Mini Project - Class of Machine Learning for NLP

Quentin Marret
ENSAE Paris
quentin.marret@ensae.fr

Abstract

This study investigates multilingual multilabel classification of clinical case reports using Medical Subject Headings (MeSH) categories. We explore and compare classical and transformer-based NLP approaches on a large open-access corpus of over 10,000 English clinical cases, with the goal of automatically predicting MeSH-C "disease" categories. Our models include a naive One-vs-Rest classifier with TF-IDF features, BioBERT pretrained on biomedical literature, and a multilingual BERT model. To address low-resource scenarios in French, we augment the training set with synthetic clinical cases generated using Gemini and evaluate performance on machine-translated French cases. Results show that the naive TF-IDF model performs competitively with transformer models on English data (micro F1 = 0.70), while the multilingual BERT model demonstrates strong generalization to French cases (micro F1 = 0.57). These findings confirm the effectiveness of multilingual language models for cross-lingual clinical classification and validate the use of synthetic data generation as a viable strategy in low-resource medical NLP settings.

1 Introduction

The rapid digitization of healthcare data and proliferation of electronic health records (EHRs) have led to an exponential increase in the volume of available medical information. While this vast repository of medical data holds significant promise, manually analyzing and extracting structured clinical information from unstructured text has become impractical. Consequently, Natural Language Processing (NLP) has emerged as a key approach for automating the extraction of essential medical information, such as physiological characteristics of patients, symptoms, diagnoses, and prescribed treatments. NLP systems offer the potential to significantly enhance medical data structuring, indexing, and retrieval, facilitating advanced medical applications like pharmacovigilance or cohort identification for clinical trials.

2 Brief State-of-the-Art

The extraction of structured medical information has traditionally relied on specialized NLP systems tailored to specific tasks or domains. Historically, these systems frequently employed lexicon dictionaries combined with regular expressions (regex). For instance, rEHR was developed to manage and analyze Electronic Health Records (EHR) data, ctrdata focused on analyzing clinical trial data, and medExtractR was designed specifically for extracting medication-related information. Although these regex-based systems proved effective for straightforward tasks, such as extracting demographic data (e.g., age or gender), they struggled significantly with more complex and nuanced clinical text, including tasks like identifying objects causing injuries or interpreting negations and context-dependent statements (Sciannameo et al., 2024).

Furthermore, traditional NLP systems often faced significant challenges in text classification tasks involving medical text. Text classification involves categorizing an entire document by assigning one

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

or more predefined labels. This has broad applications in the biomedical domain, such as biomedical literature indexing, automatic diagnosis code assignment, classification of public health-related tweets, and patient safety report categorization.

A representative example of text classification in clinical case reports is Medical Subject Headings (MeSH) indexing. This task involves assigning MeSH labels that best correspond to the symptoms described in clinical cases or identifying relevant medical specialties related to the presented cases. The MeSH terminology is structured hierarchically as a tree; thus, labels become increasingly specific and detailed as one moves down the branches. Historically, in the early 2010s, this indexing was primarily conducted manually by experts at the National Library of Medicine (NLM) (Shengwen Peng et al., 2016).

Automating this process using classical NLP tools, especially regex-based systems, faced three primary challenges:

- The multiclass classification involved a high number of classes, often dozens to thousands.
- The number of labels assigned varied significantly depending on the text content.
- Expressions involving negation could be particularly misleading, especially when relying solely on regex approaches.

To overcome these limitations, recent NLP methodologies have increasingly transitioned towards machine learning (ML) and deep learning approaches. For example, DeepMeSH (Shengwen Peng et al., 2016) leveraged deep semantic representations to enhance large-scale MeSH indexing primarily based on article abstracts, achieving approximately 63% precision. ML-Net (Jingcheng Du et al., 2019) introduced an efficient and scalable multilabel classification system, eliminating the need to construct individual classifiers for each label, thus significantly reducing manual feature engineering. ML-Net reported an F1 score of 0.829 for predicting 10 classes and 0.753 for a prediction task involving 32 classes. Nevertheless, these systems utilized only article abstracts for their classifications and did not leverage full document texts.

More recently, transformer-based approaches like BERTMeSH (Ronghui You et al., 2020) have significantly improved multilabel MeSH indexing performance by incorporating entire medical documents rather than merely abstracts. BERTMeSH achieved around 69% precision, clearly demonstrating the advantage of exploiting deeper, context-rich textual information. Additionally, approaches such as NewsMeSH highlighted the benefits of adjusting MeSH label granularity, showing improved predictive performance when limiting tree depth. Specifically, at a depth of 2, NewsMeSH achieved approximately 74% precision with an F1-score of 55% in annotating medical texts.

3 Proposal and Justification of Experiment

A recent and influential contribution to MeSH classification in low-resource settings is the work of Gérardin et al. (2022), who addressed the challenge of limited annotated training data in French by developing a multilingual pipeline based on the DEFT 2021 dataset. This dataset includes 275 de-identified clinical case reports in French, with only 167 labeled with MeSH-C chapter categories. Their method combined named entity recognition (NER), distant supervision using UMLS-based synonym expansion, and transformer-based classification. They first extracted medically relevant mentions (e.g., diseases, symptoms) using a BERT-based NER model, filtered out negated or hypothetical mentions, and then associated those text spans with MeSH-C categories using synonym lists. These synonym lists, derived from French and English UMLS entries, allowed the construction of over 300,000 training pairs without manual annotation. The classification model was fine-tuned using either CamemBERT or multilingual BERT, with the best configuration reaching a micro-F1 score of 0.811 on the DEFT test set.

Inspired by this approach, our experiment seeks to explore and compare different strategies for MeSH multilabel classification using a larger and more diverse dataset. Since the DEFT dataset is not publicly available, we use the MultiCaRe dataset, which contains over 10,000 full-text clinical case reports in English sourced from open-access PubMed articles. This dataset, in contrast to earlier work that used only abstracts, captures richer clinical narratives. Our goal is to evaluate the impact of model architecture and data representation on MeSH classification performance, using both English and French data where possible.

We compare three models:

- **Naive multilabel classifier (One-vs-Rest with TF-IDF)**: A simple baseline that uses a TF-IDF bag-of-words representation combined with logistic regression trained independently for each label. This model serves to contextualize the gains achieved by more advanced, context-aware NLP architectures.
- **BioBERT**: A domain-specific transformer model pretrained on large-scale biomedical literature. It represents the current state-of-the-art for many clinical NLP tasks and is expected to perform well on MeSH-based classification due to its specialized medical vocabulary and contextual understanding.
- **BERT-base-multilingual-cased**: A bilingual transformer model pretrained on both French and English corpora. This model allows us to assess cross-lingual adaptability and evaluate performance on both original English cases and synthetic French cases.

Unlike Gérardin et al., we feed the entire case reports directly into the model, without using a separate NER step. This decision is motivated by four factors: we lack human-annotated entities; our dataset is significantly larger; BioBERT is already trained to recognize biomedical language; and we aim to compare full-text classification versus mention-level classification.

In a second experiment, we propose to enrich the training data with synthetic French case reports generated using a large language model (LLM), specifically Gemini. Rather than translating existing English cases into French, we generate French clinical cases de novo to simulate real-world physician-authored notes. To ensure comparability with the study by Gérardin et al., we limited the French training dataset to only 167 synthetic cases, mirroring the size of their training set. This approach avoids introducing systematic biases caused by translation artifacts and ensures our method can generalize to truly French-written clinical texts. Furthermore, this LLM-based data augmentation technique is scalable and could be used to generate examples for underrepresented MeSH categories in future applications.

Since we currently lack real French clinical notes for evaluation, we propose to build a test set by translating a subset of the English cases into French. While this does introduce the limitations of machine translation, we assume that the semantic integrity of the cases is preserved well enough for evaluation purposes.

All models were trained using a strategy that optimizes the micro F1 score. To determine the best model configuration, we used a validation set representing 10% of the full dataset, corresponding to 1,072 clinical cases. In addition, all BERT-based models were trained for a fixed number of 10 epochs.

Finally, we assess the models using standard multilabel evaluation metrics similar to those reported in by Gérardin et al. (2022) in their article to allow direct comparison:

- Micro-averaged Precision: Evaluates the proportion of correctly predicted labels among all predicted labels, aggregated over all classes.
- Micro-averaged Recall: Assesses the proportion of correctly predicted labels among all true labels, aggregated over all classes.
- Micro-averaged F1 score: Measures the harmonic mean of micro-precision and micro-recall, providing a balanced view of overall performance across all labels.
- Exact match ratio: Indicates the percentage of clinical cases for which the predicted set of labels exactly matches the true set of labels.

4 Analysis of the Dataset

A total of 27,367 MeSH major category labels from category C ("Disease") were assigned to the 10,718 clinical cases in our dataset, resulting in an average of 2.72 labels per clinical case. We observed a significant imbalance among the 23 MeSH classes[1]. For instance, 2,998 clinical cases were labeled under class C04 ("Neoplasms"), whereas class C21 ("Disorders of Environmental Origin")

[1]As of the 2020 MeSH update, the National Library of Medicine merged "Bacterial Infections and Mycoses" (C01), "Virus Diseases" (C02), and "Parasitic Diseases" (C03) into a single category "Infections" (C01), reducing

had no cases assigned, and class C22 ("Animal Diseases") included only 7 cases. Nevertheless, 20 out of the 23 classes were used to index at least 100 clinical cases, and 13 classes were used to index over 1,000 cases. We believe these disparities fairly represent the overall distribution of major MeSH terms in medical literature. Furthermore, the analysis of correlations between classes reveals that most correlations are negligible. Out of the 231 correlations calculated among the 22 non-empty categories, 208 had absolute values lower than 0.10. Only one correlation exceeded 0.30, specifically between categories C16 ("Congenital, Hereditary, and Neonatal Diseases and Abnormalities") and C20 ("Immune System Diseases"). Thus, we can conclude that category prevalences are largely independent.

Regarding textual characteristics, clinical cases had an average length of 524 words, with 50% of cases (interquartile range Q1-Q3) containing between 300 and 650 words. Among all texts, the most frequently occurring word was "patient" (42,208 occurrences), followed by "showed" (21,150 occurrences), and then "figure." Medical vocabulary appearing in the top 20 most frequent words included "blood" (13,655 occurrences), "examination" (12,687 occurrences), "treatment" (11,966 occurrences), and "cells" (9,616 occurrences) (see Appendix A for details on the top 20 general words).

To more precisely identify medical terms within the texts, a secondary analysis was performed using the SpaCy model `en_core_sci_sm` along with SciSpacy's medical term linker. This analysis, conducted on a random sample of 250 lines from the clinical case texts, revealed logically high frequencies of medical terms such as "patient/patient's" (1,018 occurrences), as well as temporality-related terms—"day," "month," "week," and "year"—which are considered medical by the model, with 370, 287, 190, and 186 occurrences, respectively.

Furthermore, the analysis frequently identified terms closely associated with diagnosis and treatment, such as "treatment," "diagnosis," "diagnose," and "lesion," as well as hospitalization-related terms like "admission" and "hospital." Also prevalent were quantifiers often used in medical assessments, including "level," "increase," "negative," "positive," "decrease," and "figure."

For additional information on the top 20 most frequent medical terms identified in this process, see Appendix A.

Analysis of Results

Table 1: Evaluation results for various models and data configurations

Model and Vectorizer	Trainset	Test Set	F1	Prec.	Rec.	HLoss	EM (%)
One-VS-Rest (2-gram)	English	English	0.70	0.65	0.75	0.07	20.52
One-VS-Rest (2-gram)	French*	French**	0.30	0.31	0.30	0.13	1.33
BioBERT	English	English	0.67	0.75	0.60	0.06	24.72
BERT Multilabel	English	English	0.66	0.74	0.59	0.06	23.79
BERT Multilabel	English French*	French**	0.57	0.68	0.49	0.07	20.00
BERT Multilabel	English	French**	0.52	0.65	0.43	0.07	15.00

* Synthetic data: French clinical cases generated using the Gemini 2.0 Flash model.
** Translated data: French clinical cases obtained by translating English cases using Google Translate.

Several noteworthy insights emerge from our experiments. Firstly, contrary to initial expectations, the naive One-VS-Rest classifier performed similarly to or even slightly better than the large language models BioBERT and BERT Multilabel (micro F1 scores: 0.70 vs. 0.67 and 0.66, respectively). This outcome is likely due to these models primarily leveraging specific vocabulary within clinical cases for classification. For example, the two most predictive words for the category "C07 Respiratory Tract Diseases" were "sars" and "covid." This result suggests that additional syntactic and semantic context provided limited added value for this specific classification task. Nevertheless, it is surprising that the

the number of disease subcategories from 26 to 24. Source: `https://www.nlm.nih.gov/pubs/techbull/nd19/nd19_medline_data_changes_2020.html`

BERT-based models did not notably surpass the naive models, especially considering their potential advantage in better filtering negated or hypothetical mentions. Such mentions can easily mislead models relying solely on vocabulary-based approaches (bag-of-words). These results reinforce the methodology proposed by Gérardin et al., who first extracted medically relevant mentions (e.g., diseases and symptoms) using a BERT-based NER model, filtered out negated or hypothetical mentions, and linked these filtered mentions to MeSH-C categories via synonym lists. Their approach achieved a significantly higher micro F1 score of 0.811 on the DEFT test set compared to our best result of 0.70, obtained using only raw text as input.

The second important observation concerns the models' ability to maintain predictive performance on French clinical cases, despite deliberately restricting the French training set to only 167 synthetic clinical cases generated by the Gemini 2.0 Flash model, mirroring the dataset constraints used by Gérardin et al. Here, the multilingual BERT model clearly outperformed the naive One-VS-Rest approach trained solely on the limited French data. Notably, the multilingual BERT model exhibited only a moderate performance drop when tested on translated French data (micro F1 of 0.57) compared to its performance on English data (micro F1 of 0.66), despite being predominantly trained on English data (8,722 English cases vs. 167 French synthetic cases). This highlights the robust few-shot learning capabilities of multilingual language models (Cahyawijaya et al., 2024).

Finally, the third insight relates to the effectiveness of synthetically generated data. Our results show that incorporating synthetic French cases improved model performance compared to the BERT Multilabel model trained only on English data (micro F1 of 0.57 vs. 0.52). This indicates that the model successfully learned from synthetic data generated by Gemini, validating the approach known as knowledge distillation. Knowledge distillation involves using a larger, powerful "teacher" model to generate labeled data or predictions, subsequently training a smaller "student" model to replicate the teacher's behavior. This method allows the student model to achieve comparable performance more efficiently (Xu et al., 2024). Our experiment validates this approach and underscores its potential for effectively training models in scenarios with limited available data. This is especially relevant in medicine, where constraints related to patient privacy, medical confidentiality, and regulations such as GDPR significantly restrict the creation of large-scale, realistic clinical datasets necessary for training advanced language models.

Conclusion

In this study, we evaluated several approaches for multilingual multilabel classification of clinical case reports using MeSH-C disease categories. Our results demonstrate that a simple TF-IDF-based classifier can perform competitively with transformer-based models on English data, highlighting the strength of lexical cues in clinical narratives. However, transformer models, particularly multilingual BERT, showed superior generalization to French data—especially when augmented with synthetic cases generated via large language models. These findings confirm the promise of multilingual transformers in low-resource settings and illustrate the practical value of synthetic data generation for enhancing model performance. Future work may explore more sophisticated filtering of negated mentions and broader MeSH coverage, as well as validate results on real, non-synthetic French clinical cases.

References

[1] Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-Net: Multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11), 1279–1285. https://doi.org/10.1093/jamia/ocz085

[2] Gérardin, C., Wajsbürt, P., Vaillant, P., Bellamine, A., Carrat, F., & Tannier, X. (2022). Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, 128, 102311. https://doi.org/10.1016/j.artmed.2022.102311

[3] Offidani, M. A. N., & Delrieux, C. A. (2023). Dataset of clinical cases, images, image labels and captions from open access case reports from PubMed Central (1990–2023). *Data in Brief*, 52, 110008. https://doi.org/10.1016/j.dib.2023.110008

[4] Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., & Zhu, S. (2016). DeepMeSH: Deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12), i70–i79. https://doi.org/10.1093/bioinformatics/btw294

[5] You, R., Liu, Y., Mamitsuka, H., & Zhu, S. (2020). BERTMeSH: Deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, 37(5), 684–692. https://doi.org/10.1093/bioinformatics/btaa837

[6] Sciannameo, V., Pagliari, D. J., Urru, S., Grimaldi, P., Ocagli, H., Ahsani-Nasab, S., Comoretto, R. I., Gregori, D., & Berchialla, P. (2024). Information extraction from medical case reports using OpenAI InstructGPT. *Computer Methods and Programs in Biomedicine*, 255, 108326. https://doi.org/10.1016/j.cmpb.2024.108326

[7] Cahyawijaya, S., Lovenia, H., & Fung, P. (2024). LLMs are Few-Shot In-Context Low-Resource Language Learners. *Proceedings of NAACL 2024*, 405–433. https://doi.org/10.18653/v1/2024.naacl-long.24

[8] Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., Yan, B., & Chen, Y. (2024). Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3699518

A Appendix / supplemental material

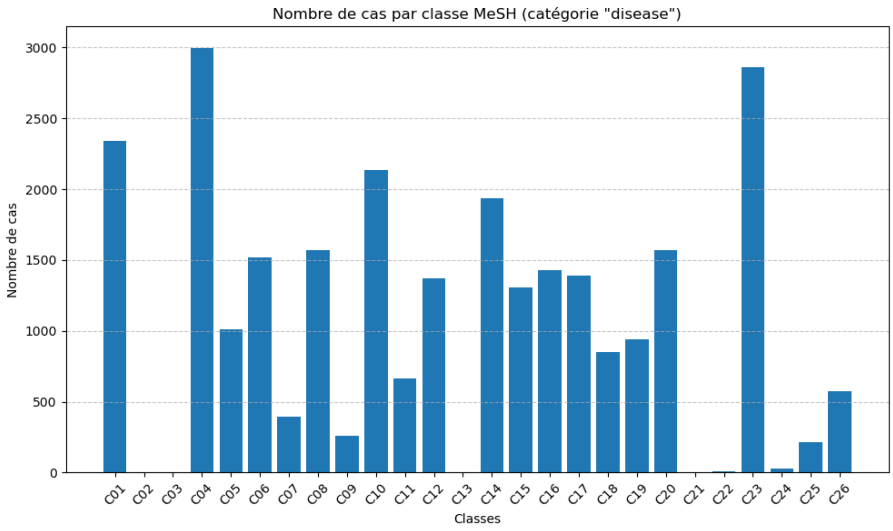

Figure 1: Number of cases per MeSH category (subcategory "disease").

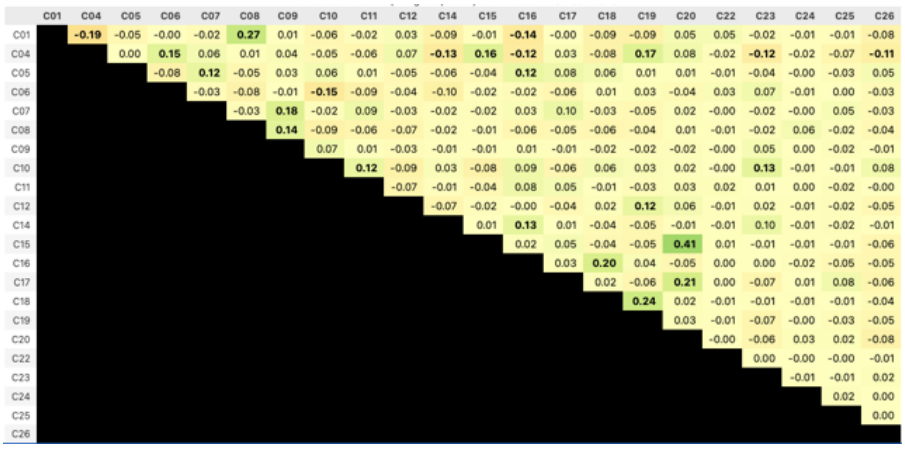

Figure 2: Correlation matrix between MeSH categories C01 to C26.

Table 2: Top 20 most frequent general terms in the full corpus.

Word	Frequency
patient	42,208
showed	21,150
figure	19,926
normal	17,939
revealed	16,054
left	15,717
right	15,197
day	14,040
blood	13,655
examination	12,687
treatment	11,966
year	11,738
performed	10,850
old	10,724
fig	10,646
history	10,253
months	10,033
cells	9,616
days	8,983
also	8,185

Table 3: Top 20 medical terms identified using SciSpacy on a sample of 250 clinical case lines.

Word	Frequency
patient	936
day	370
month	287
increase	200
treatment	197
week	190
year	186
level	180
negative	179
diagnosis	154
figure	147
diagnose	132
admission	125
lesion	125
patient's	122
finding	117
case	115
positive	114
hospital	103
decrease	102

Figure 3: Word cloud from the general vocabulary.

Figure 4: Word cloud from medical terms (sample of 250 lines).

Table 4: Top-2 tokens most predictive of each MeSH-C class in the English corpus, as learned by the One-VS-Rest model. Coefficients indicate feature importance.

Code	MeSH Category	Top Tokens (Coefficient)
C01	Bacterial Infections and Mycoses	`covid (1.20), abscess (0.93)`
C02	Virus Diseases	Not applicable
C03	Parasitic Diseases	Not applicable
C04	Neoplasms	`metastatic (1.68), metastases (1.11)`
C05	Musculoskeletal Diseases	`consent (0.67), short (0.62)`
C06	Digestive System Diseases	`hepatitis (1.05), gastrointestinal (0.81)`
C07	Stomatognathic Diseases	`facial (1.26), having (1.24)`
C08	Respiratory Tract Diseases	`sars (1.85), covid (1.81)`
C09	Otorhinolaryngologic Diseases	`nasal (1.70), external (1.25)`
C10	Nervous System Diseases	`temporal (0.69), brain (0.66)`
C11	Eye Diseases	`acuity (1.37), replacement (1.36)`
C12	Urologic and Male Genital Diseases	`hiv (1.36), pelvic (0.84)`
C13	Female Genital Diseases and Pregnancy Complications	Not applicable
C14	Cardiovascular Diseases	`aneurysm (1.27), echocardiography (1.15)`
C15	Hemic and Lymphatic Diseases	`lymphoma (1.43), marrow (0.89)`
C16	Congenital, Hereditary, and Neonatal Diseases	`genetic (0.71), cov (0.68)`
C17	Skin and Connective Tissue Diseases	`breast (0.83), prednisolone (0.69)`
C18	Nutritional and Metabolic Diseases	`glucose (0.83), resolved (0.81)`
C19	Endocrine System Diseases	`adrenal (1.62), diabetes (1.24)`
C20	Immune System Diseases	`lymphoma (1.41), hiv (0.93)`
C21	Disorders of Environmental Origin	Not applicable
C22	Animal Diseases	`diarrhea (0.41), children (0.39)`
C23	Pathological Conditions, Signs and Symptoms	`magnetic (0.57), hemorrhage (0.34)`
C24	Occupational Diseases	`exposure (0.96), small (0.51)`
C25	Chemically Induced Disorders	`mental (1.45), maintained (1.13)`
C26	wounds and injuries	`injury (1.57), trauma (1.24)`

A Prompt used for synthetic data generation

Prompt utilisé pour la génération de cas cliniques en français à partir de codes MeSH

Je vais te fournir des codes MeSH en entrée. Tous les codes MeSH appartiennent à la catégorie C « Disease » et je ne te fournis que des codes des catégories de « niveau 1 », donc uniquement des codes compris entre C01 et C26.

Pour rappel, les catégories MeSH entre C01 et C26 sont les suivantes :

- C01 : Infections bactériennes et mycoses
- I listed all the MeSH categories from C01 to C26...
- C26 : Blessures et traumatismes

À partir des codes MeSH que je te fournis ci-dessous, écris en **français** un exemple de cas clinique réaliste tel qu'on pourrait le retrouver dans la littérature scientifique. Le texte doit comporter entre **300 et 600 mots**.

Attention : un cas clinique peut être associé à plusieurs catégories MeSH.

Avant de commencer, je vais te donner des exemples rédigés en anglais. Bien que tu doives produire les cas cliniques en français, ces exemples te servent de modèle de structure et de ton. (Les exemples suivent dans la section suivante.)

Maintenant, écris des cas cliniques similaires **en FRANÇAIS** à partir de la liste de codes MeSH que je vais te donner. Les cas doivent avoir une longueur comprise entre 300 et 600 mots