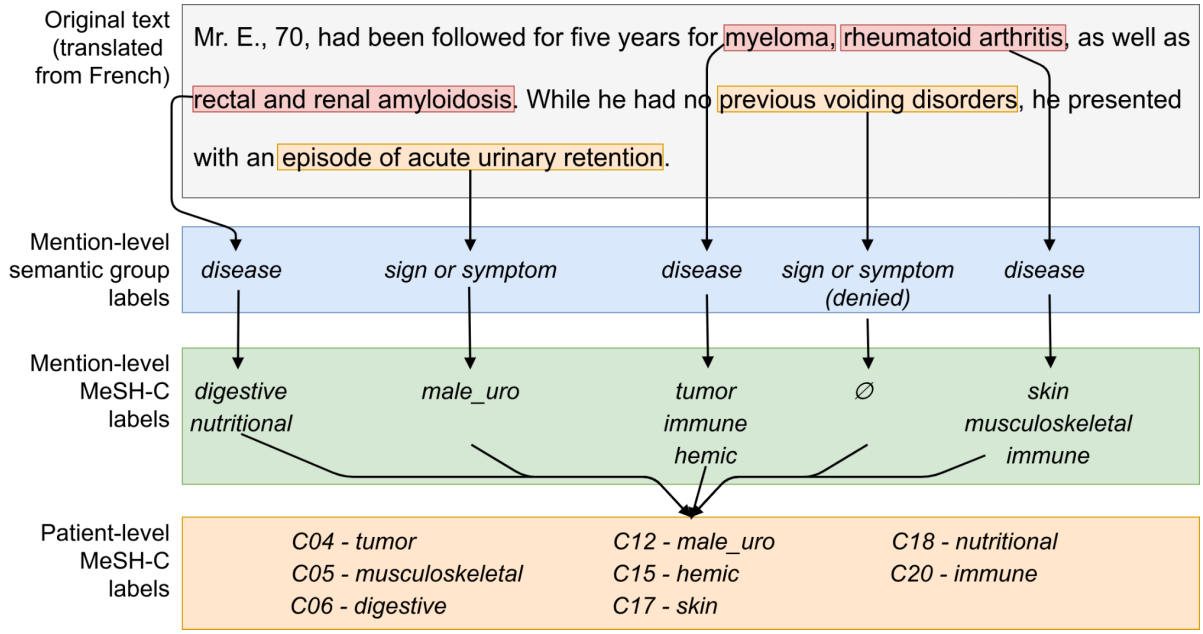


*Figure 1: Annotations provided in the DEFT 2021 corpus. For each medical concept of interest (highlighted), there is an entity label "disease", "sign or symptom" and the negation/hypothesis/link to someone else attribute. Each of the positive entities can be mapped to several MeSH-C chapter headings (corresponding to the "Mention-level MeSH-C label", i.e., the label for each concept). For instance, the extracted mention "myeloma" is labeled with the labels "tumor", "immune" and "hemic". The patient-level MeSH-C labels (bottom) are the labels that we seek to predict for each original text.*

| **MeSH-C level** | **Chapter name** | **Label** |
|---|---|---|
| C01 | Infections | infections |
| C04 | Neoplasms | tumors |
| C05 | Musculoskeletal diseases | musculoskeletal |
| C06 | Digestive System Diseases | digestive |
| C07 | Stomatognathic Diseases | stomatognathic |
| C08 | Respiratory Tract Diseases | respiratory |
| C09 | Otorhinolaryngologic Diseases | ENT |
| C10 | Nervous System Diseases | nervous |
| C11 | Eye Diseases | eye |
| C12 | Male Urogenital Diseases | male_uro* |
| C13 | Female Urogenital Diseases and Pregnancy Complications | female_uro* |
| C14 | Cardiovascular Diseases | cardiovascular |
| C15 | Hemic and Lymphatic Diseases | hemic |
| C16 | Congenital, Hereditary, and Neonatal Diseases and Abnormalities | congenital |
| C17 | Skin and Connective Tissue Diseases | skin |
| C18 | Nutritional and Metabolic Diseases | nutritional |
| C19 | Endocrine System Diseases | endocrine |

| | | |
|---|---|---|
| C20 | Immune System Diseases | immune |
| C21 | Disorders of Environmental Origin | *(missing in the dataset)* |
| C22 | Animal Diseases | *(missing in the dataset)* |
| C23 | Pathological Conditions, Signs and Symptoms | path_sosy |
| C24 | Occupational Diseases | *(missing in the dataset)* |
| C25 | Chemically Induced Disorders | chemical |
| C26 | Wounds and Injuries | injuries |

*Table 1: List of MeSH-C descriptive headings and the short names used in this paper[4]. * male_uro and female_uro are grouped together into a urogen class in our first-step classification.*

| | **Number of documents** | **Number of words** |
|---|---|---|
| Training dataset | 167 | 57,174 |
| Test dataset | 108 | 34,258 |
| Total | 275 | 91,432 |

*Table 2: DEFT 2021 corpus statistics.*

Figure 2 shows the distribution of labels in the training dataset for illustrative purposes. The label *path_sosy* (*Pathological Conditions, Signs and Symptoms*) appears in 141 texts, while *stomatognathic* is present in only 3 texts. The number of labels per document is also presented (median = 3).

Annotations from this training DEFT set will be used to train the named entity recognition NER algorithm, train the multilabel classifier and train the gender classifier (steps 1, 2, and 3 in Figure 3).

[4] To rely on the latest version of the NIH MeSH, we merged the three classes “infectious disease”, “viral disease” and “parasitic disease” into one, which was not the case in the DEFT 2021 challenge. The results and comparison with other participants are still possible since the DEFT test dataset only contained 4 “viral” terms and 1 “parasitic” term. In any case this difference led to an underestimation of our results.

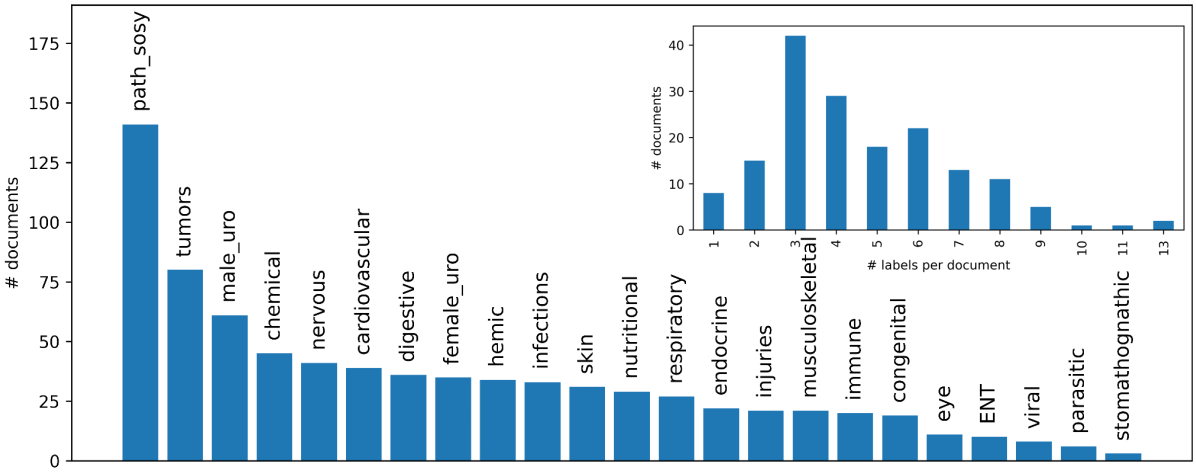


*Figure 2: Distribution of labels in the DEFT training dataset. The y-axis represents the number of documents, and all labels presented are listed in Table 1. The thumbnail represents the number of labels per document.*

### *2.2. Terminological resources (term sets)*

Due to this unbalanced distribution and the small volume of the DEFT 2021 training dataset, we also used terms related to MeSH-C from the UMLS terminology. From now on, we will refer to this resource as the *term set*. The Unified Medical Language System® (UMLS®) brings together three knowledge sources: a metathesaurus, a semantic network and a specialist lexicon and lexical tools. In this work, we only worked on the metathesaurus that unifies concepts from more than 200 vocabularies in the biomedical domain [13]. A *concept* is an entry of a particular terminology and corresponds to a specific notion of this terminology. Each concept is mapped to one or more *terms* (or *synonyms*), possibly in different languages. A unique concept identifier (CUI) is assigned to each concept in the UMLS. For example, the MeSH concept "Breast Neoplasms" (from branch C04 - tumors) is associated with the terms "breast carcinoma", "breast cancer", "mammary carcinoma", "cancer du sein" (French), etc. This MeSH concept is also mapped to its equivalent UMLS concept "Breast Carcinoma" (C0678222), which can lead to other terms from other terminologies.

To obtain synonyms to augment our training term set, we first retrieved the concept unique identifier (CUI) of the MeSH-C terms from the UMLS and then extracted all synonyms related to the CUI in French and English. A complete list of all ontologies used to construct our training term sets can be found in Appendix 1. The bilingual databases were built using PymedTermino2 [14], a Python package that provides easy access to key medical terminologies. We also experimented with an automatic machine translation into French from English terms. For this, we used a state-of-the-art pretrained translation system “opus-mt-en-fr” [15] from the Hugging face library [16].

These term sets will be used to train both the monolingual and multilingual multilabel term classifiers (step 3 in Figure 3).

Table 3 lists all the term sets used, along with the model they trained, synthesizing the three main approaches described above:

- French only (FR): the set of terms in the DEFT dataset and all the French UMLS vocabularies listed in Appendix 1 mapped to MeSH terms.
- multilingual with French and English terms (FR-EN): all terms from the DEFT dataset terms and all the French and English UMLS vocabularies mapped to MeSH terms.
- French terms and translated English terms (FR-tr): the same as the previous set but with all the English terms translated.

| **Multilabel classifier training term sets** | **(number of term/ label couples)** | **Model trained** |
|---|---|---|
| French synonyms (FR) | 42,912 | camemBERT |
| English and French synonyms (FR-EN) | 308,043 | camemBERT and multilingual BERT |
| English Translated and French synonyms (FR-tr) | 209,145 | camemBERT |

*Table 3: Different term sets used for training the multilabel classifier in our experiments. "French synonyms" correspond to the DEFT dataset annotated terms, and all French UMLS vocabularies correspond to MeSH terms. For the "English and French synonyms" set, we added English UMLS vocabularies mapped to MeSH terms. For the "English translated and French synonyms", the same English terms were translated. All models mentioned will be described in Section 3.*

## 3. Methods

### *3.1. System overview*

Figure 3 describes the general architecture of the proposed system. First, a named entity recognition system extracts mentions of the entities: "disease" and "sign or symptom (sosy)" (step 1 in Figure 3). We consider these entities as clues for MeSH-C labels at the patient level. From these mentions, we discard:

- concepts that are negated, hypothetical or associated with someone other than the patient;

- concepts corresponding to negative outcomes (e.g., normal exam, negative analysis).

For this system, we merge the entities “disease” and “sosy” into one to reach the entities to be extracted: “sosy disease”, “sosy disease absent” (i.e., negated), “sosy disease hypothetical”, and “sosy disease non associated” (i.e., relative to another person). This merger is justified, in our opinion, by the semantic proximity of the two entities. Indeed, in MeSH-C, many terms are found in both categories. For example, "*amnesia*", "*amblyopia*", and "*hearing loss*" are cited both in the section "Diseases of the nervous system" and in the section "Pathologic conditions, signs and symptoms". This fusion also has the advantage of grouping the syntactic contexts related to negation, hypothesis, and family medical history to ensure better learning of these non trivial notions. We will show in Section 4 that this assumption is also supported by preliminary results obtained by our NER algorithm, which performed better with than without this fusion step.

In addition, MeSH Chapters C12 (female urogenital diseases) and C13 (male urogenital diseases) can sometimes be distinguished only by the gender of the patient (for example, *anuria, adrenal tumor, pyelonephritis*). Therefore, it is necessary to build a classifier that predicts the gender of the patient from the content of the report (step 2 in Figure 3).

Once the terms of interest are extracted by the system, a classifier predicts the MeSH-C chapters related to each term (step 3). This is, thus, a multilabel classifier (each term can be labeled by none, one or several of the 22 classes represented in the dataset, aggregating female and male in *urogen*). We trained this classifier with the annotated terms of the DEFT training dataset but also with the FR, FR-EN or FR-tr term sets described in Section 2 based on our experiments.

Finally, we aggregate the extracted term-level information to predict document-level classes.

The following sections detail each of these steps.

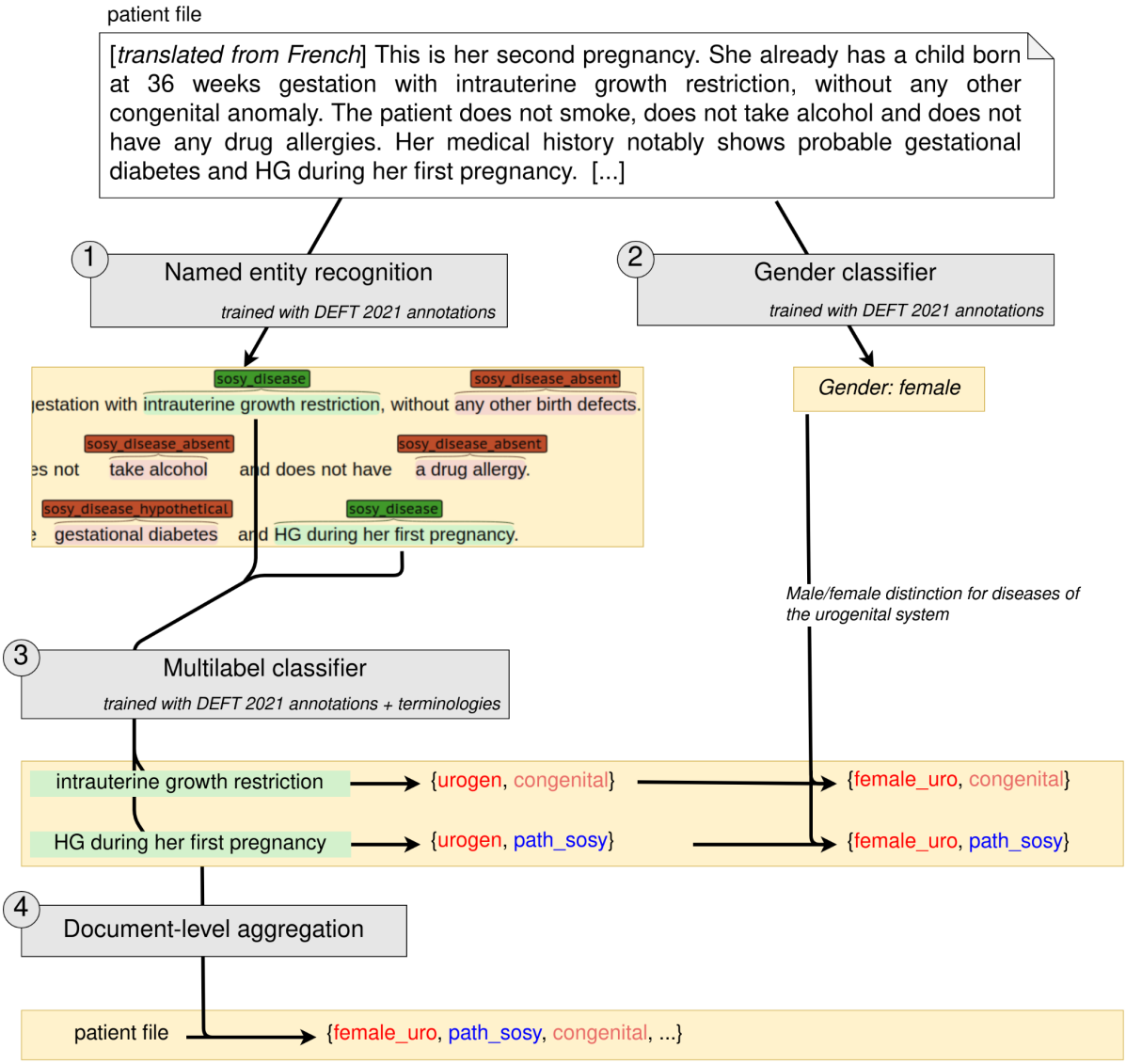


*Figure 3: General system architecture. First, named entity recognition is performed, trained on the annotated DEFT dataset to extract positive medical concepts (1). In parallel, the gender classifier, also trained on the DEFT dataset, determines the written gender of the patient (2). Then, a multilabel classifier assigns a MeSH-C label to each extracted term (3). This multilabel classifier is trained with DEFT annotations and French and English UMLS vocabularies mapped to MeSH-C terms. Finally, all MeSH-C labels are aggregated at the document level for each patient observation (4).*

### *3.2. Named entity recognition (NER)*

The named entity recognition model is illustrated in Figure 4. The model exhaustively keeps scores of all possible spans before prediction; it consists of a BERT transformer [4], which has become a standard way to represent the textual input of a neural network, followed by a bidirectional long short-term memory LSTM [17], similar to the method in [18]. The extracted spans are triplets (begin, end, label).

Each word in the text is first split into word pieces and passed through the transformer. The representations of the last 4 BERT layers are averaged with learnable weights, and the word pieces of a word are max-pooled to build its representation. Char-CNN encoding [19] of the word is concatenated to the max-pooled representation to obtain the word representation.

These word representations are passed through a three-layer highway LSTM with learnable gating weights [20]. We apply the sigmoid function to obtain probabilities. During the prediction, we select the triplets (begin, end, label) that have a probability greater than 0.5.

The model is trained via a binary cross-entropy objective with the Adam optimizer [21]. We use a linear decay learning rate schedule with a 10% warm-up and two initial learning rates: $4.10^{-5}$ for the transformer and $9.10^{-3}$ for the other parameters.

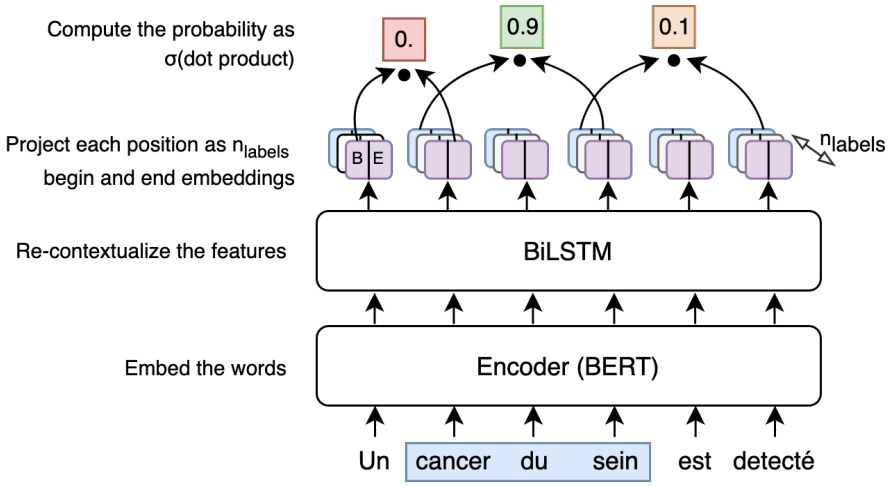


*Figure 4: NER system architecture. Each word is projected into $n_{labels}$ to begin representations and $n_{labels}$ to end representations. Finally, each triplet (B, E, L) is scored as a dot product between the begin representation of label L at position B and the end representation of label L at position E.*

### *3.3. Gender Classification*

All DEFT documents were labeled with gender. To train a classifier to determine gender, we extracted a large number of candidate features and assessed their relevance. An observation of the documents first

determined that in the vast majority of cases, in this type of document, the information describing the patient is found in the first sentence. Therefore, we weighted the variables by their distance from the beginning of the text (according to a weighting function of the order number of the sentence in the document, starting at 1 for the first sentence and decreasing linearly to 0.5 for the last). We then identified the variables that seemed significant during a first qualitative survey of the training corpus.

The most significant feature is (1) the gender of the word patient (in French, "*patient*" is a male while "*patiente*" is a female). The other significant features are, in order of importance: (2) the gender of adjectives applied to humans; (3) The number of occurrences of morphemes referring to sex-specific biological or medical concepts (e.g., peni-, uter-, testi-, vagin-, with a list built from MeSH terms, made available with the code; (4) The gender of civil honorifics ("M. ", " Mr", "Mr"), (" Ms", " Ms. "); (5) The gender of common nouns frequently used to designate a human individual (woman, man, child...); (6) The gender of personal third-person singular pronouns used in the text; (7) The explicit indication of gender.; and (8) the gender of first names, determined from an INSEE[5] reference list of the most frequently given first names in France and of the associated gender.

We extracted the morphosyntactic categories (POS, gender, number) and the syntactic dependencies using the stanza library [22].

We trained a supervised classifier, AdaBoost [23], based on these data to determine the gender prediction function from a text document.

To validate this approach, we trained the classifier on 80% of the provided training data and validated it on a 20% set.

### *3.4. Multilabel classification*

We perform a preliminary filter on the NER output to remove the physiological findings (normal exam, negative analysis[6]). Indeed, these items are often annotated as *sosy* in the DEFT dataset but should not result in a MeSH-C annotation, since MeSH-C classification focuses only on pathological information. For example, "negative HIV serology" or "normal cardiovascular examination" were annotated as "signs and symptoms" in the DEFT dataset. These terms do not correspond to a pathological condition or disease and therefore do not belong to the MeSH-C classification, thus needing to be removed. This filtering is provided by simple regular expressions. An example of this filtering is shown in Appendix 2. This is a minor step different from the negation detection performed by the NER step.

[5] https://www.insee.fr/fr/statistiques/2540004?sommaire=4767262

[6] This is different from negated or hypothetical concepts, in which processing is included into the supervised NER system as described in Section 3.1.

The MeSH-C classification model consists of a pretrained transformer [24] including a final linear output layer. We used either BERT embeddings [4] trained on French data only (CamemBERT [25], model camembert-large) or a multilingual "bert-base-multilingual-cased", both from the HuggingFace library [16]. To enable a multilabel classification, the loss function is the binary cross-entropy, summed over all classes. We used an Adam optimizer [21] with a linear decreasing training step, starting at $1.10^{-5}$. For the prediction, the scores are calculated by the sigmoid function as output.

We used the terminology training sets shown in Table 3 to have the classifier learn to map each entity extracted by NER to its label(s).

We used a 20% validation set (see next section) to choose the best number of epochs and the logit threshold above which a class is positive. The threshold retained for the final prediction maximizes the precision score on the validation set, which leads to better preliminary results. This metric is preferred to the F1 score because the document-level step (step 4 in Figure 3) aggregates possibly redundant information, which mechanically increases recall.

### *3.5. Validation set*

Given the unbalanced representation of each label in the DEFT training dataset (see Figure 2), we chose to build the validation term set with the same class distribution as the DEFT training dataset (as opposed to a random selection which would have led to a distribution similar to the classes inside the UMLS, i.e., unrepresentative of the real documents). Once the best model is selected and the threshold is computed on the validation term set, we use a last step of fine-tuning the classification model for 10 epochs on the validation term set. This last step enables the model to "see" the whole vocabulary at least once.

### *3.6. Experimental setups*

Our set of experiments aims to show how the volume and the language of the terminologies used influence the results. Thus, we propose the results of the system trained on the three term sets described in Section 2.1 (i.e., "FR", "FR-EN", "FR-tr"). For the bilingual FR-EN training term set, we compare two pretrained embeddings: the French CamemBERT model (*camembert-large*) and a multilingual BERT (*bert-multingual-base*; note that there is no "large" multilingual BERT available).

We also compare our results to those of other DEFT participants: a system based on a list of terms manually curated specifically for the DEFT dataset [12], a direct multilabel classification system, i.e., taking the entire text as input, without using the intermediate notion of concept mention [26].

Finally, we performed ablation studies to estimate the impact of the different steps in our system:

- As a gold standard reference for the NER model, the DEFT organizers provide the annotations for the entity types “disease” and “sosy" in the test dataset, enabling us to assess separately the NER performances. For each experiment, we then add a run called “gold mentions”, which uses gold standard named entities instead of the step 1 NER system.
- We also provide results without the final fine-tuning on the validation set (“no FT”).
- Finally, we removed part of the FR-EN term set from the DEFT training dataset to show the results obtained in an unsupervised setup (i.e., only terms from terminologies, none from a human annotation). We called this run “FR-EN no DEFT”.

We evaluated our system using three scores for training, validation and test performance: microprecision, microrecall and micro-F1 score, the most common metrics for multilabel classification. All scores presented in this paper are the average of 5 runs performed with different random seeds to mitigate the effect of initialization and training order.

We also provide carbon footprint estimates for each configuration, as provided by the CarbonTracker tool [27].[7]

## 4. Results

The results are presented in Table 4, where our main runs constitute the runs with our “end-to-end” algorithm: our NER system associated with different classifiers. For each different experimental setup, we show the average score results over 5 runs. Our best results are obtained with the bilingual approach with an F1 score of 0.811 for NER extraction and an F1 score of 0.819 for Gold mentions. The last fine-tuning step improves the F1-score by 1.1 percentage points on average over the 5 experiments (from +0.007 to +0.016).

Note that the threshold selected for classification from the sigmoid output was almost always the same (0.99), which is a good outcome for the robustness of the system.

We also evaluated the performances of intermediate steps 1, NER and 2, gender classification. The NER system detects the “disease, sign and symptom” mentions, excluding the negation, hypothesis and information not related to the patient, with a precision of 0.93 and a recall of 0.88 (F1 score: 0.90). As mentioned in Section 3.1, the NER system detected the merged mentions of “sign and symptom” and “disease”, improving the F1 score by 0.1 on the validation set. The gender classification obtains a perfect score (no error).

[7] Note that these estimates remain very approximate, taking into account neither the execution environment nor the method of energy production at the place of the experiments. CarbonTracker computes its estimates by using the average carbon intensity in the European Union in 2017.

For the DEFT challenge, we initially only used a restricted term set for the classifier, containing only the name of each concept in French, without synonyms, leading to a training term set of 9,363 terms. The official results obtained were F = 0.770 with our NER extraction and a CamemBERT-large classifier and F = 0.775 with the Gold mentions.

We also compared the different carbon footprints: interestingly, the multilingual BERT with 110 million parameters has a much lower approximate carbon footprint than the CamemBERT-large, which has a total of 340 million parameters to train. We also see that the carbon footprint is directly related to the size of the training term set, even though the number of training epochs required is higher for smaller term sets.

| | Recall | Precision | F1 | Carbon footprint (eq $CO_2$) |
|---|---|---|---|---|
| Our main runs | (averaged over 5 runs) | | | |
| FR | 0.801 | 0.812 | 0.807 | 507 g |
| FR-EN (CamemBERT) | 0.832 | 0.788 | 0.809 | 1300 g |
| FR-EN (multilingual BERT) | 0.809 | 0.814 | **0.811** | 239 g |
| FR-tr | 0.833 | 0.763 | 0.797 | 957 g |
| Ablation runs (tradeoff with the main run) | (averaged over 5 runs) | | | |
| FR-EN no DEFT (unsupervised) | 0.828 (-0.4) | 0.688 (-10) | 0.752 (-5.7) | 916 g |
| Gold mentions - FR | 0.813 (+1.2) | 0.809 (-0.3) | 0.811 (+0.4) | |
| Gold mentions - FR-EN (CamemBERT) | 0.847 (+1.5) | 0.793 (+0.5) | 0.819 (+0.8) | |
| Gold mentions - FR-EN (mult. BERT) | 0.815 (+0.6) | 0.811 (-0.3) | 0.813 (+0.2) | |
| Gold mentions - FR-tr | 0.851 (+1.8) | 0.770 (+0.7) | 0.806 (+0.9) | |
| No FT - FR | 0.800 (-0.1) | 0.786 (-2.6) | 0.791 (-1.6) | |
| No FT - FR-EN (CamemBERT) | 0.835 (+0.3) | 0.761 (-2.7) | 0.796 (-1.3) | |
| No FT - FR-EN (mult. BERT) | 0.812 (+0.3) | 0.789 (+0.1) | 0.800 (-1.1) | |
| No FT - FR-tr | 0.839 (+0.6) | 0.746 (-1.7) | 0.790 (-0.7) | |
| Other DEFT participants systems | | | | |
| Manually curated list [12] | 0.750 | 0.888 | 0.814 | |
| Document classification [26] | 0.730 | 0.558 | 0.633 | |

*Table 4: Results of our different experimental setups. The names of the runs are detailed in the previous section. "Gold mentions" uses gold standard named entities (*i.e., *manually annotated) instead of the step 1 NER system. "No FT" corresponds to the results without the final fine-tuning on the validation set.*

Examples of multilabel misclassification are shown in Appendix 3 with the model trained on the translated terms (FR-tr). Examples of erroneous results include "*fever at 39.1 degrees C*" mislabeled "*infections*" (Line 1); "*ureteral valve in the form of an endoluminal transverse fold…* " mislabeled "*cardiovascular", most* likely because of the terms "*valve*" and "*endoluminal*" (Line 2); and "*proliferation index assessed by anti-ki67 antibodies is high*" mislabeled "*immune*", most likely because of the term "*antibodies*".

In addition to the expected label-level results for the shared task corresponding to the objective of our work, we also calculated the number of patients in the dataset for whom we were able to correctly assign all labels. These results range from 32.3% with the worst of our model to 46% with the best. These results are to be expected given the large number of labels to be found in a case (see Figure 2).

Table 5 shows the results for each class with our best model (i.e., multilingual BERT on the French and English term sets). We can see that our system gives homogenous results from one class to another even if the initial distribution is very heterogeneous.

| | Recall | Precision | F1 |
|---|---|---|---|
| Results for each class | | | |
| injuries | 0.684 | 0.722 | 0.703 |
| cardiovascular | 0.926 | 0.735 | 0.820 |
| chemical | 0.636 | 0.700 | 0.667 |
| digestive | 0.864 | 0.613 | 0.717 |
| endocrine | 0.786 | 0.786 | 0.786 |
| path_sosy | 0.960 | 0.951 | 0.956 |
| female_uro | 1.000 | 0.842 | 0.914 |
| congenital | 0.500 | 0.500 | 0.500 |
| hemic | 0.920 | 0.719 | 0.807 |
| male_uro | 1.000 | 0.947 | 0.973 |
| immune | 0.636 | 0.778 | 0.700 |
| infections | 0.704 | 0.679 | 0.691 |
| nervous | 0.717 | 0.825 | 0.767 |
| nutritional | 0.870 | 0.833 | 0.851 |
| eye | 0.667 | 0.857 | 0.750 |
| musculoskeletal | 0.773 | 0.810 | 0.791 |
| skin | 0.812 | 0.619 | 0.703 |
| respiratory | 0.882 | 0.882 | 0.882 |
| stomatognathic | 1.000 | 0.429 | 0.600 |
| tumors | 0.824 | 0.875 | 0.848 |
| GLOBAL EVALUATION | 0.840 | 0.804 | 0.821 |

*Table 5: Results class by class with the best model.*

## 5. Discussion

Although the differences between the four main runs are not very high, it is interesting to note that the joint use of terms in both languages with multilingual embeddings is the most efficient. It is particularly noteworthy that a monolingual space with translated terms performs worse than a multilingual space. This is especially true since the French model used is a “large” model (340 M parameters), while the

multilingual model is a “base” model (110 M parameters). The large models generally outperform the base models in almost all tasks.

*5.1. Comparison of the different experiments*

The results obtained by the multilingual version show that our method could easily be adapted to any other similar language to obtain better performance by taking advantage of the large vocabularies of biomedical concepts of UMLS in English.

It is encouraging to see that the unsupervised setup (“FR-EN No DEFT”) leads to an acceptable F1-score of 0.75, showing that it is possible to obtain reasonable results without any annotated data. This observation is also reinforced by the fact that the results per class are relatively similar, with few exceptions, as shown in Table 5. This would not have been the case in a classical supervised learning approach, where an expected result is that underrepresented classes obtain much worse results than the others.

It is also interesting to observe that the use of gold-standard mentions (experiments “Gold-mentions”) increases the overall results by only a small margin. The NER results are not perfect (F1 = 0.90), but this small difference can be explained by the fact that the redundancy of mentions in a document can help erase some NER errors through the document-level aggregation step.

Finally, our best models lead to results very similar to those of hand-curated terminology matching [12][8], with a much better generalization potential. In that study, the authors used the MeSH lexicon and manually processed this lexicon, removing terms leading to false negatives and positives in the training corpus. Our training resources can be built quickly by a few queries in the UMLS database without correction, which makes our approach easily adaptable to other classes or languages. As we have shown, it can even run with decent performance without any annotated data, while they are needed for curating a terminology through a trial-and-error methodology.

Because our algorithm is based on the MeSH-C classification, we had to determine the gender of the patient as an intermediate step. This has two major drawbacks. First, gender as a social construct is

[8] To rely on the latest version of the NIH MeSH, we merged the three classes “infectious disease”, “viral disease” and “parasitic disease” into one, leading our results to be underestimated when compared to the original benchmark. However, with only 5 occurrences of “parasitic disease” and “viral disease” in the test set, this underestimation is marginal.

used to determine a biological trait. Second, it does not address intersex or transgender urological or gynecological issues and may lead to sexual reductionism, as described in [28].

In other experiments, inspired by the performance of the BioBERT [29] and clinicalBERT [30] models, we tried to fine-tune the CamemBERT-large language model on the 4,000 French open access biomedical articles on EuropePMC[9], but this did not result in major improvements. This is probably because the CamemBERT-large model is already trained on a large volume of heterogeneous data. Moreover, the volume of 4,000 articles was probably insufficient to allow a real contribution to the model. Unfortunately, as with most languages except for English, there are often too few accessible biomedical resources available to improve performance, which justifies the need to use multilingual models.

This work enables one to automatically detect medical categories from clinical narratives. The next step of this work will be to directly create a representation of the patient from the embeddings of the labeled terms. For instance, in the case of a text explaining that a patient with glaucoma has the flu, the labels with our algorithm would be 'eye' and 'infection', and a relevant representation of the patient would be the concatenation of the 'glaucoma' and 'flu' embeddings. This representation can lead to a finer phenotyping of the patient and enables, for example, computation of the similarity of patients. This representation is inspired by the "Deep-Patient" model [31], except that our features are based on transformer embedding and filtered by a classification algorithm.

### *5.2. Comparison with previous work*

As mentioned above, the extraction of the main pathological characteristics of a clinical case corresponds to a phenotyping of the patient. In recent years, several studies have been carried out on the phenotyping of patients from the EHR.

Gerhmann et al. [32] compared deep learning- and concept extraction-based methods for patient phenotyping in English. More precisely, they assess the performance of convolutional neural networks for narrative-based patient phenotyping, comparing it to cTAKES (Mayo clinical Text Analysis and Knowledge Extraction System) [33] to predict 10 disorders. They obtained an improvement of the F1 score ranging from 2 to 26 points (except for one disorder).

Yang et al. [34] proposed a method combining a CNN-based deep learning neural network and natural language processing to predict ten disorders from English clinical narratives. The CNN processes inputs at the word and sentence levels. Similar to our approach, the authors used different sample sizes of the

[9] https://europepmc.org/

training dataset. The authors also used word2vec [35] word embeddings. The obtained results range from an F1 score of 63% for “Chronic Pain” to 86% for “Depression”.

Aside from the other participants in the DEFT 2021 challenge, no other articles have the exact same objective. However, Weng et al. [36] proposed a classification of clinical notes into medical subdomains. Their Natural Language Processing (NLP) pipeline is based on cTAKES [33] and on the UMLS metathesaurus. The best performing algorithm was a convolutional recurrent neural network with neural word embeddings (fastText [37]), with AUCs of 0.975 and 0.995, respectively, for each of their datasets, and F1 scores of 0.845 and 0.870. Using two different datasets, the overall prediction portability from one dataset to another gave an F1 score of 0.7.

Compared to the abovementioned studies, the originality of our work lies in the fact that our classification is not as broad as the medical subdomain classification task or as narrow as the disease classification task but rather in between, enabling the rapid detection of pathological characteristics with good performance in the French language using a multilingual system.

[38] shares the same objective of exploring the possibility of combining multilingual resources in the same space for concept classification. Their application task is different, but their conclusions on this topic align with ours: they also found that a multilingual approach performs better than a translated approach and constitutes a good alternative for languages other than English. However, the variety of English models remains much higher, especially for the sciences and medical fields (BioBERT [29], clinicalBERT [30]), and annotated data remain massively more important in English, thus requiring NLP in other languages to continue to progress.

## 6. Conclusion

In this work, we proposed a multilabel classification of clinical narratives with all the headings of MeSH-C chapters, leading to a 22-label classification with good performance. This multilabel classification allows rapid extraction of the pathological domain for the phenotyping of patients. We tested several vocabularies to train our classifiers. Interestingly, our bilingual approach with UMLS English and French vocabularies leads to the best results, suggesting that our method could be used for any other similar language.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**

[1] C. Grouin, N. Grabar & G. Illouz, Classification de cas cliniques et évaluation automatique de réponses d'étudiants: présentation de la campagne DEFT 2021 (Clinical cases classification and automatic evaluation of student answers: Presentation of the DEFT 2021 Challenge), Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles, Atelier DÉfi Fouille de Textes (DEFT)(2021) 1-13.

[2] J. Pennington, R. Socher R, CD. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (2014) 1532-1543.

[3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365, 2018.

[4] J. Delvin, W. Changm., Leek. & Toutanovak, BERT : Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human LanguageTechnologies - Proceedings of the Conference 1 (2019) 4171-4186.

[5] D. Zhang, J. Thadajarassiri, C. Sen, E. Rundensteiner, Time-Aware Transformer-based Network for Clinical Notes Series Prediction, Machine Learning for Healthcare Conference (2020) 566-588.

[6] S. Soni, K. Roberts, Patient Cohort Retrieval using Transformer Language Models, AMIA Annual Symposium Proceedings (2020) 1150.

[7] J. Feng, C. Shaib, F. Rudzicz, Explainable clinical decision support from text, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020) 1478-1489.

[8] J. Shang, T. Ma, C. Xiao, J. Sun, Pre-training of graph augmented transformers for medication recommendation, arXiv preprint arXiv:1906.00346, 2019.

[9] L. Chen, Y. Gu, X. Ji, C. Lou, Z. Sun, H. Li, Y. Gao, Y. Huang, Clinical trial cohort selection based on multi-level rule-based natural language processing system, Journal of the American Medical Informatics Association 11 (2019) 1218-1226.

[10] S. Bayer, C. Clark, O. Dang, J. Aberdeen, S. Brajovic, K. Swank, L. Hirschman, R. Ball, ADE Eval: An Evaluation of Text Processing Systems for Adverse Event Extraction from Drug Labels for Pharmacovigilance, Drug safety 1 (2021) 83-94.

[11] 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.

[12] N. Hiot, A.L. Minard, and F. Badin, DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques, Traitement Automatique des Langues Naturelles ATALA (2021) 41-53.

[13] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res 1;32 (Database issue) (2004) 267-270.

[14] J.B. Lamy, A. Venot, C. Duclos,PyMedTermino: an open-source generic API for advanced terminology services, Studies in health technology and informatics 210 (2015) 924-928.

[15] J. Tiedemann, S. Thottingal, OPUS-MT–building open translation services for the world, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (2020) 479-480.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac,..., and A. M. Rush, Transformers : State-of-the-art natural language processing, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations, Online: Association for Computational Linguistics (2020) 38-45.

[17] S. Hochreiter and J. Schmidhuberj, Long Short-Term Memory, Neural Computation 9(8) (1997) 1735-1780.

[18] J. Yu, B. Bohnet and M. Poesio, Named Entity Recognition as Dependency Parsing, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics (2020) 470-476.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural Architectures for Named Entity Recognition, Proceedings of the 2016 Conference of theNorth American Chapter of the Association for Computational Linguistics : Human LanguageTechnologies, Stroudsburg, PA, USA : Association for Computational Linguistics (2016) 260–270.

[20] K. Jaeyoung, E.K. Mostafa and L. Jungwon, Residual LSTM: Design of a Deep Recurrent Architecture for Distant Speech Recognition, Interspeech (2017).

[21] D.P. Kingma, J. Ba, Y. Bengio and Y. LeCun, Adam: A method for stochastic optimization, 3rd International Conference on Learning Representations, ICLR (2015).

[22] P. Qi, Y Zhang, Y. Zhang, J. Bolton and C.D. Manning, Stanza: A Python natural language processing toolkit for many human languages, Proceedings of the 58th Annual Meetingof the Association for Computational Linguistics : System Demonstrations (2020).

[23] Y. Freund and R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55(1) (1997) 119-139.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, …, and I. Polosukhin, Attention is all you need, Advances in neural information processing systems (2017) 5998-6008.

[25] L. Martin, B. Muller, P.J.O. Suárez, Y. Dupont, L. Romary, E. De La Clergerie, D. Seddah and B. Sagot, CamemBERT : a tasty French language model, Proceedings of the 58th Annual Meeting of the

Association for Computational Linguistics, Online : Association for Computational Linguistics (2020) 7203–7219.

[26] M.B. Billami, L. Nicolaieff, C. Gosset and C. Bortolaso, Participation de Berger-Levrault (BL.Research) à DEFT 2021 : de l'apprentissage des seuils de validation à la classification multi-labels de documents, Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT) (2021) 82-94.

[27] L.F. Wolff Anthony, B. Kanding and R. Selvan, Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, CML Workshop on Challenges in Deploying and Monitoring Machine Learning Systems (2020).

[28] F. Hamidi, M.K. Scheuerman and S.M. Branham, Gender Recognition or Gender Reductionism ? The Social Implications of Embedded Gender Recognition Systems, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery 8 (2018) 1-13.

[29] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36(4) (2020) 1234-1240.

[30] E. Alsentzer, J.R. Murphy, W. Boag, W.H. Weng, D. Jin, T. Naumann and M. McDermott, Publicly available clinical BERT embeddings, arXiv preprint arXiv:1904.03323, 2019.

[31] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific reports 6(1) (2016) 1-10.

[32] S. Gehrmann, F. Dernoncourt, Y. Li, E.T. Carlson, J.T. Wu, J. Welt, .. and L.A. Celi, Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, *PloS one*, *13*(2) (2018) e0192360.

[33] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, Journal of the American Medical Informatics Association, 17(5) (2010) 507-513.

[34] Z. Yang, M. Dehmer, O. Yli-Harja and F. Emmert-Streib, Combining deep learning with token selection for patient phenotyping from electronic health record, Scientific Reports 10(1) (2020) 1-18.

[35] T. Mikolov, W. Yih, and G. Zweig, Linguistic regularities in continuous space word representations, Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (2013) 746-751.

[36] W.H. Weng, K.B. Wagholikar, A.T. McCray, P. Szolovits, H.C. Chueh, Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach, BMC medical informatics and decision making 17(1) (2017) 1-13.

[37] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135-146.

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

*Appendix 1: UMLS vocabularies used for the training set*[10]

| UMLS abbreviation | Vocabulary | Language |
|---|---|---|
| BI | Beth Israel Problem List | EN |
| CHV | Consumer Health Vocabulary | EN |
| CSP | CRISP Thesaurus | EN |
| CST | COSTART | EN |
| CVX | Vaccines Administered | EN |
| DRUGBANK | DrugBank | EN |
| HPO | Human Phenotype Ontology | EN |
| ICD10 | International Classification of Diseases and Related Health Problems, Tenth Revision | EN |
| ICD10CM | International Classification of Diseases, Tenth Revision, Clinical Modification | EN |
| ICPC2P | ICPC-2 PLUS | EN |
| ICPCFRE | ICPC French | FR |
| LNC | LOINC | EN |
| LNC-FR-FR | LOINC Linguistic Variant - French, France | FR |
| MDR | MedDRA | EN |
| MDRFRE | MedDRA French | FR |
| MEDCIN | MEDCIN | EN |
| MMX | Micromedex | EN |
| MSH | MeSH | EN |
| MSHFRE | MeSH French | FR |
| MTHICD9 | ICD-9-CM Entry Terms | EN |

[10] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

| UMLS abbreviation | Vocabulary | Language |
|---|---|---|
| BI | Beth Israel Problem List | EN |
| CHV | Consumer Health Vocabulary | EN |
| CSP | CRISP Thesaurus | EN |
| CST | COSTART | EN |
| CVX | Vaccines Administered | EN |
| DRUGBANK | DrugBank | EN |
| MTHMSTFRE | Minimal Standard Terminology French (UMLS) | FR |
| NCBI | NCBI Taxonomy | EN |
| NCI | NCI Thesaurus | EN |
| NCI_CDISC | CDISC Terminology | EN |
| NCI_CTRP | Clinical Trials Reporting Program Terms | EN |
| NDDF | FDB MedKnowledge | EN |
| OMIM | Online Mendelian Inheritance in Man | EN |
| PDQ | Physician Data Query | EN |
| RCD | Read Codes | EN |
| SNMI | SNOMED Intl 1998 | EN |
| SNOMEDCT_US | SNOMED CT, US Edition | EN |
| SRC | Source Terminology Names (UMLS) | EN |
| WHO | WHOART | EN |
| WHOFRE | WHOART French | FR |

*Appendix 2: Code used to filter "normal" or "negative" terms.*

```
indexNorm = df2[(df2['term'].str.contains("normaux")) | (df2['term'].str.contains("normales"))
        | (df2['term'].str.contains("normal")) | (df2['term'].str.contains("normale"))].index
        df2.drop(indexNorm, inplace=True)

        indexNeg = df2[(df2['term'].str.contains("négatif")) | (df2['term'].str.contains("négative"))
        | (df2['term'].str.contains("négatifs")) | (df2['term'].str.contains("négatives"))].index
        df2.drop(indexNeg, inplace=True)
```

*Appendix 3: Examples of term misclassification.*

| | ***Wrong label (false-positive)*** | ***term (translated from French)*** | ***source*** |
|---|---|---|---|
| **1** | *infections* | *fever at 39.1 degree C* | *filepdf-292-3-cas.ann* |
| **2** | *cardiovascular* | *ureteral valve in the form of an endoluminal transverse fold including smooth muscle fibers throughout its surface* | *filepdf-156-1-cas.ann* |
| **3** | *cardiovascular* | *heart rate at 80 per minute* | *filepdf-71-2-cas.ann* |
| **4** | *skin* | *voluntary ingestion of a black shoe dye* | *filepdf-519-cas.ann* |
| **5** | *musculoskeletal* | *literally from French "lumbar contact", corresponding to the palpation of an enlarged kidney in the back* | *filepdf-184-cas.ann* |
| **6** | *hemic* | *Benign proliferation, formed by both lobules of mature adipocytes and normal hematopoietic tissue* | *filepdf-256-cas.ann* |
| **7** | *immune* | *proliferation index assessed by anti-ki 67 antibodies is high* | *filepdf-42-cas.ann* |
| **8** | *digestive* | *sphincter insufficiency* | *filepdf-54-2-cas.ann* |