Multilingual Classification of Clinical Case Reports

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 derived from a subset of the English dataset. 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 (Gerardin et al., 2022).

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

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

one or more predefined labels (Jurafsky and Martin, 2025). This task has broad applications in the biomedical domain, such as biomedical literature indexing (Huang et al., 2011; Peng et al., 2016), automatic diagnosis code assignment (Perotte et al., 2013; Baumel et al., 2017), classification of public health-related tweets (Du et al., 2018a; Du et al., 2018b; Bian et al., 2017), and patient safety report categorization (Liang and Gong, 2017).

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) (Mork et al., 2013).

Automating this process using classical NLP tools, especially regex-based systems, faced three primary challenges: (1) the multiclass classification involved a high number of classes, often dozens to thousands; (2) the number of labels assigned varied significantly depending on the text content; and (3) 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 (Pita-Costa et al., 2021).

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 comprises 275 de-identified clinical case reports in French, among which 167 were manually annotated and used as the training set. The annotations include mentions of signs, symptoms, and disease-type entities, along with their contextual attributes, such as negations, hypothesis status, or references to individuals other than the patient. For some of these mentions, MeSH-C labels were directly associated with the entity, while at the document level, all MeSH-C labels mentioned at least once were aggregated to form the multilabel annotation. Notably, the annotation is confined to the top-level disease categories within Chapter C of the MeSH hierarchy, specifically at depth level 2. This restriction encompasses 24[1] categories, ranging from C01 (Infections) to C26 (Wounds and Injuries), thereby focusing the classification task on broad disease 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

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

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 with few manual annotations. 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 exactly the same set of 24 MeSH Chapter C "disease" categories (depth level 2) as defined in Gérardin et al. (2022) (see Appendix A for the presentation of the 24 MeSH categories). We apply this framework to a larger and more diverse dataset, the MultiCaRe corpus, which originally includes over 70,000 full-text clinical case reports in English from open-access PubMed articles (Offidani and Delrieux, 2023), from which we retained more than 10,000 after applying various filtering steps[2]. This dataset, in contrast to earlier work that used only abstracts, captures richer clinical narratives.

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: (1) we lack human-annotated entities; (2) our English dataset is significantly larger; (3) BioBERT is already trained to recognize biomedical language; and (4) 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, following the same label distribution as the English dataset, 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 biases introduced by translation and ensures our method generalizes to genuinely French-written texts. Furthermore, this LLM-based data augmentation technique is scalable and could be used to generate examples for underrepresented MeSH categories in future applications. However, as we lack real French clinical notes, we translated a subset of the English dataset to build the French test set. 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.

The datasets were split into training (80%), validation (10%), and test (10%) sets, resulting in 8,574, 1,072, and 1,072 clinical cases for the English dataset, and 167 synthetic cases for training, 19 for validation, and 300 translated cases for testing in the French dataset.

To assess how model architecture influences MeSH classification performance across both the large English dataset and the smaller French one, we compare three models:

- **Naive multilabel classifier (One-vs-Rest with TF-IDF)**: A TF-IDF bag-of-words model with a tokenizer extracting unigrams and bigrams. Each label is predicted independently using logistic regression. This model should serve as a baseline to contextualize the gains achieved by more advanced, context-aware NLP architectures.
- **BioBERT**: A domain-specific transformer model pretrained on large-scale biomedical literature, but restricted to the English language only. 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.

All models were fine-tuned using the validation set to optimize the micro F1 score. Additionally, 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.

[2]The filtering mainly involved removing clinical cases that lacked MeSH annotations for Chapter C "diseases", as well as discarding cases extracted from PubMed articles reporting multiple case studies where it was not possible to unambiguously link MeSH terms to individual cases.

- **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.

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 24 MeSH classes. For instance, 2,998 clinical cases were labeled under class C04 ("Neoplasms"), whereas classes C13 and C21 had no cases assigned, and class C22 ("Animal Diseases") included only 7 cases. Nevertheless, 20 out of the 24 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 (see Appendix A for the correlation matrix between MeSH labels). 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), and then "figure." Medical vocabulary appearing in the top 20 most frequent words included "blood" (13,655 occurrences), "examination" (12,687), "treatment" (11,966), and "cells" (9,616) (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 quantifiers often used in medical assessments, including "level," "increase," "negative," "positive," "decrease," and "figure." Also prevalent were hospitalization-related terms like "admission" and "hospital." 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.
One-VS-Rest (2-gram)	English	English	0.70	0.65	0.75
One-VS-Rest (2-gram)	French*	French**	0.30	0.31	0.30
BioBERT	English	English	0.67	0.75	0.60
BERT multilingual	English	English	0.66	0.74	0.59
BERT multilingual	English, French*	French**	0.57	0.68	0.49
BERT multilingual	English	French**	0.52	0.65	0.43
Gérardin et al. (2022) - BERT multilingual	**French, English**	**French**	**0.81**	**0.81**	**0.81**

* 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 multilingual (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 "C08 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 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 trained on human-annotated data, filtered out negated or hypothetical mentions, and linked these filtered mentions to MeSH-C categories via synonym lists, prior to training their BERT-based classification models (see Appendix A). 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. Despite its domain-specific pretraining on large biomedical corpora, BioBERT did not significantly outperform the general-purpose multilingual BERT model. This observation suggests that domain adaptation alone may be insufficient when the classification task is driven more by vocabulary presence than by deep contextual comprehension.

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. 148 French synthetic cases). Even more surprisingly, the multilingual BERT model achieved a micro F1 score of 0.52 on French clinical cases, despite being trained exclusively on English data, 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 multilingual 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 sconstraints 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 of depth 2. 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 how to use large language models such as Gemini to generate machine-annotated training data for MeSH entity recognition. This would reduce reliance on costly human annotations while enabling the use of state-of-the-art architectures using BERT-based NER systems for MeSH classification tasks.

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] Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2017). Multi-label classification of patient notes: a case study on ICD code assignment. *arXiv*. https://arxiv.org/abs/1709.09587

[5] Bian, J., Zhao, Y., Salloum, R. G., Guo, Y., Wang, M., Prosperi, M., et al. (2017). Using social media data to understand the impact of promotional information on laypeople's discussions: A case study of Lynch Syndrome. *Journal of Medical Internet Research*, 19(12), e414. https://doi.org/10.2196/jmir.9266

[6] Costa, J. P., Rei, L., Stopar, L., Fuart, F., Grobelnik, M., et al. (2021). NewsMeSH: A new classifier designed to annotate health news with MeSH headings. *Artificial Intelligence in Medicine*, 114, 102053. https://doi.org/10.1016/j.artmed.2021.102053

[7] Du, J., Tang, L., Xiang, Y., Zhi, D., Xu, J., Song, H., & Tao, C. (2018). Public perception analysis of tweets during the 2015 measles outbreak: Comparative study using convolutional neural network models. *Journal of Medical Internet Research*, 20(7), e236. https://doi.org/10.2196/jmir.9413

[8] Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making*, 18(S2). https://doi.org/10.1186/s12911-018-0632-8

[9] Huang, M., Névéol, A., & Lu, Z. (2011). Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5), 660–667. https://doi.org/10.1136/amiajnl-2010-000055

[10] Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3e éd.). *Prentice Hall eBooks*. https://nats-www.informatik.uni-hamburg.de/pub/CDG/JurafskyMartin00Comments/JurafskyMartin00-Review.pdf

[11] Liang, C., & Gong, Y. (2017). Automated classification of multi-labeled patient safety reports: A shift from quantity to quality measure. *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-61499-830-3-1070

[12] Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013). The NLM Medical Text Indexer system for indexing biomedical literature. *BioASQ 2013*.

[13] 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

[14] Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2013). Diagnosis code assignment: Models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2), 231–237. https://doi.org/10.1136/amiajnl-2013-002159

[15] Sciannameo, V., Pagliari, D. J., Urru, S., Grimaldi, P., et al. (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

[16] Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., et al. (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

[17] 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

[18] 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

A Appendix / supplemental material

Table 2: MeSH Disease Categories (C01–C26)

Code	MeSH Category Name
C01	Infections*
C04	Neoplasms
C05	Musculoskeletal Diseases
C06	Digestive System Diseases
C07	Stomatognathic Diseases
C08	Respiratory Tract Diseases
C09	Otorhinolaryngologic Diseases
C10	Nervous System Diseases
C11	Eye Diseases
C12	Urologic and Male Genital Diseases
C13	Female Genital Diseases and Pregnancy Complications
C14	Cardiovascular Diseases
C15	Hemic and Lymphatic Diseases
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
C17	Skin and Connective Tissue Diseases
C18	Nutritional and Metabolic Diseases
C19	Endocrine System Diseases
C20	Immune System Diseases
C21	Disorders of Environmental Origin
C22	Animal Diseases
C23	Pathological Conditions, Signs and Symptoms
C24	Occupational Diseases
C25	Chemically Induced Disorders
C26	Wounds and Injuries

**Note: As of the 2020 MeSH update, categories C02 (Virus Diseases) and C03 (Parasitic Diseases) have been merged into C01 (Infections).*

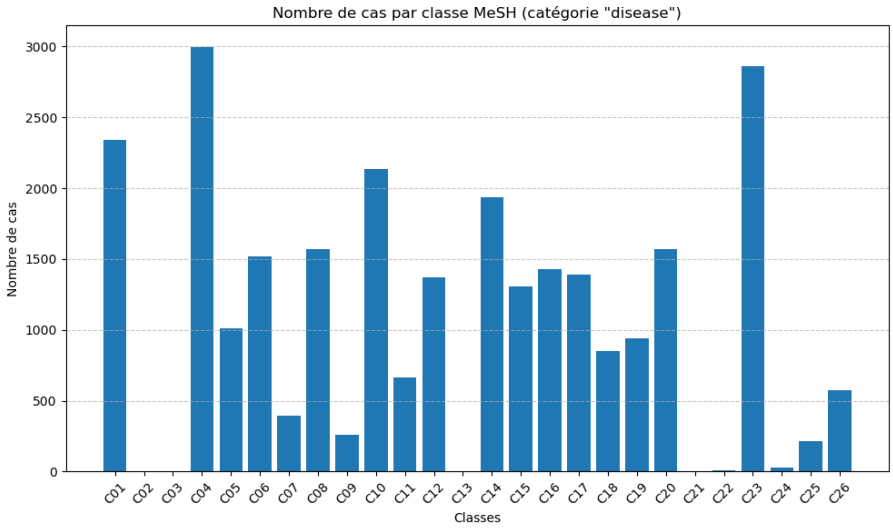
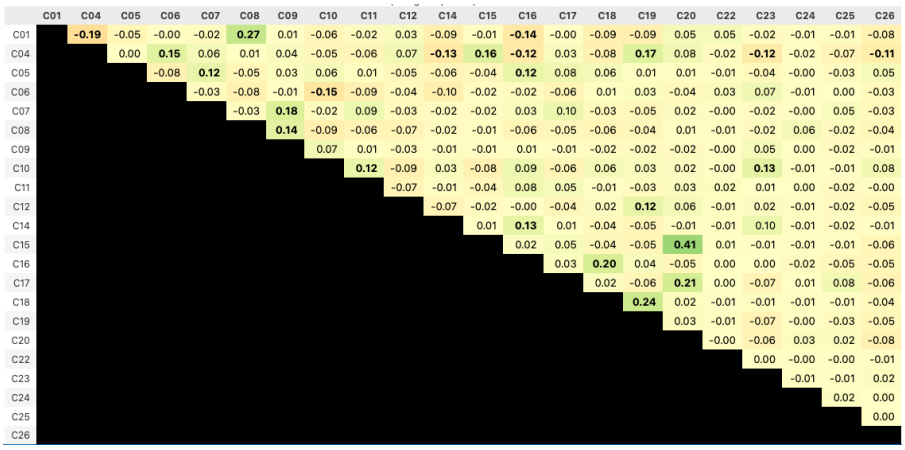


Figure 1: Number of cases per MeSH category (subcategory "disease").



	C01	C04	C05	C06	C07	C08	C09	C10	C11	C12	C14	C15	C16	C17	C18	C19	C20	C22	C23	C24	C25	C26
C01		-0.19	-0.05	-0.00	-0.02	0.27	0.01	-0.06	-0.02	0.03	-0.09	-0.01	-0.14	-0.00	-0.09	-0.09	0.05	0.05	-0.02	-0.01	-0.01	-0.08
C04			0.00	0.15	0.06	0.01	0.04	-0.05	-0.06	0.07	-0.13	0.16	-0.12	0.03	-0.08	0.17	0.08	-0.02	-0.12	-0.02	-0.07	-0.11
C05				-0.08	0.12	-0.05	0.03	0.06	0.01	-0.05	-0.06	-0.04	0.12	0.08	0.06	0.01	0.01	-0.01	-0.04	-0.00	-0.03	0.05
C06					-0.03	-0.08	-0.01	-0.15	-0.09	-0.04	-0.10	-0.02	-0.02	-0.06	0.01	0.03	-0.04	0.03	0.07	-0.01	0.00	-0.03
C07						-0.03	0.18	-0.02	0.09	-0.03	-0.02	-0.02	0.03	0.10	-0.03	-0.05	0.02	-0.00	-0.02	-0.00	0.05	-0.03
C08							0.14	-0.09	-0.06	-0.07	-0.02	-0.01	-0.06	-0.05	-0.06	-0.04	0.01	-0.01	-0.02	0.06	-0.02	-0.04
C09								0.07	0.01	-0.03	-0.01	-0.01	0.01	-0.01	-0.02	-0.02	-0.02	-0.00	0.05	0.00	-0.02	-0.01
C10									0.12	-0.09	0.03	-0.08	0.09	-0.06	0.06	0.03	0.02	-0.00	0.13	-0.01	-0.01	0.08
C11										-0.07	-0.01	-0.04	0.08	0.05	-0.01	-0.03	0.03	0.02	0.01	0.00	-0.02	-0.00
C12											-0.07	-0.02	-0.00	-0.04	0.02	0.12	0.06	-0.01	0.02	-0.01	-0.02	-0.05
C14												0.01	0.13	0.01	-0.04	-0.05	-0.01	-0.01	0.10	-0.01	-0.02	-0.01
C15													0.02	0.05	-0.04	-0.05	0.41	0.01	-0.01	-0.01	-0.01	-0.06
C16														0.03	0.20	0.04	-0.05	0.00	0.00	-0.02	-0.05	-0.05
C17															0.02	-0.06	0.21	0.00	-0.07	0.01	0.08	-0.06
C18																0.24	0.02	-0.01	-0.01	-0.01	-0.01	-0.04
C19																	0.03	-0.01	-0.07	-0.00	-0.03	-0.05
C20																		-0.00	-0.06	0.03	0.02	-0.08
C22																			0.00	-0.00	-0.00	-0.01
C23																				-0.01	-0.01	0.02
C24																					0.02	0.00
C25																						0.00
C26																						

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

Table 3: 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 4: 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).

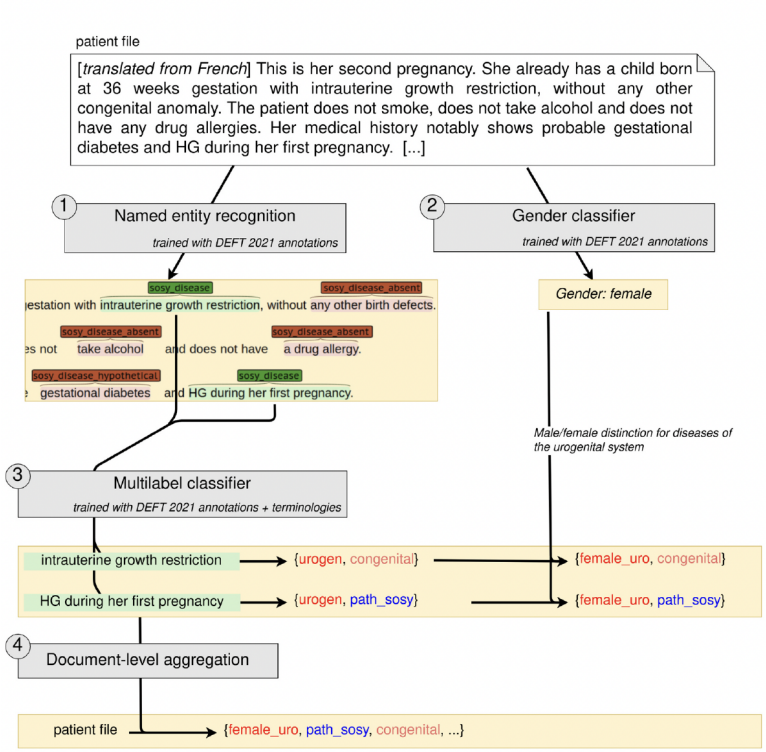


Fig. 3. General system architecture. First, named entity recognition is performed, trained on the annotated DEFT dataset to extract positive medical concepts (1). In parallel, the gender classifier, also trained on the DEFT dataset, determines the written gender of the patient (2). Then, a multilabel 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).

Figure 5: MeSH classifier architecture from Gérardin et al., 2022.

Table 5: 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	Infections	`covid (1.20), abscess (0.93)`
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)`