# Développement d'un outil d'aide à la décision thérapeutique pour quatre maladies rares : le lupus érythémateux, la sclérodermie systémique, la maladie de Takayasu, le syndrome des anti-phospholipides

Christel Gérardin

ÉCOLE DOCTORALE PIERRE LOUIS DE SANTÉ PUBLIQUE ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE (ED n° 393)

Titre : **Aide à la décision thérapeutique par extraction automatique de cohortes de patients similaires à partir des comptes-rendus médicaux en contexte de maladies auto-immunes : une analyse sur l'entrepôt de données de Santé de l'AP-HP**

Thèse d'informatique médicale

par Christel Gérardin

Soutenue publiquement le 29 septembre 2023

**Jury:**

- Pr Marc Cuggia (rapporteur)
- Pr Pierre Zweigenbaum (rapporteur)
- Pr Sophie Georgin-Lavialle (membre du jury)
- Pr Anne-Sophie Jannot (membre du jury)
- Dr Marie Franck (membre du jury)
- Pr Fabrice Carrat (directeur de thèse)
- Pr Arsène Mékinian (co-directeur de thèse)

# Remerciements

Je remercie en tout premier lieu mon directeur de thèse, professeur Fabrice Carrat, avec qui j'ai eu l'honneur de travailler pendant ces trois années de thèse. Je lui suis particulièrement reconnaissante pour son exigence et sa bienveillance.

Je remercie également le professeur Arsène Mekinian pour avoir co-dirigé cette thèse avec un souci constant de placer les questions médicales au cœur du sujet.

Je suis particulièrement reconnaissante auprès des professeurs Marc Cuggia et Pierre Zweigenbaum qui nous ont fait l'honneur, par leur grande expertise sur le sujet, de la rapporter.

Je souhaite également adresser mes plus sincères remerciements au professeur Sophie Georgin-Lavialle, le professeur Anne-Sophie Jannot et le Docteur Marie Frank, pour avoir accepté d'être membre du jury.

Je tiens aussi à exprimer ma profonde reconnaissance au professeur Xavier Tannier, qui, par ses conseils et son soutien constant, a permis la réalisation d'une grande partie du travail présenté ici.

Un grand merci aussi à mes collègues du laboratoire et co-doctorants pour avoir été mes compagnons de travail pendant ces trois années : Nathanael, Isabelle, Laurent, Bertrand, Elhadji, Frederic, Gregory, Paul, Emmanuelle, Maria, Benjamin.

Je souhaite également remercier spécialement mes collègues de l'Entrepôt de Données de Santé : Noël (y compris pour sa relecture précieuse du manuscrit), Alice, Perceval, Thomas, Adam, David, Edouard, Damien, Nicolas, Christophe, Christel.

Je remercie également mes amies Océane, Justine, Eline, Nihal, Gabriel et Charlotte, Danaé et Sébastien pour leur soutien infaillible cette année.

Je remercie aussi ma belle-famille si attentionnée, mes parents et mes trois sœurs Claire, Marine et Lucile pour leur écoute et leurs encouragements quotidiens.

Et enfin et surtout Maxime et nos deux petits soleils.

## Publications :

- Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artificial Intelligence in Medicine. 2022 Jun 1;128:102311.

- Gérardin C, Mageau A, Mékinian A, Tannier X, Carrat F. Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study. JMIR Medical Informatics. 2022 Dec 19;10(12):e42379.

- Gérardin C, Xong Y, Mekinian A, et al. AB1767-HPR DOCUMENT SEARCH IN LARGE RHEUMATOLOGY DATABASES: ADVANCED KEYWORD QUERIES TO SELECT HOMOGENEOUS PHENOTYPES. Annals of the Rheumatic Diseases 2023;82:2117-2118.

**Résumé** :

Contexte : Les dossiers médicaux électroniques et les entrepôts de données de santé donnent accès à un grand volume d'informations cliniques riches mais souvent non structurées (c'est-à-dire en texte brut). Les maladies auto-immunes sont rares dans la population générale et il est donc difficile de mener des essais thérapeutiques en raison de la taille trop réduite des échantillons. Pour fonder la décision médicale, la présentation de l'analyse de cas similaires peut être une aide précieuse, permettant un raisonnement par analogie.

Objectif : L'objectif de la thèse est l'extraction automatique de cohortes de patients similaires, à partir d'informations directement présentes dans les notes cliniques (symptômes, comorbidités, résultats de biologie et d'imagerie, etc..), avec un enjeu central d'interprétabilité des résultats pour les cliniciens.

Méthode : La question initiale pour la construction des cohortes de patients similaires est la validation d'une modélisation du patient permettant le calcul d'une similarité cliniquement interprétable, à partir des informations des comptes-rendus.

Un prototype d'extraction automatique de cas similaires à partir de l'entrepôt de données de santé de l'AP-HP a été développé pour 4 maladies auto-immunes : le lupus érythémateux disséminé, la sclérodermie, le syndrome des antiphospholipides et la maladie de Takayasu.

Les cas similaires présentent des maladies et des symptômes semblables à ceux du patient traité, et sont sélectionnés à partir des informations présentes dans les rapports médicaux. Le prototype comprend plusieurs étapes :

1/ Présélection des patients d'intérêt à l'aide d'une méthode d'expansion de requête basée sur un algorithme de type Transformer (sur 20 phénotypes, précision moyenne 0,93 [0,90 ; 0,96] et rappel moyen 0,78 [0,71 ; 0,85]), ayant fait l'objet d'une soumission (article 2).

2/ Extraction des concepts médicaux d'intérêt (symptômes, maladies, valeurs biologiques et traitements) par un algorithme de "deep learning" avec un score F1 moyen de 0,81 [0,79 ; 0,82], avec une méthode originale d'annotation en signes et symptômes pathologiques versus physiologiques pour fournir un modèle du patient proche de la clinique.

3/ Classification des concepts extraits selon les principaux domaines médicaux (neurologie, hématologie, cardiologie, etc.), correspondant aux en-têtes des sous-chapitres de la branche C-maladie du MeSH [5], permettant de calculer une similarité avec le patient index selon les domaines d'intérêt. Ce classifieur multi label est basé sur un transformer BERT entraîné sur l'ensemble des synonymes des termes MeSH de l'UMLS et un jeu de concepts annotés. Les performances de ce classifieur ont été testées sur un jeu de données cliniques externes avec une mesure F1 de 0.809 à 0.811 en fonction des modèles testés et a fait l'objet d'une publication (article 4). Les résultats biologiques sont également extraits et comparés par leur z-score permettant un

calcul de similarité multimodale. Un prétraitement des textes cliniques avec analyse des différentes sections pertinentes du texte (“Histoire de la maladie”, “Traitement à l’entrée” etc..). La méthode de construction d'une cohorte de patients similaires a été validée pour des phénotypes complexes tels que la pneumopathie interstitielle dans la sclérodermie (précision allant de 0,65 [0,58 ; 0,72] à 0,98 [0,97 ; 0,99]) et publiée (article 5).

4/ Pour ces cas similaires, les traitements et leur posologie sont extraits et standardisés selon leur classe thérapeutique (ATC).

Ce projet a par ailleurs été soumis pour un appel à projet INSERM (MESSIDORE) en collaboration avec cinq équipes - dont deux équipes cliniques de médecine interne - pour la finalisation du développement du prototype. Il s’agira, au cours des réunions multidisciplinaires thérapeutiques, d’ajouter l’analyse rétrospective de cohorte de patients similaires à l’expertise des différents spécialistes pour aider à la décision thérapeutique. Une première étude pilote consistera à évaluer l’impact d’un tel outil sur la décision thérapeutique (changement de molécule, posologie, durée …).

**Abstract**:

Background: Electronic medical records and health data warehouses provide access to a large volume of rich but often unstructured (i.e. plain text) clinical information. Autoimmune diseases are rare in the general population, making it difficult to conduct therapeutic trials due to small sample sizes. In order to provide a basis for medical decisions, the analysis of similar cases can be a valuable aid, enabling reasoning by analogy.

Objective: The aim of this thesis is therefore the automatic extraction of cohorts of similar patients, from information directly present in clinical notes (symptoms, comorbidities, biological and imaging results, etc.), with a central issue of interpretability of results for clinicians.

Method: The main initial question for the construction of cohorts of similar patients is the validation of a model of the patient enabling the calculation of a clinically interpretable similarity, based on the information in the reports.

A prototype for automatic extraction of similar cases from the AP-HP healthcare data warehouse has been developed for 4 autoimmune diseases: systemic lupus erythematosus, scleroderma, antiphospholipid syndrome and Takayasu disease.

Similar cases present diseases and symptoms similar to those of the patient being treated, and are selected on the basis of information contained in medical reports. The prototype comprises several stages:

1/ Pre-selection of patients of interest using a query expansion method based on a transform-type algorithm (on 20 phenotypes, mean PPV 0.93 [0.90; 0.96] and mean sensitivity 0.78 [0.71; 0.85]), submitted (article 2).

2/ Extraction of medical concepts of interest (symptoms, diseases, biological values and treatments) by a deep learning algorithm with an average F1 score of 0.81 [0.79; 0.82]. With an original method of annotation into pathological versus physiological signs and symptoms to provide a clinically close model of the patient.

3/ Classification of extracted concepts into broad domains (neurology, hematology, cardiology, etc.), corresponding to the headings of the sub-chapters of the C-disease branch of MeSH, making it possible to calculate a similarity with the index patient according to the domains of interest. This multi-label classifier is based on a BERT transformer trained on the set of synonyms of MeSH terms from UMLS and a set of annotated concepts. The performance of this classifier was tested on an external clinical dataset with an F1 measure of 0.809 to 0.811 depending on the models tested, and was the subject of a publication (article 4).

Biological results are also extracted and compared by their z-score, enabling a multimodal similarity calculation.

A pre-processing of clinical texts with analysis of the various relevant sections of the text ("History of illness", "Treatment at entry" etc.) is currently being submitted.

The method of constructing a cohort of similar patients has been validated for complex phenotypes such as interstitial lung disease in scleroderma (PPV ranging from 0.65 [0.58; 0.72] to 0.98 [0.97; 0.99]) and published (article 5).

4/ For these similar cases, the treatments and their dosage are extracted and standardized according to their therapeutic class (ATC).

This project has also been submitted to an INSERM call for projects (MESSIDORE) in collaboration with five teams - including two internal medicine clinical teams - to finalize the development of the prototype. During multidisciplinary therapeutic meetings, the retrospective analysis of a cohort of similar patients will be added to the expertise of the various specialists to help with therapeutic decisions. An initial pilot study will evaluate the impact of such a tool on therapeutic decisions (change of molecule, dosage, duration, etc.).

**Mots-clés :** Traitement automatique des langues, Phénotypage, Patients similaires

**Plan :**

# 1. Introduction

## 1.1. Contexte

En pratique clinique, la médecine fondée sur les preuves, c'est-à-dire basée sur des essais contrôlés randomisés, ne répond pas à toutes les questions thérapeutiques. C'est particulièrement le cas lorsque la maladie est rare ou lorsque les patients sont peu inclus dans les essais cliniques : enfants, patients âgés ou patients présentant des comorbidités (par exemple, l'insuffisance rénale). Dans ce contexte, les avis d'experts basés sur des études de moindre niveau de preuve restent l'alternative la plus utilisée, peinant parfois à convaincre le clinicien.

Pour ces patients, il est en outre souvent difficile de déterminer les risques immédiats et à moyen ou long terme associés à leur état et de proposer le meilleur traitement. La recherche d'articles publiés sur ces situations cliniques rares est également souvent infructueuse.

Les entrepôts de données peuvent alors fournir des éléments cliniques et pronostiques aidant à la prise en charge, en identifiant des patients ayant le même profil et en présentant leur évolution clinique dans différents contextes de prise en charge.

Une expérience particulièrement intéressante est fournie par une équipe médicale de Stanford en charge d'un patient atteint de lupus pédiatrique avec syndrome néphrotique et péricardite, qui a interrogé sa base de données "STRIDE" pour évaluer son risque de thrombose [1]. L'extraction manuelle et en temps réel des données a permis d'identifier une dizaine de patients pertinents et d'évaluer en temps réel le risque pour le patient, avec comme impact une décision thérapeutique préventive d'anticoagulation à dose curative.

Depuis, plusieurs équipes ont proposé des algorithmes d'extraction de cas similaires pour aider à la décision thérapeutique, en oncologie par exemple [2], ou pour d'autres pathologies fréquentes comme le diabète [3].

Nous poursuivons ici cette voie de l'automatisation de l'identification des patients similaires, en nous appuyant sur la base de données de santé de l'Assistance publique - Hôpitaux de Paris (AP-HP).

## 1.2. Questions et objectifs de la recherche

L'objectif du projet est d'élaborer un algorithme d'extraction automatique de patients similaires à partir des comptes-rendus médicaux, en contexte de maladies auto-immunes. Le prototype fournit

une synthèse statistique d'études de cas pour permettre un raisonnement par analogie, centrée sur un cas, c'est-à-dire un patient présentant un problème spécifique.

Le projet tient notamment compte des objectifs/principes suivants :

- prise en compte de la multiplicité des facteurs de similitude : l'algorithme de rapprochement des patients tient compte de leurs comorbidités, symptômes, résultats d'examens, âge, sexe, etc.

- interprétabilité et modularité du modèle : les étapes de développement de l'algorithme s'appuient systématiquement sur le raisonnement clinique et les symptômes et les maladies sont également classés par domaine médical - ou spécialité- (infectieuse, cardiovasculaire, hématologique, etc.) afin que les cliniciens puissent choisir sur quel domaine comparer les patients ;

- lisibilité pour les cliniciens, en portant une attention à la visualisation des résultats lorsque, une fois construite, la cohorte de cas similaires est analysée, notamment pour chaque classe thérapeutique présente dans les rapports. Une analyse du risque relatif pour les principales complications (infections, défaillances d'organes, thromboses, etc.) est en effet réalisée, ainsi que pour le pronostic du patient (réhospitalisation, décès, retour à domicile, etc.).

Pour les besoins de l'exercice, le patient à traiter dans les différentes expérimentations présentées est tiré au sort dans la base de données. Pour une première preuve de concept, l'étude est restreinte à quatre pathologies : Lupus, maladie de Takayasu, sclérodermie et syndrome des antiphospholipides. Par la suite, le prototype fera l'objet d'une étude pilote pour l'évaluer en pratique clinique, à partir des observations de patients à traiter directement.

Il vise à démontrer l'utilité des algorithmes de traitement automatique des langues, et en particulier des modèles de langage récemment développés et très performants (notamment BERT [4] ou GPT-3 [5] à l'origine de Chat GPT [6]). Il cherche également à préserver l'interprétabilité du modèle en s'appuyant systématiquement sur le raisonnement clinique pour les étapes de développement.

# 2. Etat de l'art

## 2.1. Traitement du langage naturel

### 2.1.1. Généralités

Une partie importante de l'information clinique est présente directement dans les textes bruts des dossiers patients informatisés, notamment les symptômes, l'ensemble des pathologies, les antécédents familiaux etc… Ces textes bruts sont également appelés *données non structurées* dans la littérature scientifique, par opposition aux données structurées, issues des tables, correspondant par exemple aux données du codage PMSI (Programme de médicalisation des systèmes d'information) de valorisation des soins tels que les actes CCAM (Classification commune des actes médicaux), les codes CIM10 (Classification internationale des maladies) ou aux données de biologie.

L'apport de l'analyse des données non structurées a notamment été montré pour la prédiction clinique : une revue de littérature récente [7] portant sur 126 études conclut à une amélioration des résultats par l'utilisation couplée des données structurées et non structurées. L'utilisation des textes bruts a également permis l'obtention de bonnes performances pour d'autres applications cliniques telles que l'identification de cohortes [8] ou l'aide à la décision diagnostique et thérapeutique [9, 10].

L'analyse automatisée des textes nécessite l'utilisation d'outils du Traitement automatique des langues (TAL), également appelé traitement automatique de la langue naturelle ou NLP (pour Natural Language Processing en anglais). Le TAL est un domaine de l'informatique, de l'intelligence artificielle et de la linguistique qui s'intéresse aux interactions entre les ordinateurs et les textes en langage humain (naturel) [11]. Il vise à doter les programmes informatiques de la capacité de traiter et de comprendre des données textuelles non structurées.

Deux éléments essentiels au TAL appliqué à la clinique sont, d'une part, les ressources linguistiques telles que les terminologies et les ontologies (bases de connaissances médicales) et, d'autre part, les outils algorithmiques. Pour permettre une meilleure compréhension pour la suite, ces deux éléments seront définis dans cette première partie avant de développer plus précisément l'état de l'art des différentes étapes de construction de la cohorte de patients similaires.

### 2.1.2. Les bases de connaissances médicales

Dans le domaine biomédical, il existe différentes ressources linguistiques dont les ontologies, les vocabulaires organisés et les classifications.

Les ontologies sont des bases de connaissances organisées de façon formelle et structurée pour décrire un domaine scientifique spécifique. Elles permettent de décrire des concepts et leurs propriétés d'une part et les relations entre ces concepts d'autre part. On peut citer en exemple la SNOMED_CT (Sytematized Nomenclature of Medicine –Clinical Terms) [12], qui est une collection de termes médicaux (médicaments, pathologies, symptômes, tests biologiques etc…) systématiquement organisée. Elle contient quatre composantes principales : Les *codes numériques* des concepts organisés hiérarchiquement, les *descriptions textuelles* des concepts, les *relations* entre les codes conceptuels et les *ensembles de références* utilisés pour regrouper les concepts et permettre une mise en correspondance avec les autres classifications.

La distinction entre les ontologies et les autres ressources linguistiques tient essentiellement dans le degré d'organisation des connaissances. Pour faire un parallèle, une ontologie peut être comparée à la classification périodique des éléments de Mendeleïev, au sens où celle-ci a permis d'identifier les éléments qui restaient à découvrir. Pour reformuler, l'architecture d'une ontologie en elle-même contient la connaissance. A contrario, les listes organisées de termes correspondent davantage à des dictionnaires.

En dehors des ontologies, il existe des listes organisées de termes (ou terminologies) tel que le MeSH [14] par exemple qui correspond à un lexique organisé en arborescence de concepts biomédicaux, et dont les grandes catégories sont par exemple : l'anatomie, les maladies, les médicaments, etc. Il a été créé par la Bibliothèque Nationale de Médecine des Etats-Unis (National Library of Medicine, US NLM), et permet notamment l'indexation des publications médicales par mots-clés. Les concepts médicaux (environ 30 000 en anglais en 2022) sont répertoriés par catégorie, et associés à une description textuelle et des synonymes.

Un autre vocabulaire organisé pour les médicaments est celui de l'ATC [15] (Anatomical Therapeutic Chemical Classification System), système de classification contrôlé par l'Organisation Mondiale de la Santé. Les médicaments y sont classés par groupes selon l'organe sur lequel ils agissent.

Enfin, la NLM a également développé une ressource centrale regroupant un grand nombre de vocabulaires et d'ontologies médicales : l'UMLS (Unified Medical Language System) [16] . En 2023, elle regroupe plus d'une centaine de terminologies, avec une représentation de plus de vingt-cinq langues différentes. Cette ressource linguistique permet de faciliter l'intégration et l'interopérabilité (c'est-à-dire l'échange standardisé) des informations médicales. En effet, au sein de l'UMLS [16] chaque concept- issu de différents vocabulaires et associé à ses synonymes est relié à un identifiant unique : le CUI (Concept Unique Identifier), permettant une standardisation du concept. A titre d'exemple, le concept "paracétamol" (CUI C0000970) correspond au concept "acetaminophen" dans

l'UMLS avec, comme synonymes "N-Acetyl-p-aminophenol", "Hydroxyacetanilide", "paracetamol", etc…

### 2.1.3. Les différentes tâches de TAL dans les applications cliniques

Dans le domaine médical et de la santé, les techniques de TAL sont utilisées pour traiter le volume important d'information : par exemple, pour résumer les notes cliniques, pour l'analyse des traitements, pour l'extraction et la récupération d'informations à partir des comptes-rendus de sortie, et pour la compréhension sémantique des requêtes des patients [11] .

L'étendue des tâches cliniques du TAL ne cesse de croître à mesure que le domaine du TAL évolue avec les progrès des systèmes linguistiques, ces tâches sont majoritairement les suivantes :

- la *reconnaissance d'entités nommées* qui correspond à l'identification et à l'extraction des éléments textuels faisant référence à des concepts médicaux spécifiques comme les médicaments, les maladies, les procédures diagnostiques et thérapeutiques. Les types de concepts retenus sont fréquemment ceux proposés par l'UMLS [17] . Cette tâche est celle qui sera la plus utilisée dans ce travail. La reconnaissance d'entités nommées est particulièrement complexe en contexte clinique notamment du fait de l'importance de la présence des acronymes et des abréviations, mais également en raison d'une importante part d'implicite dans les textes. La reconnaissance d'entités nommées correspond habituellement à deux tâches : celle de l'extraction de la mention elle-même ("diabète", "paracétamol", "pace-maker", etc…) et celle de l'*attribut* de la mention lorsqu'elle est présente. Un exemple fréquent d'attribut est, par exemple, le caractère nié d'un terme ("le patient n'a pas de diabète", "paracétamol suspendu", "pas d'indication à un pace-maker"). Un autre attribut peut également être le caractère hypothétique du terme ("le patient aurait un diabète") ou bien le fait que la mention du concept est rattaché à une autre personne que le patient ("son frère a un diabète", "sa soeur a fait une intoxication au paracétamol" etc…). Ces attributs peuvent être de complexités variées et les approches pour les extraire automatiquement sont parfois différentes de celles utilisées pour extraire les entités.

- la *normalisation des entités,* également appelée "entity linking" en anglais, qui consiste à associer les concepts médicaux à leur identifiant unique dans une base de connaissances médicales, telle que l'UMLS[16] par exemple ou le MeSH [14] . Il s'agit de pouvoir relier

l'ensemble des synonymes, acronymes et abréviations dans le texte à une référence commune, permettant une harmonisation des informations provenant des différents comptes rendus médicaux pour faciliter l'analyse et la comparaison des données médicales, et la communication et l'échange entre les systèmes et acteurs en santé (base de données, registre, épidémiologiste, institution nécessitant des relevés précis d'indicateurs, etc…).

- La *classification des documents* médicaux, qui correspond à relier les documents à certaines catégories prédéfinies.

- La *pseudonymisation/dé-identification*: en contexte médical, l'anonymisation est un enjeu majeur pour la sécurité des données des patients et reste un défi important. La dé-identification totale reste difficile du fait notamment de la singularité des histoires cliniques et des contextes décrits dans les comptes-rendus. Néanmoins, l'ensemble des données complètement identifiantes tels que les noms, lieux etc.. peuvent être filtrées avec de très bonnes performances [18] .

D'autres tâches du traitement automatisé du langage sont également actuellement couramment appliquées aux données de santé telles que la traduction, la synthèse des textes cliniques en résumés et la tâche de "Question-Réponse" (Question Answering).

### 2.1.4. Les modèles de langages

Une part importante du travail présenté dans ce manuscrit est basée sur des algorithmes de modèles de langages. L'objectif de ce paragraphe est d'en fournir une vision schématique pour permettre une compréhension de leur fonctionnement.

#### 2.1.4.1. Modèle d'apprentissage profond et réseaux de neurones

Un *modèle d'apprentissage profond* est un type de modèle conçu pour apprendre à partir de données. Il peut être entraîné pour différentes tâches : la prédiction, la classification, la reconnaissance d'images, la traduction etc…

Un *réseau de neurones profonds* est un type de modèle d'apprentissage automatique dont l'origine schématique s'inspire du fonctionnement des neurones biologiques (Wikipédia 2023). Il s'agit d'un système dont les unités fonctionnelles - les neurones - fonctionnent en réseau, c'est-à-dire échangent

de l'information. Ces neurones artificiels s'inspirent du fonctionnement biologique en recevant un signal d'entrée et en émettant un signal de sortie en fonction de la quantité d'information accumulée.

La traduction mathématique de cette unité fonctionnelle est représentée par le schéma figure 1 ci-après:

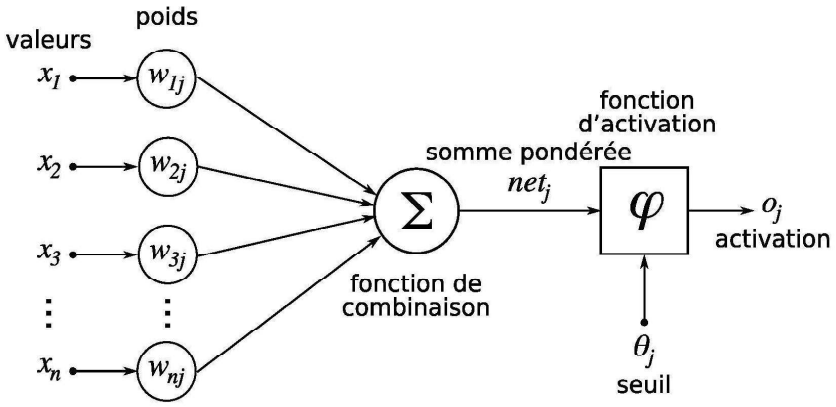



**Figure 1 : Schéma d'un neurone artificiel.** La cellule effectrice (à droite) reçoit un signal d'entrée (à gauche), constitué de plusieurs variables ou valeurs ; ce signal est ensuite pondéré par une suite de poids qui sont combinés (combinaison linéaire) ; cette somme pondérée (ou signal accumulé) est ensuite transmise à la fonction d'activation qui "passe" ou non le signal en fonction d'un certain seuil(source : Wikipédia).

Le neurone artificiel est donc constitué de deux parties : une partie de combinaison linéaire des différentes valeurs du signal d'entrée et une partie de fonction d'activation qui ne passe l'information qu'au-dessus d'un certain seuil (partie non linéaire).

L'architecture finale est ensuite constituée de plusieurs couches de neurones avec, pour chacun d'eux, un nombre de paramètres associés et en particulier les *poids* pondérant les valeurs d'entrée du signal. Ces *poids* sont les paramètres qui vont être *appris* itérativement par l'algorithme en fonction de la tâche sur laquelle il aura été entraîné.

A titre d'exemple, supposons que l'on veuille entraîner un modèle prédisant la probabilité de précipitation dans l'heure. Le modèle dispose pour son entraînement d'un échantillon de signaux d'entrée constitués par exemple de plusieurs variables : humidité ambiante, température, géolocalisation, etc… et, associés à ces entrées, les signaux de sortie pluie ou absence de pluie. L'apprentissage consistera alors à *optimiser* itérativement les *poids* pour que le modèle sache prédire la pluie ou l'absence de pluie en fonction des conditions météorologiques d'entrée.

Les étapes de cet apprentissage sont les suivantes :

1. Initialisation des poids : initialement, les poids associés aux neurones sont en général définis aléatoirement. Ils sont ensuite modifiés au fur et à mesure de l'apprentissage.
2. Passe *avant* : les signaux d'entrées sont fournis au modèle, dans toutes les couches du réseau de neurones, fournissant les premiers signaux de sortie du modèle.
3. Calcul de l'erreur : une fonction d'erreur est utilisée pour calculer l'erreur entre la sortie du modèle et la sortie attendue. Cette fonction d'erreur permet d'évaluer les performances du modèle sur les données d'entraînement.
4. Rétropropagation de l'erreur - aussi appelée rétropropagation du gradient : la rétropropagation du gradient est utilisée pour calculer le gradient de l'erreur par rapport à chaque poids du modèle. Le gradient indique la direction et le taux de changement nécessaires pour réduire l'erreur.
5. Mise à jour des poids : les poids du modèle sont mis à jour en utilisant un algorithme d'optimisation, tel que la descente de gradient. L'algorithme utilise le gradient calculé précédemment pour ajuster les poids dans la direction qui réduit l'erreur.
6. Répétition des étapes 2 à 5 : les étapes de passe avant, de calcul de l'erreur, de rétropropagation du gradient et de mise à jour des poids sont répétées sur plusieurs itérations, ou époques, pour améliorer progressivement les performances du modèle.
7. Validation et évaluation : une fois l'entraînement terminé, le modèle est évalué sur des données de validation et de test pour estimer ses performances sur des exemples qu'il n'a pas vus pendant l'entraînement. Cela permet de vérifier la généralisation du modèle.

Le schéma de ces étapes peut être résumé ci-après, les couches intermédiaires de neurones entre l'entrée et la sortie sont appelées "couches cachées" :

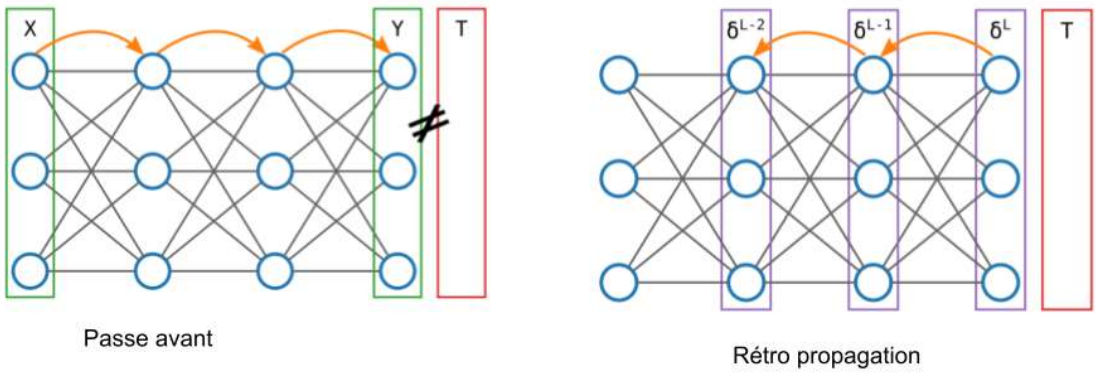



**Figure 2 : Réseau de neurone : illustration de la rétro-propagation de l'erreur**. Celle-ci permet l'adaptation itérative des poids de chaque neurone pour que le modèle apprenne à prédire

correctement. La passe avant (étape 2 détaillée ci-dessus), à gauche, et la rétropropagation et modification des poids à droite et (étape 4 et 5) à droite. Chaque cercle correspond à un neurone, l'entrée est le vecteur X (signal d'entrée), la sortie calculée du modèle correspond au Y et la vérité que l'on souhaite prédire est représentée par le vecteur T. (Source : https://www.miximum.fr/blog/introduction-au-deep-learning-2/)

#### 2.1.4.2. Mesures de performances des modèles

***Généralités sur l'entraînement des modèles***

En apprentissage supervisé, le jeu de données est donc découpé en jeu d'entraînement, jeu de validation et jeu de test. Dans le contexte du réseau de neurones, le jeu *d'entraînement* permet de modifier itérativement les poids du modèle, le jeu de *validation* est celui à partir duquel on modifie les hyperparamètres (vitesse de convergence de la descente du gradient, nombre de fois où le jeu de données est "vu" par le réseau de neurone, architecture du réseau, graine d'initialisation des poids aléatoires, etc..) et le jeu de *test* est celui utilisé pour calculer les performances du modèle. La répartition habituelle est de conserver 80% du jeu de données pour l'entraînement, 10% pour la validation, 10% pour le test.

Un des inconvénients des méthodes d'apprentissage supervisé est le risque de *sur-apprentissage* ("overfitting" en anglais) : lorsque le modèle apprend de façon trop exacte sur le jeu d'entraînement, il n'est pas capable de généraliser ensuite correctement. Pour contrôler ce *sur-apprentissage*, il ne doit pas y avoir de redondance d'information entre le jeu d'entraînement et le jeu de test, qui n'est jamais "vu" par le modèle y compris pour la recherche des meilleurs hyperparamètres.

Pour évaluer au mieux la robustesse du modèle face à ce risque de sur-apprentissage, une méthode appelée *validation croisée* (ou "cross validation" en anglais) est habituellement utilisée, elle correspond à découper le jeu d'entraînement du modèle en différents "plis" ou "fold" de répartition jeu d'entraînement/validation qui permet de montrer que le modèle ne dépend pas trop du jeu sur lequel il a été entraîné. Les modèles étant lourds à entraîner d'un point de vue computationnel, le nombre de validations croisées est en général autour de 5.

Enfin, lors de l'initialisation de tous les poids d'un modèle, il est fréquent d'utiliser la même *graine aléatoire* ("random seed" en anglais, qui permet de générer des poids de façon pseudo-aléatoire et déterministe) pour assurer la reproductibilité des résultats d'une expérience à l'autre. Néanmoins, comme le précise Bethard et al. [19], il est également recommandé de tester différentes *graines aléatoires*, également pour s'assurer de la robustesse du modèle.

***Métriques utilisées et intervalle de confiance***

Pour s'assurer que le modèle "converge" bien, c'est-à-dire que la fonction d'erreur diminue bien et atteint un plateau, plusieurs métriques sont suivies en parallèle (VP correspond à Vrais Positifs, FP à Faux Positifs, FN à Faux Négatifs):

- La *précision* (ou valeur prédictive positive) Précision(P) = VP/(VP + FP) : correspond au taux de prédictions correctes des caractéristiques d'intérêt parmi l'ensemble des caractéristiques trouvées. Il s'agit d'une mesure de la pertinence ;
- Le *rappel* (ou sensibilité) Rappel(R) = VP/(VP+ FN) : correspond au taux de prédictions correctes des caractéristiques d'intérêt parmi l'ensemble des prédictions de ces caractéristiques. Il s'agit d'une mesure de l'exhaustivité.
- Le *score F1* qui combine ces deux métriques, il s'agit de la moyenne harmonique des deux :

$$F = 2\ (P*R)/(P+R)$$

Qui mesure donc à la fois la pertinence et l'exhaustivité.

Ces mêmes métriques sont ensuite utilisées sur le jeu de validation pour le choix des hyperparamètres, puis également ensuite sur le jeu de test pour évaluer définitivement le modèle.

En fonction des applications, plusieurs méthodes de calcul des intervalles de confiance sont possibles. Mais la méthode de calcul de l'intervalle de confiance la plus utilisée dans la littérature est le "bootstrapping" qui correspond à une méthode statistique pour estimer la variabilité d'une métrique (précision, rappel, F1 score par exemple). Des échantillons aléatoires du jeu de données sont sélectionnées itérativement. Les métriques d'intérêt sont ensuite calculées sur chaque échantillon. Ces étapes de tirages et calculs des métriques sur échantillons sont répétés un grand nombre de fois, permettant d'estimer un intervalle de confiance autour de la métrique d'intérêt. L'intérêt principal de cette méthode est de ne pas avoir d'a priori sur la distribution de la métrique.

#### 2.1.4.3. Les modèles de langage

***Word2vec [20]***

Les premiers modèles de langage à partir de réseau de neurones profonds ont été développés dans les années 2010 (notamment Word2vec[20] en 2013 et GloVe[21] en 2014). Ils représentent les mots

par des vecteurs numériques (donc des suites de chiffres, de taille fixe), qui sont appelés “embeddings” de mots.

L’objectif de ces modèles est d’apprendre à prédire les mots en fonction de leur contexte (c’est-à-dire l'ensemble des autres mots environnants, à la manière d’un texte à trous), à partir d’un large corpus de texte.

Schématiquement, les étapes de ces modèles sont les suivantes :

1. Il existe une première étape de *collecte des données*, un corpus - le plus volumineux possible- de texte est constitué à partir de pages web (notamment l’ensemble de Wikipédia[ref] par exemple), des articles de presse, des livres etc… Ces textes sont fournis au modèle sans étiquetage, donc sans supervision.
2. Il existe ensuite une deuxième étape de *pré-traitement des données,* le corpus de textes est nettoyé (suppression des caractères spéciaux etc…) et pré-traité. Le pré-traitement correspond à une étape de *tokenisation* (division du texte en unité linguistiques-appelées *tokens*- telles qu’un mot, une ponctuation, un symbole, un morceau de mot etc.. ), à une étape de découpage du corpus en suite de phrases, à la suppression de la ponctuation, etc.
3. La troisième étape correspond à la constitution du vocabulaire, à partir des données pré-traitées. Ce vocabulaire correspond à l’ensemble des mots uniques auxquels sont associés un identifiant unique et est utilisé pour créer la représentation vectorielle des mots.
4. La quatrième étape correspond à la création des *contextes* : pour chaque mot du texte, on associe un contexte c’est-à-dire un nombre de mots avant et après le mot en question. La taille de ce contexte est un paramètre réglable. Par exemple dans la phrase : “ Le patient diabétique est traité par insulinothérapie”, le mot “diabétique” a pour contexte “le”, “patient”, “est”, “traité”, “par”, “insulinothérapie”.
5. La cinquième étape correspond à l’apprentissage du modèle. L’objectif du modèle étant d’apprendre à prédire un mot en fonction de son contexte (dans le schéma “continuous bag of words” comme présenté figure 3), la fonction de coût (qui calcule l’erreur) cherche à maximiser la probabilité d’apparition du mot en fonction des autres mots environnants. Ceci permet d’obtenir, après itérations sur tout le corpus de texte, la meilleure représentation vectorielle pour chaque mot.
6. La dernière étape consiste à l’obtention des représentations vectorielles de mots (correspondant à la dernière couche cachée (couche avant la couche “Y” dans le figure 2) en sortie du réseau de neurones entraînés), pour les mots du vocabulaire.

Ces vecteurs ont ensuite des propriétés particulièrement intéressantes : en particulier, les synonymes sont notamment proches dans l'espace en termes de distance vectorielle.

L'inconvénient de ces modèles réside essentiellement dans le fait qu'ils sont entraînés sur un corpus fixe et ne savent pas fournir de représentation de nouveau mot qui serait hors de leur vocabulaire d'entraînement. Par ailleurs, les homonymes ont la même représentation vectorielle en dépit du fait qu'ils n'ont pas le même sens.

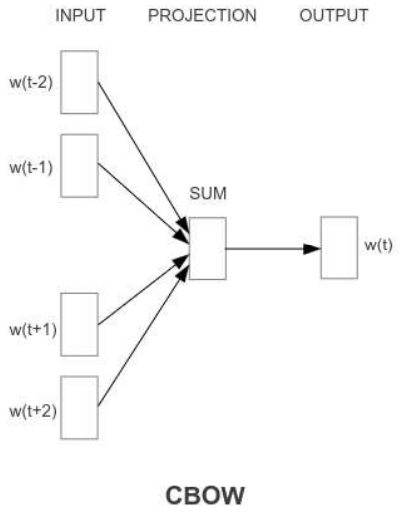



**Figure 3**: Architecture "continuous-bag-of-word" du modèle Word2vec. L'objectif de l'entraînement est l'apprentissage d'une représentation vectorielle des mots qui permettent de prédire un mot (w(t)) en fonction de son contexte (c'est-à-dire des mots proches dans le texte : w(t-2), w(t-1), w(t+1), w(t+2)) [20]

***Le modèle FastText [22]***

Pour pallier la dépendance au vocabulaire, une solution apportée par Bojanowski et al. [22] propose l'apprentissage de la représentation vectorielle non seulement des mots mais également des parties de mots, c'est-à-dire toutes les suites de chaînes de caractères dans les mots - également appelé *n-gram*. Ainsi, pour le mot "eating" présenté figure 4, il existe une représentation vectorielle pour tous les 3-gram (suite de 3 caractères consécutifs dans le mot), le 4-gram etc.. L'avantage de la modélisation des morceaux de mots est que d'une part, cette méthode permet d'obtenir une représentation y compris pour les mots hors vocabulaire, mais qu'elle permet également d'obtenir des représentations plus proches pour les mots comportant la même racine, par exemple "néphrite" en médecine, sera relativement proche des mots "pyélonéphrite", "glomérulonéphrite", etc…

Néanmoins, ce modèle de langage présente le même défaut d'identité des représentations des homonymes.


| Word | Length(n) | Character n-grams |
|---|---|---|
| eating | 3 | <ea, eat, ati, tin, ing, ng> |
| eating | 4 | <eat, eati, atin, ting, ing> |
| eating | 5 | <eati, eatin, ating, ting> |
| eating | 6 | <eatin, eating, ating> |

**Figure 4** : Exemple de division des mots en partie de mots à la base de modèle de langage FastText[22] (source: https://amitness.com/2020/06/fasttext-embeddings/)

Ce modèle a été utilisé dans notre article présenté section 4.6.

***Le modèle BERT [4]***

Le modèle BERT [4] (Bidirectional Encoder Representations from Transformers) permet de pallier à la fois la dépendance au vocabulaire d'entraînement mais également le problème des homonymes. Le principe global de ce modèle est que la représentation vectorielle du mot est cette fois calculée à partir des informations de toute la phrase. Par exemple, dans la phrase “je vais à la banque retirer de l’argent” et dans la phrase “j’utilise une volumineuse banque de données cliniques”, la représentation vectorielle des deux mots “banques” sera différente.

Comme les modèles précédents, ce modèle est pré-entraîné sur un très grand corpus de texte. Ce pré-entraînement est non supervisé (sans supervision manuelle) et a pour objectif la prédiction de mots *masqués* dans une phrase en se basant sur le contexte des mots environnants.

L’architecture du réseau de neurones est un modèle Transformer correspondant à un type de réseau de neurones dont l’élément central est la couche d’attention[23]. La couche d’attention ou *self-attention* ou *attention multi-tête* permet au modèle de prendre en compte les relations entre les mots (ou tokens) dans une phrase, pour chaque mot individuellement traité. Concrètement, les scores d’attention sont obtenus par produit scalaire entre le vecteur représentant le mot actuel et les vecteurs représentant les autres mots dans la phrase. Ces scores sont ensuite passés à une fonction *softmax*, qui normalise les scores en leur donnant une interprétation de probabilité.

Pour permettre ce fonctionnement d’observation à l’échelle de la phrase, trois informations d’entrée sont encodées et fournies au modèle : l’embedding du token, l’embedding de la phrase - avec une taille maximale de 512 tokens, et l’embedding de position du token dans la phrase.

Sur l'exemple de la figure 5 schématisant ce mécanisme, dans la phrase "the rabbit quickly hopped" le score d'attention est maximal entre la représentation du vecteur du mot "rabbit" (lapin) et celui du mot "hopped" (sauter). De même le terme entre le terme "crawled" (ramper) et "turtle" (tortue).

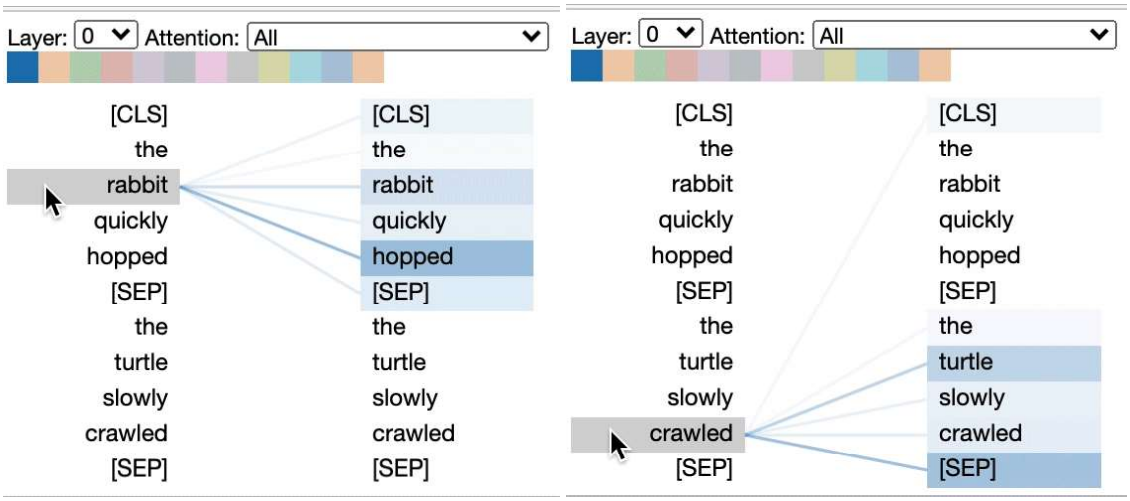



**Figure 5**: illustration du mécanisme d'attention (https://raw.githubusercontent.com/jessevig/bertviz/)

Cette étape d'attention dans le modèle fournit donc une représentation vectorielle du mot en fonction des mots environnants. En utilisant des couches d'attention empilées, l'architecture Transformer peut intégrer des dépendances à plus longue distance dans les séquences de texte, ce qui permet de capturer l'information syntaxique et sémantique des phrases. Néanmoins, ces dépendances à plus longue distance sont limitées par la taille maximale de la phrase définie par construction à 512 tokens (pour que la couche d'attention, qui réalise donc un produit scalaire avec des vecteurs de taille 512 ne soit pas trop lourde computationnellement). Le contexte considéré pour la prédiction d'un mot ne dépasse donc pas cette longueur. Cette limite est discutée par la suite pour la représentation vectorielle des documents à l'aide de cette architecture.

Le pré-entraînement du modèle correspond à l'étape d'apprentissage de la représentation vectorielle des mots dans leur contexte, et permet d'obtenir un modèle avec tous ses poids calculés. Après ce pré-entraînement, le modèle est adapté à des tâches spécifiques (comme la classification de texte, la reconnaissance d'entités nommées, le résumé automatique) grâce à un processus appelé "*fine-tuning*" ou *affinage*. Dans cette étape, les poids des dernières couches du modèle sont modifiés sur un jeu de données annoté. Le modèle affiné peut ensuite être utilisé pour générer des prédictions. L'ensemble de ces étapes sont illustrées sur la figure 6.

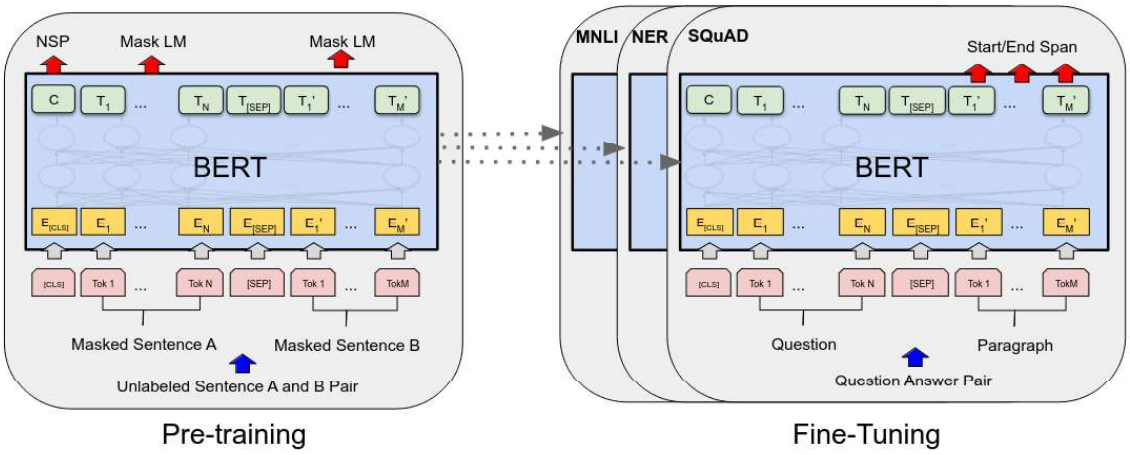



**Figure 6**: Représentation schématique du pré-entrainement (à gauche) et de l'affinage en fonction de la tache (à droite) du modèle BERT [4].

Ce modèle BERT appartient au modèle de langage qu'on appelle "masqué" (Masked language Model) qui apprend à prédire des mots masqués au sein d'une phrase, à partir des mots précédents et suivants. Cet apprentissage est différent des modèles de langage entraînés pour prédire le mot suivant seulement à partir des mots précédents tel que GPT-3 [5] et qu'on appelle modèle auto-régressif. ChatGPT[6]est la version affinée par apprentissage supervisé de GPT-3 [5].

Pour donner une idée d'échelle, le modèle BERT [4] possède 340 millions de paramètres et a été pré-entraîné sur un corpus de 3.3 milliards de mots ; GPT-3 [5] possède 175 milliards de paramètres et a été pré-entraîné sur 300 milliards de mots.

Ce modèle a été utilisé, sous différentes versions, dans l'ensemble des articles présentés sections 4, notamment pour les étapes d'extraction d'entités nommées et pour la classification et la normalisation des termes.

#### 2.1.4.4. Modèles de langage spécifiques au domaine biomédical :

Les modèles présentés ci-dessus ont été pré-entraînés sur des domaines linguistiques généraux (comme les articles de presse, Wikipédia etc, ..). En fonction de la langue du corpus d'entraînement, ces modèles peuvent être spécifiques à une langue. En français, deux modèles ont été développés : CamemBERT [24] et FlauBERT [25] ou multilingue comme le BERT multilingue [4].

Au-delà de la spécialisation d' une langue, ces modèles peuvent également être entraînés spécifiquement sur des domaines scientifiques - qui possèdent leur propre sémantique- avec une importante amélioration des performances sur différentes tâches de TAL. Ces modèles peuvent être soit issus d'un ré-entraînement des poids d'un premier modèle de langage, soit entraînés depuis le départ (*from scratch*), avec des poids initiaux aléatoires.

Dans le domaine biomédical les premiers modèles développés sont notamment le BioBERT [26] développé sur l'ensemble des articles biomédicaux (PubMed abstracts et PMC fill-text articles) et le *Clinical BERT,* affiné par Alsentzer et al. [27] sur 2 millions de notes cliniques d'une importante base de données américaine accessible : MIMIC III [28] avec des performances nettement améliorées, notamment pour l'extraction d'entités nommées.

Au cours de nos analyses présentées section 4, en particulier section 4.4., nous avons pu comparer ces modèles anglais avec nos propres modèles.

S'agissant du domaine biomédical en langue française, l'accès à des modèles publics performants reste plus complexe du fait d'un bien moins grand volume de données accessibles pour l'entraînement. Des modèles prometteurs ont été proposés récemment tels que le CamemBERT-bio [29] et le Dr-BERT [30]. Un modèle BERT entraîné "from scratch" c'est-à-dire depuis l'initialisation sur 27 millions de documents de l'entrepôt de données de santé de l'AP-HP [31] nous a également été accessible pour la dernière année de la thèse et utilisé pour les algorithmes présentés section 4.7.

D'autres modèles de transformer, plus larges, tel que le longformer [32], capable de dépasser la limitation des 512 token du BERT classique- qui correspond donc à un fenêtrage plus court, ont également été ré-entraîné sur des corpus médicaux [33] avec des résultats supérieurs à l'état de l'art jusqu'alors, mais n'existe pas en français à notre connaissance d'une part et ont un coût computationnel très élevé d'autre part : l'expérience de Li et al. [33] a par exemple été réalisée sur 4 GPUs de 32 Gigabytes (GB) de mémoire, tandis que dans notre espace, nous disposons d'un GPU de 32 GB. Ces modèles plus larges n'ont pas pu être testés sur notre espace restreint sur le cluster de l'entrepôt de données de santé.

Après ces développements sur les modèles de langages appliqués à la médecine, il convient de détailler l'état de l'art pour les différentes étapes de l'algorithme d'extraction automatique de cohorte de patients similaires.

## 2.2. Extraction de concepts médicaux

### 2.2.1. Extraction des concepts médicaux

L'extraction de concepts médicaux depuis les textes correspond à la tâche spécifique du TAL de *reconnaissance d'entités nommées*. Dans le cadre médical, il s'agit d'extraire des mentions dans le texte relevant de types sémantiques d'intérêt tel que les traitements, les signes et symptômes, les comorbidités, les procédures diagnostiques et thérapeutiques, les constantes cliniques (poids, taille, pression artérielle, etc…), *etc.*. Un exemple d'extraction d'entités dans le texte est fourni figure 7.

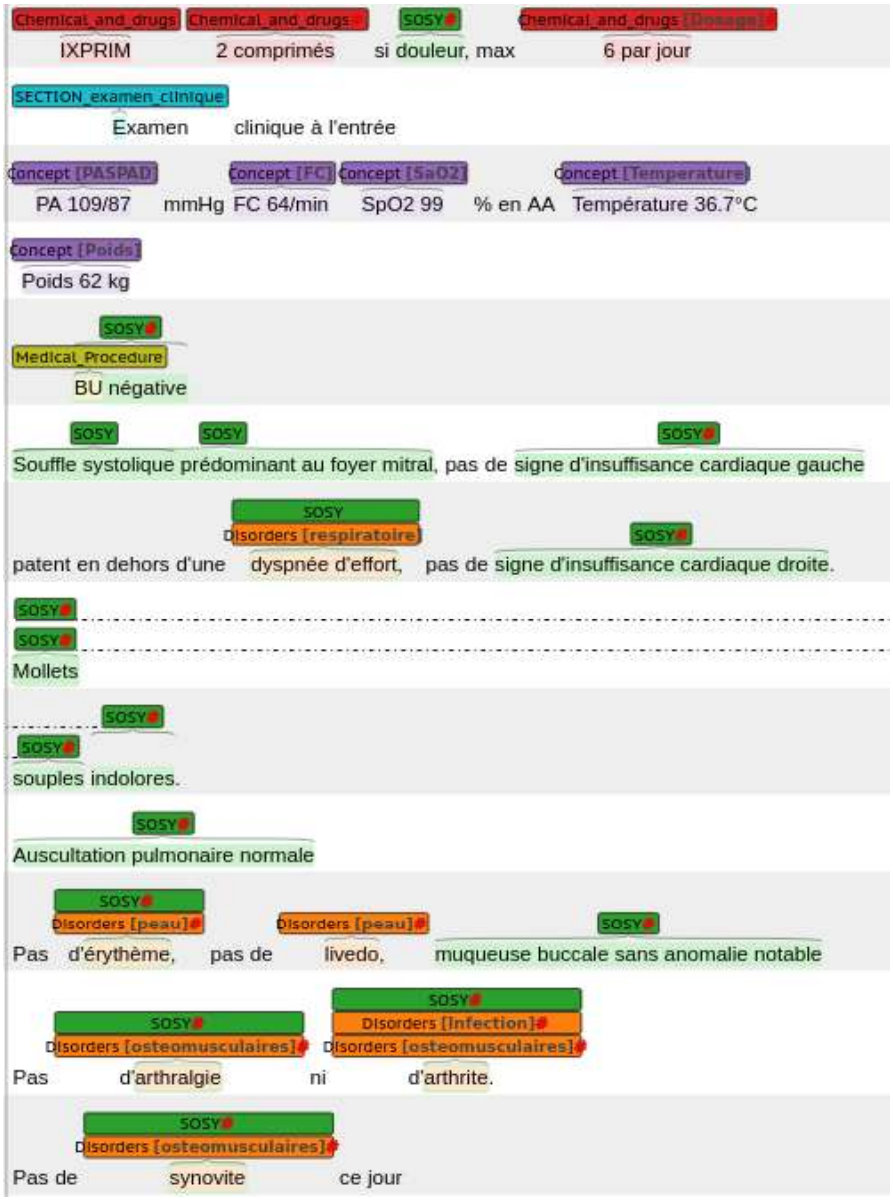


**Figure 7**: exemple d'extraction d'entités dans le texte clinique, ici, à partir d'une annotation manuelle. "*Chemical and drugs"* sont les traitements, "*Disorders*" les pathologies, "*concept*" les constantes cliniques, "*SOSY*" les symptômes.

#### 2.2.1.1. L'encodage du texte

Plus concrètement, l'identification des parties de texte (ou *spans*) d'intérêts consiste à délimiter le début et la fin des mentions dans la phrase. Le problème peut donc être vu comme une problème combinatoire de classification de tous les couples possibles (localisation/type sémantique) pour chaque mot dans le texte où la localisation est encodée de plusieurs façons dans la littérature : une des plus connues est l'encodage B-I-O pour Begin-Inside-Out (début, milieu et extérieur au concept à extraire). Il existe également un schéma dérivé de ce dernier : le schéma B-I-O-U-L pour Beginning (début de la mention)-Inside (milieu de la mention)-Out(hors mention)-Unit (mention unitaire)-Last(fin de mention) qui permet un encodage plus complet des mentions constituées de plusieurs mots. Par exemple, dans la phrase suivante : "Le patient présente une glomérulonéphrite extra-membraneuse lupique et est traité par Endoxan." l'encodage est le suivant :

-Le (O)- patient(O)- présente (O)- une (O)- glomérulonéphrite (B-Maladie)- extra-membraneuse (I-Maladie) - lupique (L-Maladie)-et(O)-est(O)-traité(O)- par(O)- Endoxan (U-Médicament).

#### 2.2.1.2. Les différentes approches :

Les approches les plus courantes pour réaliser cette tâche sont les suivantes :

1. Les approches basées sur des *règles heuristiques*, qui peuvent être basées sur un ensemble d'expressions régulières ou des modèles linguistiques. Les exemples dans le domaine biomédical sont par exemple les algorithmes tel que cTakes (clinical Text Analysis and Knowledge Extraction System) [34] (component implements a terminology-agnostic dictionary look-up algorithm within a noun-phrase look-up window). Ces méthodes peuvent être très performantes mais nécessitent un important travail d'ingénierie initiale pour établir les règles et les entretenir.

2. Les approches basées sur l'*apprentissage supervisé* (c'est-à-dire à partir d'une annotation dans le texte, réalisée manuellement telle que présenté figure 7). Ces approches ont évolué

au cours du temps, les premières approches d'apprentissage étaient basées sur des algorithmes tels que les *machines à vecteurs de support* (ou support vector machine en anglais), *les chaînes de markov cachées*, les *champs aléatoires conditionnels* (ou *conditional random field*-CRF en anglais) ou des réseaux de neurones combinants des CRF [35] et des Long-Short-Term-Memory (LSTM)[36] [37,38]. Pour simplifier la lecture, les modèles de CRF[35] et de LSTM [36] sont décrits en annexes. Ces dernières approches nécessitent des descripteurs des textes d'entrée tels que les caractéristiques orthographiques, la place des majuscules, les racines des mots d'entrée *etc.*. [39]

D'autres méthodes d'extraction d'entités nommées, plus récentes, utilisent les modèles de langages. Ces méthodes sont basées sur les représentations vectorielles des mots (ou -*word embeddings*) issus des modèles de langage pré-entraînés pour encoder le texte d'entrée, cet encodage est ensuite transmis à une couche de neurone de classification entraînée à reconnaître les concepts dans le texte c'est-à-dire un couple (position de mention/type sémantique). Les modèles de langage issus de l'architecture BERT [4] tels que BioBERT[26] et ClinicalBERT[27] par exemple, sont actuellement les plus performants [39] y compris par rapport au modèle GPT-3[5] à l'origine du célèbre chatGPT [40, 41], avec une F1-score de 0.94 sur le jeu de donnée international coNLL2003 [42, 40].

Néanmoins, ces modèles sont très dépendants du volume et de la distribution des données d'apprentissage et manquent de capacité de généralisation lorsqu'ils sont appliquées à des nouvelles données [39]. La création du jeu d'apprentissage est également très coûteuse en temps et nécessite d'être réalisée par les experts du domaine. Par ailleurs, il existe une forte disparité sur le volume des jeux de données accessibles entre les différentes langues, avec une surreprésentation très importante de l'anglais, via notamment des défis annuels tel que le *national NLP clinical challenges* (n2c2)[43].

A partir de cet encodeur BERT [4], plusieurs architectures complémentaires ont été proposées pour le décodage (c'est-à-dire pour la couche de sortie de prédiction de la mention dans le texte). On peut citer Yu et al. [44], Wajsburt [45] et Bannour et al [46] qui ont notamment proposé une couche de décodage par un Bi-LSTM [36], comme présenté figure 8. L'architecture présentée est particulièrement adaptée pour extraire les entités imbriquées les unes dans les autres (on parle de *Nested NER*, à l'opposé du *flat NER* ou NER "*plat*" où chaque entité est disjointe). Ici, la dernière couche ("*bound matcher"* sur la figure) est une couche qui calcule la probabilité que la mention candidate explorée corresponde effectivement à une entité. Cette probabilité est calculée par produits scalaires multiples, un pour chaque type sémantique (médicaments, symptômes, etc.. d'intérêt).

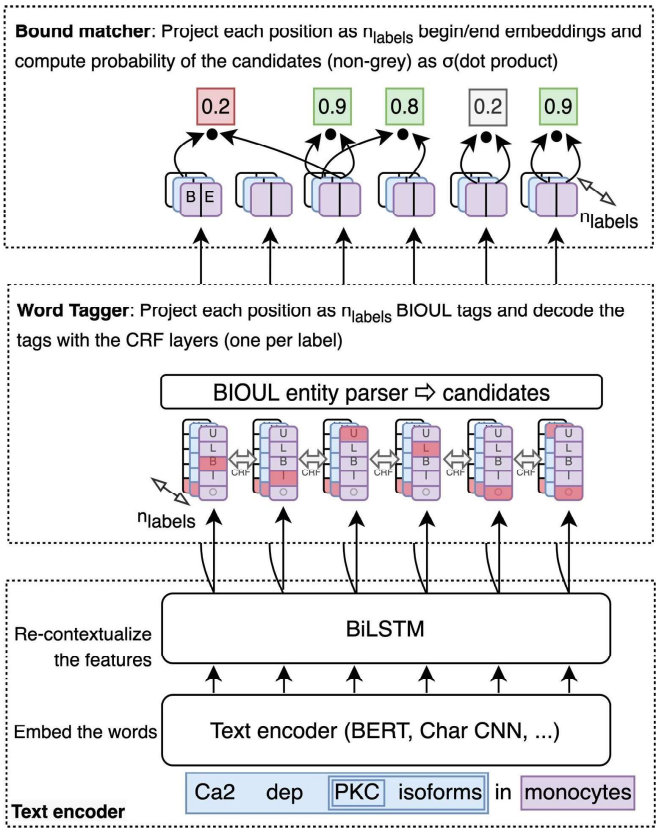



**Figure 8**: Architecture complète d'un modèle d'extraction d'entité nommée supervisé avec l'utilisation d'un encodeur BERT, d'un Bi-LSTM et d'un schéma BIOUL pour extraire les entités d'un type données ici (au nombre de *n*) dans cette figure [46]. Ce modèle a été utilisé pour l'étape d'extraction d'entités nommées de la création de cohorte de patients similaires (sections 4.4. et 4.6).

3. Les approches utilisant des *combinaisons de méthodes* comme proposé par exemple par Kraljevic et al. avec l'algorithme Medical Concept Annotation Tool [47]. Cet outil utilise une approche basée sur le dictionnaire UMLS [16] (ou la SNOMED-CT [12] ) pour permettre une extraction d'entités nommées qui se fait en plusieurs étapes : d'abord une vérification orthographique et un nettoyage des mots, puis une tokenisation et une lemmatisation. Enfin, à partir d'une fenêtre glissante sur tout le texte, chaque mot est comparé au vocabulaire UMLS [16] : si le mot est effectivement un concept de l'UMLS [16] , les représentations vectorielles du mot et du contexte (initialement à partir de l'algorithme Word2vec[20] et actuellement proposé avec un algorithme BERT [4]) sont utilisées, d'une part, pour filtrer si le concept est nié, hypothétique ou appartenant à une autre personne que le patient et, d'autre part, pour permettre une désambiguïsation (en cas d'homonyme notamment, de type sémantique différent : par exemple, "albumine" dans le texte clinique peut se référer à la fois à un test biologique ou à un traitement). L'avantage de cette approche est qu'elle permet de

traiter l'extraction de l'entité et sa normalisation dans le même temps. Les performances de ces modèles pour l'extraction restent néanmoins surtout intéressantes en langue anglaise du fait notamment du nombre bien plus important de concepts UMLS dans cette langue.

Une autre méthode proposée par Silvestri et al. [48]propose un modèle performant à partir d'un faible jeu annoté manuellement et d'une annotation à partir de terminologie pour permettre un rééquilibrage sur les types sémantiques très peu représentés.

4. Enfin, du fait du fardeau que l'annotation peut représenter, il existe également des approches basées sur l'*apprentissage non supervisé* qui ne nécessitent pas de données annotées. Il s'agit notamment du "zero-shot learning", c'est-à-dire de l'*apprentissage sans exemple* qui permet à un modèle d'apprendre à reconnaître des nouvelles classes, sans avoir reçu d'exemple spécifique. Ces modèles présentent à l'heure actuelle une moins bonne performance que les modèles en apprentissage supervisé [49, 50, 41].

### 2.2.2. Extraction des attributs

En traitement du langage naturel, il est nécessaire de qualifier les entités par des attributs. Par exemple, pour extraire les patients traités par un médicament donné, il est nécessaire de s'assurer que la mention rattachée à ce médicament dans le texte n'est pas niée ("Paracétamol suspendu") ou prescrit à un membre de la famille du patient par exemple.

Ces attributs peuvent être de complexité diverse en fonction de la question posée, les plus classiques sont donc la négation/parenté(autre personne mentionnée dans le compte-rendu)/hypothèse ; toutefois, concernant les médicaments, le dernier défi international n2c2 2022 [51] proposait un bien plus grand nombre d'attributs tels que l'*action* sur les médicaments (début, fin, dose unique, augmentation ou diminution de la posologie, etc..), l'*acteur* sur le médicament (clinicien ou patient), la *temporalité* (passée, présent, future), la *certitude* (prescription certaine, hypothétique ou conditionnelle par exemple à un symptôme) et la *négation*.

L'extraction des attributs est une tâche généralement plus complexe que l'extraction de l'entité. A nouveau, il existe plusieurs types de modèles. Negex [52] par exemple est un modèle de détection des négations basé sur des règles heuristiques.

En apprentissage supervisé beaucoup d'architectures de modèles ont été proposées. Par exemple, Nath et al. [53], ont proposé et testé trois architectures de modèles différents pour extraire les

entités, l'attribut de négation et les attributs de *modalités* pour un événement avec les types possibles suivants : se produit réellement/ est simplement proposé/ est mentionné au conditionnel/ est décrit comme possible, proposés par la tâche i2b2 2012 [54]. Ces architectures sont basées sur l'utilisation d'embeddings BERT [4] puis sur un encodage par une couche de Bi-LSTM[36] et soit de multiples CRF [35], soit de CRF[35] pour prédire l'entité et d'une couche Softmax puis *Teacher Forcing* (le *teacher forcing* est une technique d'apprentissage supervisée qui fournit, lors de la phase d'entraînement, la cible attendue comme entrée au modèle). En d'autres termes, au lieu de laisser le modèle générer une prédiction à chaque étape en se basant sur sa propre sortie précédente, on lui fournit la véritable séquence de sortie attendue comme référence à chaque étape. Cela permet d'accélérer l'apprentissage initial du modèle en lui fournissant des informations précises sur la tâche à accomplir. Malgré leur complexité, ces architectures restent peu performantes pour prédire des attributs complexes de modalités (F1 score maximum à 0.5).

Pour la tâche n2c2 2022 [51] un autre exemple de solution a été envisagé par Ramachandran et al. [55] qui propose une architecture en deux étapes : une première étape d'extraction des médicaments (les entités nommées dans cette tâche précise), et ensuite une étape d'extraction des attributs, des médicaments à partir de l'embedding de la phrase complète. L'extraction des entités est réalisée par une architecture utilisant un BERT avec une couche linéaire en sortie et une étape d'affinage pour la tâche précise. Une fois les médicaments détectés dans le texte, un embedding BERT [4] de la phrase est calculé avec ajout de token spéciaux autour de la mention du médicament-cible uniquement, avec une couche linéaire en sortie. Ces tokens spéciaux permettent de délimiter la mention d'intérêt au modèle pour prédire au mieux ses attributs permettant d'obtenir une F1 score globale de 0.83. Néanmoins, cette méthode utilisant deux architectures BERT [4] est particulièrement coûteuse en termes computationnels.

## 2.3. Normalisation des concepts médicaux

Une fois la mention d'intérêt extraite du texte, une étape importante pour permettre une homogénéisation de synonymes, acronymes etc. est l'étape de normalisation de la mention, c'est-à-dire son rattachement à un concept dans une base de connaissance. Un exemple de normalisation du terme " cancer du sein" est proposé figure 9.

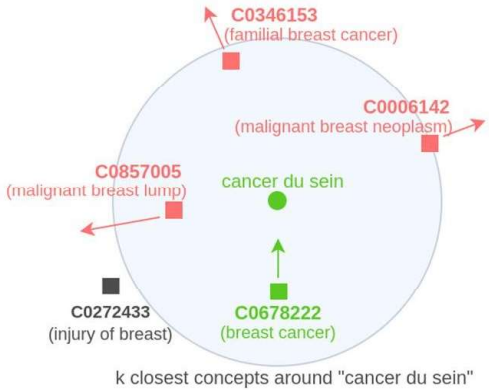



**Figure 9**: exemple de normalisation du terme “cancer du sein” à partir des concepts de la base de connaissance UMLS[16] , d’après Wasjburt et al. [56].

Là encore, plusieurs solutions existent telles qu’initialement les approches par règles heuristiques et reconnaissance de chaînes de caractères proposées par Afzal et al [57] ou D’Souza et al. [58].

Plus récemment, des méthodes basées sur l’apprentissage ont été développées. Il s’agit par exemple de DNorm[59], basé sur un algorithme d’apprentissage de classement par paire, initialement utilisé par les systèmes de recherche d’informations. L’objectif de l’apprentissage est de maximiser un score de similarité entre une représentation vectorielle de la mention extraite du texte à normaliser et une représentation vectorielle du concept auquel elle est liée dans la base de connaissance biomédicale.

Les méthodes actuelles de normalisation les plus performantes sont celles utilisant les larges modèles de langage tel que le BERT[4]. On peut citer par exemple Wajsburt et al [56] qui définit la question de normalisation non pas comme une recherche des plus proches voisins ou par paire mais comme un problème de classification à c classes où c correspond à l’ensemble des concepts dans l’UMLS [16] - c’est à dire une classification à plus d’un million de classes. L’objectif de l’apprentissage est donc d’apprendre la probabilité pour une mention *m* donnée d’appartenir à la classe *c* du concept associé. L’architecture utilisée est un encodage BERT[4] multilingue et une fonction de similarité cosinus passée à une fonction softmax en sortie. Cette méthode présente notamment l’avantage d’être multilingue et de s’appuyer sur l’ensemble des termes anglais dans l’UMLS[16] .

Enfin l'algorithme CODER [60], présenté figure 10, correspond à un modèle dont l’objectif est l’apprentissage d’embedding des termes de l’UMLS[16] pour permettre une meilleure normalisation. Cet apprentissage est réalisé par “apprentissage par contraste” ou *contrastive learning* en anglais qui vise à créer des représentations des concepts en maximisant la similarité entre des exemples positifs

(issus de la même classe ou ayant des caractéristiques similaires) tout en minimisant la similarité entre des exemples négatifs (issus de classes différentes ou ayant des caractéristiques différentes). Les exemples positifs et négatifs sont directement calculés à partir des relations entre les concepts dans l'UMLS [16]. Deux versions de cet algorithme sont disponibles : une anglaise, à partir des termes anglais de l'UMLS [16] et d'un fine-tuning du PubMedBERT[26] et une version multilingue à partir du BERT multilingue [4].

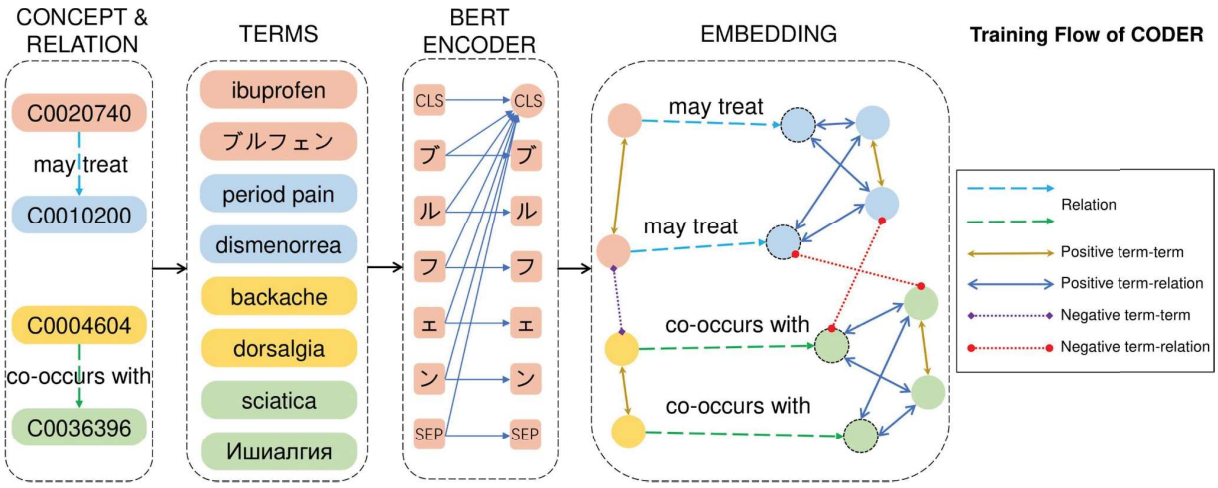



**Figure 10**: Fonctionnement de l'algorithme CODER[60] de normalisation des termes au CUI de l'UMLS[16] , à partir d'embedding appris à partir des informations textuelles et de relation.

Les deux dernières méthodes proposées ci-dessus correspondent aux meilleurs résultats de l'état de l'art, ils ont été comparés dans nos analyses présentées section 4.4. et utilisés également d'une part pour la pré-sélection de cohorte (section 4.3) et pour la normalisation des termes de biologie (section 4.7). Néanmoins, cette étape de normalisation de la mention extraite du texte est réalisée hors contexte et nécessite l'information du type sémantique de la mention pour obtenir les meilleures performances.

## 2.4. Recherche de document et similarité

En dehors du travail minutieux d'extraction et de normalisation des mentions d'intérêt dans le texte, qui présente par ailleurs un important coût computationnel, il peut être intéressant de réaliser en amont une classification moins fine, directement à l'échelle du document. Cette étape peut être vue

comme une étape de pré-screening des documents d'intérêt. A titre d'exemple, dans le jeu de données MIMIC III [28], il existe une tâche de classification des comptes-rendus anatomopathologiques des tumeurs avec différentes classes : identification du site, identification du sous-site, latéralité, comportement (tumeur invasive ou non) et grade.

Pour cette tâche, une représentation vectorielle non plus de chaque mot, mais de l'ensemble d'un texte a été développée à partir du modèle word2vec : doc2vec[61] où le modèle apprend à prédire le prochain paragraphe. Il existe également le modèle DocBERT[62], avec le même objectif d'entraînement, à partir de l'architecture BERT [4]. L'inconvénient de ces deux derniers modèles est que la longueur maximale des fenêtres de contexte est de taille limitée, notamment à 512 tokens (ou morceaux de mots) pour le modèle BERT qui peuvent donc être les 512 premiers tokens du document ou 512 pris aléatoirement. Cette limitation importante ne permet pas à ces modèles d'être plus performant que des réseaux de neurones convolutionnels simples par Gao et al. [63], et les résultats restent peu satisfaisants. Ces modèles ont été testés dans nos analyses section 4.3.

Comme expliqué par Paaß et al. [64], cette limite de taille de contexte est liée à l'architecture basée sur une couche d'auto-attention des modèles BERT[4]. "Si la longueur de la séquence T est portée à 2T, il faut calculer quatre fois plus d'associations (attentions) entre les tokens." Plusieurs solutions ont été développées pour pallier cet inconvénient, notamment des *matrices d'attention creuses* (ou "sparse matrix attention" en anglais) qui ne calcule l'attention que pour un ensemble réduit d'éléments de la séquence. Les techniques de sélection des éléments les plus importants de la séquence sont généralement déterminées en fonction de leur similarité avec l'élément courant. Cette solution est notamment utilisée par BigBird[65], longformer[32] et GPT-3[5]. Les performances de BigBird[65] et du modèle longformer sont aujourd'hui parmi les meilleures pour la tâche de classification de documents médicaux [33]. Ces trois derniers modèles nécessitent néanmoins des ressources computationnelles très importantes [33].

## 2.5. Phénotypage à partir des textes cliniques

Ces méthodes issues du traitement automatique des langues ayant été détaillées, cette partie vise à développer leurs utilisations pour permettre d'identifier des patients d'intérêt à partir de base de données multi-modales (comprenant des données démographiques, de biologie, issues du codage des pathologies et en particulier des textes des comptes-rendus médicaux et des notes cliniques).

La question posée ici est de pouvoir extraire automatiquement des patients ayant un phénotype précis. Le terme *phénotype,* qui correspond à l'ensemble des caractères apparents d'un individu, est utilisé dans un sens large dans ce contexte : il s'agit des données démographiques, des données biométriques, des informations sur les symptômes, les comorbidités etc.

L'extraction des patients ayant un phénotype donné, permet par la suite de développer des outils de prédiction du devenir patient, d'aide au diagnostic, de pré-sélection pour les essais thérapeutiques ou encore d'aide à la décision thérapeutique.

Au vu de notre objectif de recherche, les méthodes présentées ici utilisent toutes le texte clinique.

En 2018, une solution proposée par Garcelon et al [66], s'appuie sur une représentation *TF-IDF* [67] des termes dans les comptes-rendus pour explorer des associations en des phénotypes cliniques et des maladies rares, suivies à l'hôpital Necker. La méthode automatique d'extraction de phénotypes à partir des notes cliniques est la suivante : les concepts médicaux sont extraits des textes par une méthode basée sur les règles de reconnaissance d'entités issues de l'UMLS [16] ayant un type sémantique précis: "signes ou symptômes", "fonction pathologique", etc.. Le caractère nié du concept, ou rattaché à une autre personne, est filtré par un algorithme à base de règles développées par les auteurs [68]. Pour classer le caractères pertinents des concepts cliniques extraits, les auteurs ont utilisés la métrique *TF-IDF* (Term Frequency- Inverse document frequency) qui calcule la fréquence d'apparition d'un terme (*TF*) pour une maladie donnée par rapport à la fréquence de ce terme pour l'ensemble des document (*Document frequency*). L'objectif ici est donc d'extraire un concept spécifique pour une maladie donnée (ce concept, par exemple un symptôme sera a priori plus fréquent s'il est spécifique à la maladie, et donc davantage discriminant pour permettre un phénotypage précis). Le groupe des cinquante concepts ayant le meilleur TF-IDF est associé en phénotype descriptif d'une maladie donnée. Cette représentation a ensuite permis aux auteurs d'extraire automatiquement d'autres patients de la base de données ayant ce même phénotype, avec une évaluation manuelle de l'extraction par des experts cliniciens et généticiens sur la pertinence des patients extraits.

Plus récemment, également dans un objectif d'aide au diagnostic, van Haken et al. [69] proposent une méthode pour prédire un diagnostic clinique à partir du compte-rendu d'admission, à partir du jeu de données MIMIC III [28] . Les diagnostics à prédire correspondent aux trois premiers chiffres du codage ICD-9 (CIM-9 en français, 9ème version de la classification internationale des maladies), attribués au patient à la sortie de l'hôpital. L'algorithme d'apprentissage profond développé utilise une représentation vectorielle du document du patient, c'est-à-dire un encodeur de document. Pour chaque diagnostic possible, une représentation vectorielle du document prototypique associé est

apprise. L'apprentissage des vecteurs prototypes et de l'encodeur des documents est réalisé conjointement pour permettre une bonne discrimination des patients avec et sans diagnostic. La prédiction diagnostique est réalisée par mesure de similarité entre le vecteur prototype et l'encodage du document. Les auteurs ajoutent également une couche supplémentaire d'*attention* pour mettre en évidence les parties du texte ayant permis la classification diagnostique et obtiennent de bonnes performances : AUC ROC macro 87.94 +- 0.02. La figure 11 présente un schéma de l'architecture.

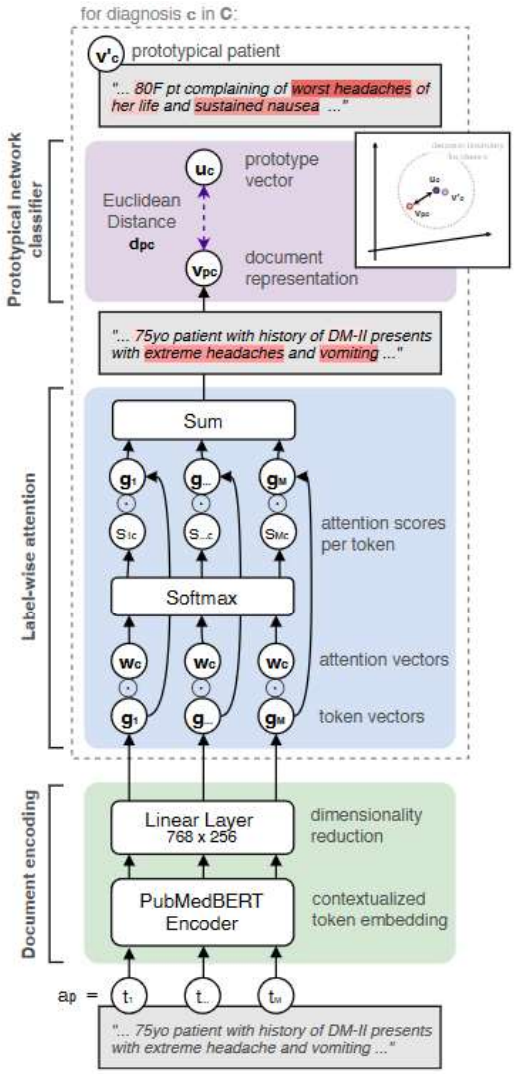



**Figure 11** : Représentation schématique du modèle de prédiction diagnostic à partir des textes développés par van Aken et al. [69].

De Freitas et al. [70] propose également un algorithme de phénotypage patient Phe2vec, qui apprend une représentation vectorielle de tous les concepts médicaux structurés (code CIM10, données de biologies etc..) et non structurés issus du texte. L'histoire clinique du patient est résumée en agrégeant l'ensemble des représentations des concepts médicaux. Cette méthode, bien que performante en terme de valeur prédictive positive pour un certain nombre de phénotypes : anévrisme de l'aorte abdominal (1.00), maladie de Crohn (0.98) , drépanocytose (0.96) etc. par comparaison avec l'évaluation manuelle des notes d'évolution clinique, manque d'interprétabilité pour le clinicien.

Dans un cas d'usage plus précis de détection du risque de démence à partir des comptes-rendus médicaux des patients, Ben Miled et al. [71] ont d'abord sélectionné les mots-clés d'intérêt pour la démence à partir d'un algorithme *TF-IDF* [67] et ont ensuite comparé deux représentations de ces mots-clés. La première, utilise la moyenne des représentations vectorielles issus du modèle de langage pré-entraîné clinical-BERT [27] de l'ensemble des mots-clés d'un compte-rendu hospitalier avec un réseau de neurones en sortie pour prédire le risque de démence à un an, et la deuxième regroupe les mots-clés en fonction des concepts UMLS[16] et utilise chaque concept comme variable d'exposition, avec un algorithme d'apprentissage automatique combinant plusieurs arbres de décision (algorithme pour la résolution des problèmes de classification basé sur une structure arborescente où chaque noeud représente une décision basée sur une caractéristique spécifique-ici le code UMLS du concept- et chaque branche représente un résultat possible). Les meilleurs résultats sont obtenus avec la seconde méthode avec un AUC de 75% pour la précision du risque de démence à un an. Les auteurs notent néanmoins qu'aucune des deux méthodes n'est généralisable à une autre institution médicale que celle pour laquelle elles ont été utilisées.

Les méthodes présentées ici reflètent la diversité et la richesse des solutions proposées ; néanmoins, peu d'applications sont directement utilisées en clinique.

## 2.6. Cohorte de patients similaires dans l'aide à la décision thérapeutique

### 2.6.1. Systèmes basés sur un moteur de recherche avec requêtage :

Une des équipes ayant le plus contribué à la construction et à l'analyse de cohorte de patients similaires est l'équipe d'informatique biomédicale de l'université de Stanford. Dès 2011, elle a proposé une preuve de concept par la recherche automatisée de patients semblables au patient index qu'elle cherchait à traiter [1]. Dix ans plus tard, cette même équipe [72, 73] a proposé le moteur de recherche ACE (Advanced Cohort Engine) qui permet, via un langage de requêtage

explicite, le phénotypage et la construction de cohorte de patients en temps réel. Ce système performant fournit en sortie l'analyse détaillée de la cohorte, y compris les informations temporelles. Par exemple, pour une requête "patients diabétiques sous glipizide sans antécédent d'AVC", chaque patient de la cohorte correspondant dans la base de donnée est représenté par une ligne avec des codes couleurs différents pour chacun des critères de la requête, comme présenté figure 12.

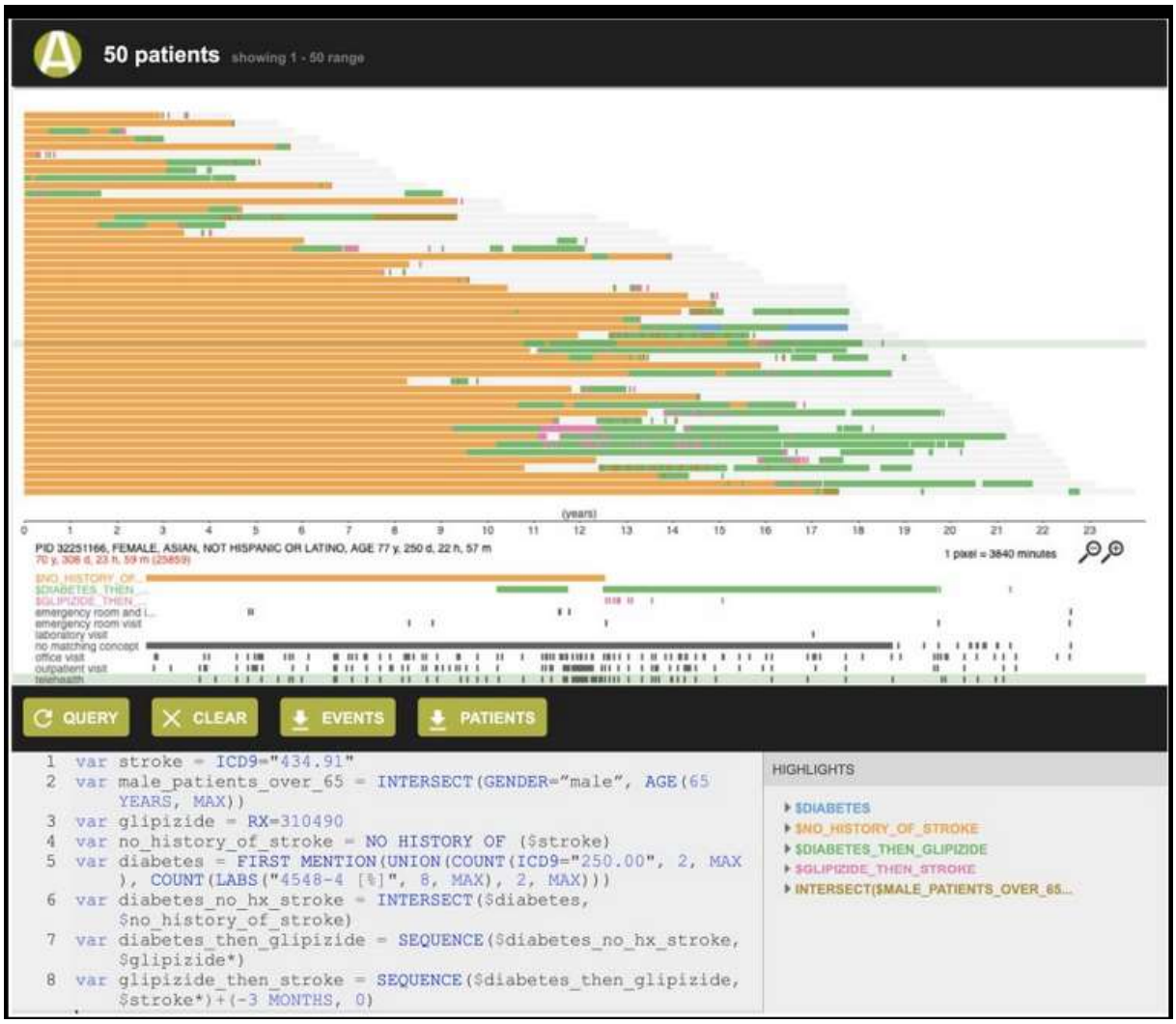



**Figure 12**: Exemple d'extraction de cohorte de patient par requête à partir du moteur de recherche ACE [73].

L'intérêt majeur de ce moteur de recherche est sa rapidité liée à sa structure de base de données interrogeable par TQL (Temporal Query Language). Néanmoins, il est essentiellement basé sur les informations en données structurées (données de codage des pathologies et des médicaments) et les utilisateurs doivent se former à un langage de requête complexe. Par ailleurs, les auteurs constatent que cet outil est davantage utilisé pour la recherche par les cliniciens avec peu d'impact sur la clinique [74]. C'est également ce que souligne un billet récent de l'institut "Human-Centered Artificial

Intelligence" de Standford [75], y compris après la mise à disposition d'un nouvel outil de représentation du dossier patient informatisé inspiré par les techniques de traitement automatisé du langage [76]. Il est important d'ajouter que ce système ACE de consultation automatisée de cohorte de patient nécessite une collaboration pluridisciplinaire d'intervenants médicaux, informaticiens et data scientistes, comme illustré figure 13.

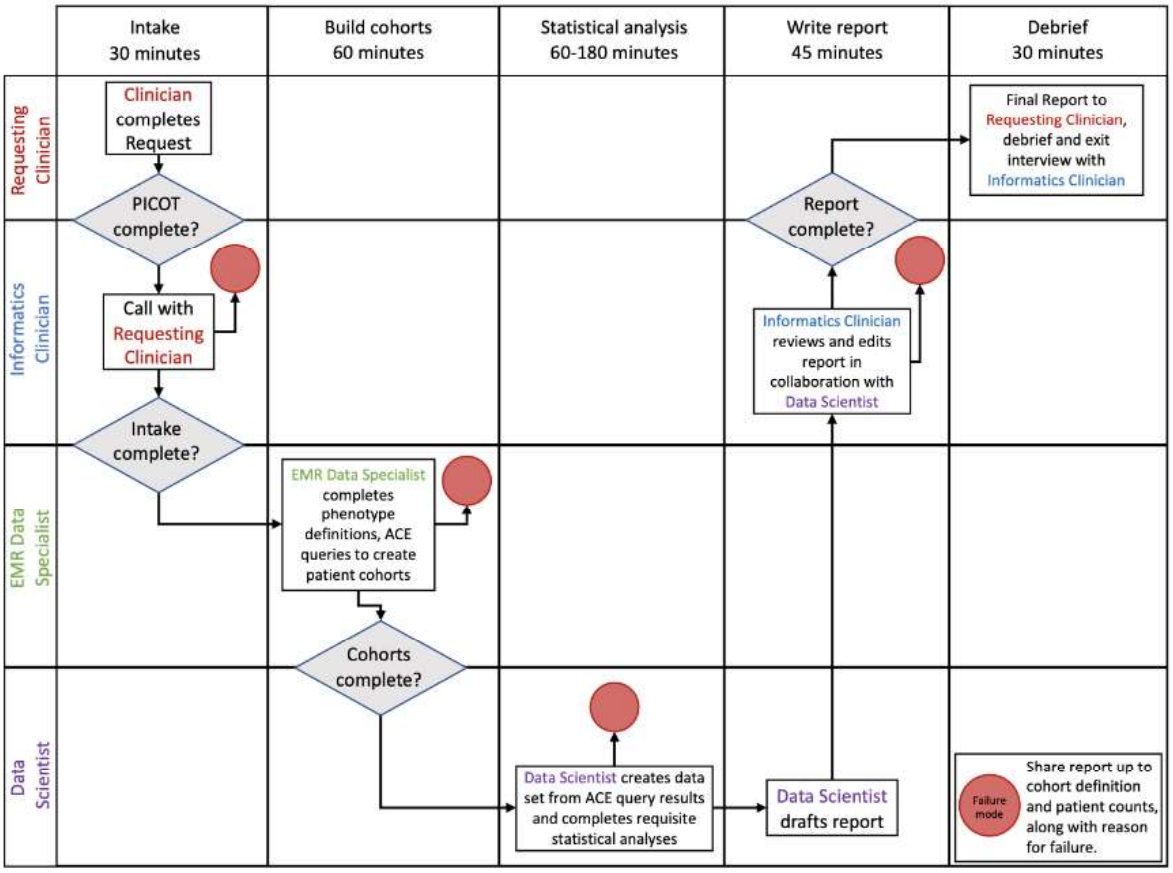



**Figure 13** : Circuit de traitement d'une demande de consultation par le système ACE, illustrant l'ordre de chaque étape, le temps nécessaire et le personnel responsable. Extrait de Callahan et al. [73]

D'autres applications ont été proposées dans des domaines spécifiques. C'est le cas notamment du prototype de Singh et al. [77] qui propose un système de consultation informatique piloté par des médecins pour améliorer la qualité des soins en orthopédie, illustré par le cas d'usage de la fracture humérale proximale. Les auteurs ont développé une interface pour les cliniciens pour définir les patients similaires ou des groupes de référence. Ils proposent ensuite une analyse visuelle des devenirs des patients en fonction des traitements entrepris. La cohorte de patients similaires est construite à partir de plusieurs filtres s'appuyant à la fois sur des données structurées pour les critères démographiques (âge, sexe) et les actes chirurgicaux (code procédure) et également sur les données non structurées, notamment pour l'extraction des caractéristiques de la fracture.

Néanmoins cette proposition reste à l'état de prototype sur des données synthétiques de cohorte de patients et l'étape du traitement du langage n'a pas encore été développée.

Les deux systèmes évoqués ci-dessus ne proposent pas directement de mesure de similarité puisque le moteur de recherche retourne tous les patients de la base de données répondant aux critères de recherche.

### 2.6.2. Systèmes basés sur des mesures de similarité :

Contrairement aux systèmes basés sur des requêtes manuelles de caractérisation de patients, les systèmes basés sur des mesures de similarité à partir d'un cas index épargnent la tâche de définition des caractéristiques du patient recherché, tâche d'autant plus fastidieuse que la recherche est fine. C'est en particulier vrai lorsque l'on souhaite utiliser les notes cliniques et qu'il faut définir par exemple une liste de mots clés et de leurs synonymes potentiels.

Certaines solutions développées proposent une mesure de similarité fixe. Dans le domaine du cancer, Lamy et al. [2] ont notamment proposé une distance multidimensionnelle en fonction du type de caractéristiques comparées : pour les variables numériques, une distance euclidienne ; pour les variables booléennes (0 ou 1) la simple différence ; et pour les variables catégorielles une distance sémantique basée sur une ontologie (ici les codes CIM-9 et CIM-10, deux codes ayant des concepts parents commun étant proche). Ensuite, la construction du groupe de patients similaires, ou cluster, est réalisée par un algorithme du type 'des plus proches voisins'.

L'apport important de cet article, dans le contexte de la construction de cohortes de patients similaires pour l'aide à la décision thérapeutique, est l'effort spécifique des outils de visualisation des résultats de cohorte. Les auteurs proposent en effet un rendu à la fois quantitatif par un diagramme de dispersion pour refléter la distance entre les patients similaires et le patient index et des diagrammes colorimétriques pour refléter les caractéristiques communes. Néanmoins, cette étude s'appuie essentiellement sur des données structurées et ne détaille pas la méthode d'extraction des candidats potentiels en amont de l'algorithme.

Enfin d'autres auteurs ont également proposé des mesures de similarité apprises. C'est le cas notamment de Suo et al. [3] qui propose un apprentissage couplé du dossier patient informatisé et d'une similarité par paire, la similarité de deux patients étant définie comme deux patients ayant la même maladie, i.e. le même code diagnostique. Le dossier patient informatisé est modélisé par une liste de codes diagnostiques pour chaque visite hospitalière ; cette matrice (ou tableau) liste de codes / numéro de visite est ensuite fournie à un réseau de neurones de type convolutif, fournissant un

premier encodage de données longitudinales. Plusieurs fonctions de similarité sont ensuite testées, avec les meilleures résultats obtenus par une fonction de classification de type *softmax* qui classe les paires similaires et dissimilaires. Cette méthode bien qu'efficace pour classer les patients ayant des pathologies fréquentes est néanmoins moins adaptée pour des maladies rares où les exemples sont plus rares pour l'apprentissage d'une part et avec une plus grande variabilité de codes.

Notre méthode bout-en-bout présentée section 4 est essentiellement centrée sur une mesure de similarité.

## 2.7. Enjeux éthiques

Il est central, dans le contexte de ce travail, de rappeler les enjeux éthiques et de s'assurer notamment du respect des quatre grands principes de la déontologie médicale : la bienfaisance, la non-maltraitance, l'autonomie de la personne et la justice.

Dans un communiqué récent, le conseil de l'ordre des médecins [78] met en avant les opportunités qu'apporte l'intelligence artificielle, en rappelant que "aujourd'hui, et plus encore demain, les personnes vivent et vivront dans une société numérique qui peut puissamment répondre au moins en partie à ces besoins". Le Conseil souligne toutefois le besoin d'encadrement de ces activités, en proposant, d'une part (i)une loi « techno-éthique », à l'instar de la loi de bioéthique qui encadre certaines activités médicales et de recherche pour s'assurer du respect de la personne humaine et (ii) d'autre part, que les futurs médecins soient formés à la fois aux outils de l'intelligence artificielle et tout autant aux sciences humaines et sociales.

Dans le même sens, le Comité Consultatif National d'Ethique pour les Sciences de la Vie et de la Santé et le Comité national pilote d'éthique du numérique ont rendu un avis commun publié en janvier 2023 sur la question du diagnostic médical [79], qui évoque les « promesses » et « avantages » des outils d'intelligence artificielle mais appelle à « se donner constamment les moyens de prendre de la distance avec le résultat fourni et [il est indispensable] de créer les conditions de la confiance."[79]

Parmi les nombreux enjeux éthiques soulevés, on peut brièvement mentionner ici la question des biais ; celle de la responsabilité ; celle de l'articulation avec la relation médecin-patient.

Pour être capable de critiquer au mieux les algorithmes d'apprentissage profonds, il est important de considérer les différentes sources de biais, c'est-à-dire les distorsions ou erreurs systématiques qui les caractérisent. Hovy et al. [80] ont proposé cinq sources de biais pour le traitement automatisé du langage par apprentissage profond : le biais lié aux données, le biais lié à l'annotation, le biais lié à la représentation en entrée (c'est-à-dire l'encodage), le biais lié au modèle et le biais lié à la conception

de la recherche.

Le biais lié aux données ou biais de sélection, est induit par les déséquilibres de distribution des différentes classes présents dans les jeux de données : cela peut-être un déséquilibre socio-démographique avec une sous-représentation de certaines minorités dans les textes par exemple. Ces jeux déséquilibrés induisent des performances bien moins bonnes pour les minorités sous-représentées, entraînant des propos possiblement sexistes ou racistes dans les suites pour l'utilisation du modèle [80]. Ce biais peut être particulièrement sensible pour la prédiction de maladies rares. De la même façon, l'annotation peut entraîner des biais, en particulier lorsqu'elle est réalisée par peu de personnes.

Est également particulièrement pertinent pour notre projet le biais de la représentation d'entrée ou biais sémantique. En effet, dans le contexte de représentation vectorielle des mots en fonction du contexte, si des préjugés sont présents dans le texte, ils influencent directement le modèle. Par exemple, si dans le corpus d'entraînement les femmes sont statistiquement plus souvent des infirmières et les hommes statistiquement plus souvent des médecins, le modèle, par construction, associera plus les femmes au métier d'infirmière et les hommes au métier de médecin.

Ces biais peuvent conduire à des résultats erronés lorsque le modèle est utilisé dans un autre contexte : il est donc absolument nécessaire de les connaître et de les considérer. Un des exemples marquant d'utilisation en pratique clinique est l'algorithme proposé par Google pour faire de la détection automatique de cancer du sein sur les mammographies. [81] précise en effet que, d'une part, les performances de l'algorithme étaient particulièrement bonnes parce que celui-ci avait été testé sur une base avec de nombreux cancers et que ,d'autre part, il était entraîné sur des mammographies de très bonne qualité : son déploiement en routine clinique a été par la suite rapidement interrompu du fait d'un important surdiagnostic de cancer du sein chez les femmes jeunes [81]. Il est donc indispensable, comme le rappelle le comité national d'éthique, pour tout algorithme déployé, de connaître, pour les cliniciens, non seulement les performances mais surtout la distribution des jeux de données sur lesquels le modèle a été entraîné [79].

Cet exemple soulève également la question de la responsabilité juridique en cas d'erreur diagnostique ou thérapeutique suite à l'utilisation d'algorithmes d'apprentissage profond, question qui n'a pas encore été tranchée.

Enfin, l'impact des prédictions par algorithme d'aide à la décision sur la relation médecin-patient n'a pas non plus été étudié comme le rappelle Virginie Im et al. [Virginie Im et al.]. Le médecin est par exemple en devoir d'information du patient sur le recours aux algorithmes d'intelligence artificielle. En outre, comme le relève le comité consultatif national d'éthique, ces algorithmes ne doivent pas être considérés comme des solutions de substitution des équipes médicales et paramédicales [79].

L'importance de ces enjeux témoigne de la pertinence, pour comprendre les outils d'intelligence

artificielle, de cette citation de Paul Valéry: “L'homme sait assez souvent ce qu’il fait, il ne sait jamais ce que fait ce qu’il fait”.

En conclusion, il existe aujourd’hui des modèles de langages, entraînés sur de volumineuses bases de données non annotées, et adaptables pour un grand nombre de tâches avec de très bonnes performances. C’est notamment le cas de l’extraction d’entités nommées réalisée par des modèles de langage masqués de type Transformers et de la génération de texte réalisée par des modèles tel que GPT-3 [5].

De très nombreuses solutions de traitement des données et de représentation de dossiers patients informatisés ont été proposées. Néanmoins, leur déploiement et leur utilisation en pratique clinique, bien que prometteuse, reste faible [75]. Plusieurs freins expliquent cette absence de déploiement : le manque de généralisation et donc de robustesse des modèles [39], leur manque d’interprétabilité [82] et également le manque d’interaction entre les développeurs de tels outils et les cliniciens.

# 3. Données de l'étude

Avant de décrire de manière plus détaillée les différentes étapes de l'algorithme, il est nécessaire de détailler le jeu de données sur lequel le prototype a été développé et sa provenance.

## 3.1. L'Entrepôt de Données de Santé de l'AP-HP

L'entrepôt de données de santé (EDS) de Assistance Publique-Hôpitaux de Paris (AP-HP) est un projet débuté en 2017, qui regroupe les informations relatives à tous les patients suivis dans les 38 hôpitaux universitaires de la région parisienne (plus de 22 000 lits, 1,5 million d'hospitalisations par an) qui utilisent un logiciel commun de dossier médical électronique (DME), ORBIS Dedalus Healthcare. L'objectif principal de l'EDS est de restituer les données médicales aux médecins chercheurs du CHU (Centre Hospitalo-Universitaire) qu'est l'AP-HP. D'autres objectifs sont visés, par exemple pour le pilotage (remonté des indicateurs de qualité etc..) ou pour l'information des patients, via l'alimentation du portail patient.

Les données collectées dans la base de l'entrepôt de données de santé (EDS) proviennent de plusieurs sources :

- les données démographiques, les données de compte-rendus médicaux (d'imagerie, de consultation, d'hospitalisation etc..) et de prescription hors chimiothérapie (à partir de 2018 seulement) proviennent par exemple directement des tables Orbis et sont intégrées tous les jours ;

- les données de biologie sont issues du logiciel GLIMS ;

- les données médico-administratives de remboursement PMSI (Programme de Médicalisation des Systèmes d'Information), disponibles uniquement pour les séjours hospitaliers - codage CIM10 des pathologies, code des actes remboursés par la sécurité sociale, code des séjours : GHM (groupe homogène de maladies) - sont issues à la fois des informations d'Orbis et de la base AREM (une base d'archive de l'historique des remontées mensuelles et des reprises PMSI, qui transmet les données aux tutelles) ;

- les données d'imagerie sont fournies par le logiciel PACS ;

- les données de décès sont fournies dans la base directement lorsque le patient décède à l'hôpital et tous les 3 à 6 mois à partir des données de l'INSEE pour les décès hors AP-HP.

La base de données de recherche suit le standard Informatics for Integrating Biology & the Bedside [83] et est en cours de migration vers un format standardisé OMOP-like développé par l' "Observational Health Data Sciences and Informatics" (OHDSI) [84].

Cette base de données regroupe les informations de 11 millions de patients, 120 millions de documents, 1.2 milliards de lignes de biologie. Ces contraintes en volume, le déploiement d'Orbis à des dates différentes dans les 38 hôpitaux, ainsi que l'apparition plus tardive du logiciel de prescription et le caractère multimodal de la donnée conduisent à une architecture à travers laquelle le transport de la donnée est complexe. Cette complexité est mise en évidence par la cartographie ci-dessous (figure 14).

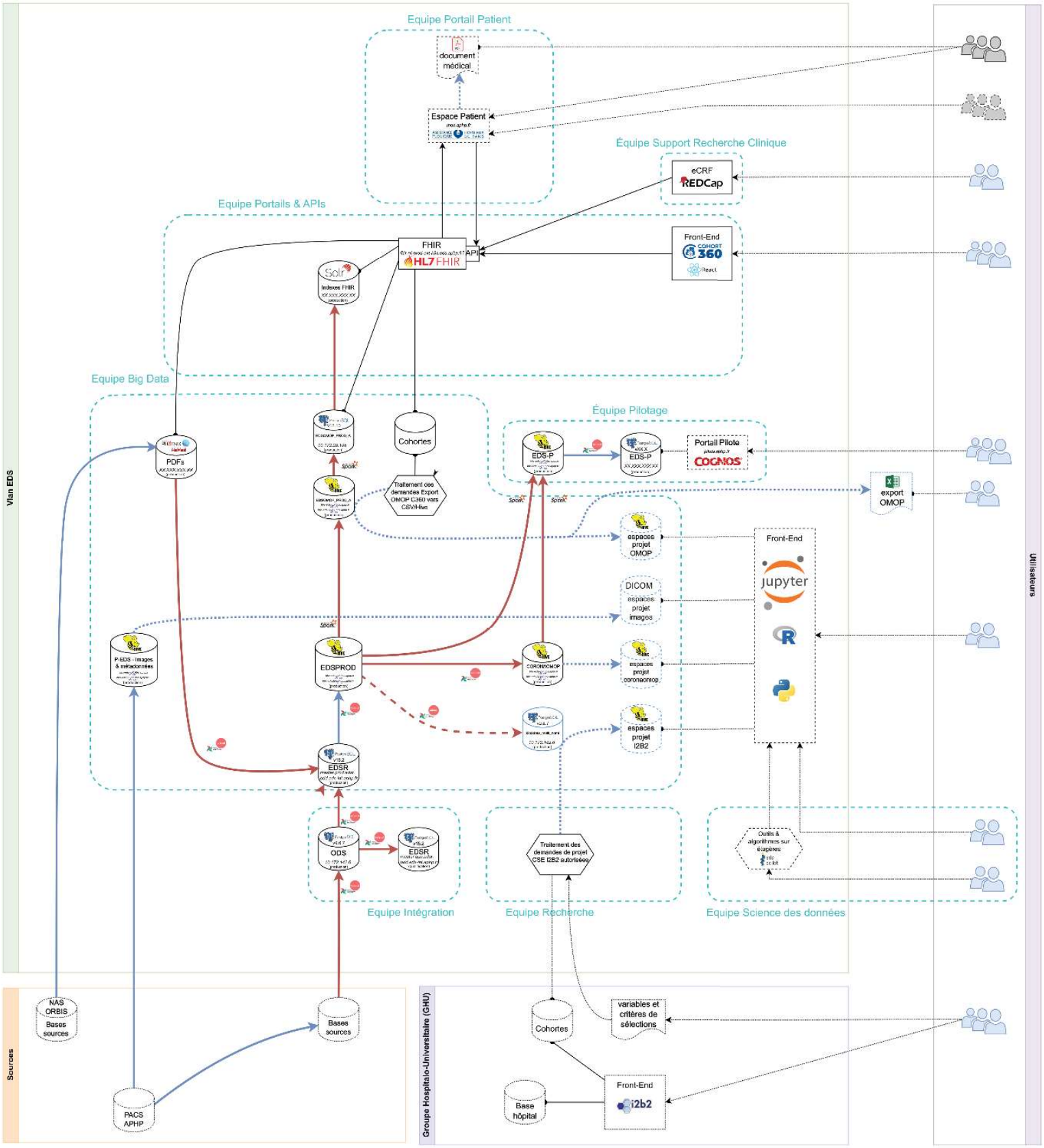



**Figure 14 :** Cartographie du transport de la donnée de l'EDS, crédit : Edouard Mahieu, fournie avec l'accord des équipes de l 'EDS.

## 3.2. Jeu de données de l'étude

### 3.2.1. Description réglementaire et critères d'inclusion

Le protocole de recherche de ce projet a été accepté en 2020 par le comité scientifique et éthique de l'EDS (décision n° 20-93). Ce jeu de données est une extraction de la base de données de recherche de l'EDS, au format i2b2 (Integrating Biology & the Bedside [83]).

Les critères d'inclusion pour l'étude sont les suivant : tous les patients âgés de plus de 15 ans atteints de lupus érythémateux disséminé, de sclérodermie systémique, de syndrome des antiphospholipides et de maladie de Takayasu ayant eu au moins une visite dans les hôpitaux de l'AP-HP initialement sur la période du 01/07/2017 au 31/12/2020. Le ciblage des patients dans la base des données était double : à partir des codes CIM 10 de ces quatre pathologies et à partir de mots-clés présents dans les comptes-rendus médicaux comme le synthétise le tableau 1.

**Tableau 1**: Ensemble des requêtes réalisées pour extraire les données du projet.

| **Pathologies** | **Codes CIM10** | **mots clés dans les notes cliniques** | **Nombre de patients correspondants** |
|---|---|---|---|
| **Lupus érythémateux disséminé** | M310, M321, M328, M329, L930, L931 (les codes CIM10 incluent les lupus cutanés : lupus cutané subaigu et lupus chronique) | "lupus" | 22 252 |
| **maladie de Takayasu** | M314 | "Takayasu" | 576 |
| **Sclérodermie systémique** | M340, M341, M348, M349 | "sclérodermie", "CREST" | 7 711 |
| **Syndrome des anti-phospholipides** | D686 | "SAPL", "syndrome des anti-phospholipides", "CAPS" | 16 401 |

Pour ces patients, les données demandées étaient :

- Les données de démographie (âge, sexe etc..) correspondant à la table “i2b2_patient” dans i2b2
- Les données des passages des patients dans les différents services (issues des tables i2b2_visit et i2b2_observation_ufr)
- Les données textuelles (issues de la table “i2b2_observation_doc”), incluant tous les comptes-rendus médicaux en texte brut (comptes-rendus de consultation, d’hospitalisation, d’examen d’actes interventionnelles, opératoire, anatomo-pathologique, d’imagerie, ordonnances, etc..). Ces textes bruts sont directement extraits des PdF issus du logiciel source Orbis
- Les données de biologie et de microbiologie
- Les données médico-administratives de codage CCAM (Classification commune des actes médicaux), GHM (Groupe homogène de maladie) et CIM10 (Classification internationale des maladies).

L’extraction portait sur l’ensemble des services médicaux pouvant potentiellement prendre en charge les patients des quatre pathologies d’intérêt : les services de médecine interne et immunologie clinique, de néphrologie, de rhumatologie, de dermatologie, de pneumologie, de neurologie, de gastro-entérologie, d’oncologie, d’hématologie, de maladies infectieuses, d’urgence et de réanimation.

Les tables i2b2 sont disponibles via une interface Jupyter sécurisée dont l’environnement de travail permet d’utiliser les langages de programmation Python 3 ou R. L’ensemble du travail présenté dans cette thèse a été programmé en Python. Un accès à de puissants processeurs graphiques (ou GPU en anglais) de type T4 et V100 est également possible dans cette interface.

Dans le but de poursuivre la création du prototype dont les principales étapes sont présentées dans ce manuscrit, un amendement au Comité scientifique et éthique CSE) de l'EDS, visant à demander une mise à jour régulière de la base de données (tous les trois mois) pour s'adapter à l'évolution des molécules thérapeutiques, et avoir accès à un groupe de patients témoins, a été évalué et accepté en octobre 2022.

### 3.2.2. Description brève des quatre pathologies auto-immunes

Le paragraphe ci-dessous résume les définitions, les données épidémiologiques et de diagnostic des quatre pathologies auto-immunes choisies pour le projet de recherche.

#### 3.2.2.1. Le lupus systémique

La définition du protocole national de diagnostic et de soins (PNDS) du lupus [85] du lupus systémique (LS) ou lupus érythémateux disséminé (LED) est “une maladie systémique protéiforme et spontanément grave caractérisée par la production d’anticorps antinucléaires dirigés en particulier contre l’ADN natif. Le LS s’associe parfois au syndrome des anticorps antiphospholipides (SAPL) défini par l’association de thromboses et/ou d’événements obstétricaux et d’anticorps antiphospholipides (aPL). Le LS pédiatrique est défini par un diagnostic posé avant l’âge de 16 ans.”

Sur le plan épidémiologique, “la prévalence du LS était en 2010 de 41/100 000 en France et l’incidence de 3 à 4 nouveaux cas pour 100 000 par an” et “chez l’adulte, le LS survient 90 fois sur 100 chez la femme, généralement en période d’activité ovarienne (pic de prévalence entre 30 et 39 ans)”[85].

Les critères diagnostiques pour cette pathologie ont été révisés en 2019 par l'ACR (American College of Rheumatology) et l’EULAR (European League Against Rheumatism) [86] et sont les suivants :

- un critère obligatoire : avoir eu au moins une fois des anticorps antinucléaires positifs
- des critères cliniques et paracliniques pondérés :
    - fièvre ( > 38.3° Celsius)
    - leucopénie (< 4 000/$mm^3$)
    - thombopénie (< 100 000/$mm^3$ )
    - anémie auto-immune
    - atteintes psychiatriques tel que le délire aigu ou la psychose
    - crises épileptiques généralisées ou partielles
    - alopécie non cicatricielle
    - ulcérations buccales
    - lupus cutané subaigu ou discoïde
    - lupus cutané aigu : rash malaire, ou rash maculo-papuleux généralisé
    - épanchement péricardique ou pleurale

- péricardite aiguë (avec au moins deux parmis : 1. douleur thoracique typique 2. frottement péricardique à l'auscultation 3. signes ECG spécifiques 4. épanchement péricardique à l'imagerie)
- atteinte articulaire : synovite, gonflement articulaire, raideur matinale
- protéinurie > 0.5g/24h
- néphropathie lupique classe II, III, IV ou V à la biopsie rénale
- positivité des anticorps anti-phospholipides
- Complément C3 ou C4 abaissé
- positivité des anticorps anti ADN double brin (ADN natif) ou anti Sm

Le diagnostic de lupus systémique est établi lorsque la somme des pondérations associée à chacun de ces critères est supérieure à 10.

#### 3.2.2.2. Le syndrome des antiphospholipides (SAPL)

Le site de la SNFMI (société française médecine interne) fournit la définition suivante pour le syndrome des antiphospholipides (SAPL): "maladie auto-immune, caractérisée par la survenue de manifestations thromboemboliques (formation de caillots de sang dans les vaisseaux, veines ou artères) et/ou la survenue de complications de la grossesse aussi appelées complications obstétricales (il s'agit de fausses couches répétées et/ou de complications plus tardives de la grossesse), et la présence, au moins à deux reprises, à trois mois d'intervalle, d'anticorps appelés anticorps antiphospholipides."[87]

Au plan épidémiologique, le SAPL primaire (ou isolé) "est très difficile à évaluer, et serait de l'ordre de 0,5% de la population générale. Lorsqu'un(e) patient(e) souffre d'un lupus systémique, il/elle a un risque de 20 à 40 % d'avoir un SAPL associé. Comme pour le lupus, ce sont les femmes qui sont le plus souvent concernées par cette maladie (environ 5 fois plus souvent que les hommes)". [87]

#### 3.2.2.3. La sclérodermie systémique

La sclérodermie systémique est définie comme "une maladie rare au cours de laquelle peuvent survenir des manifestations viscérales, en particulier vasculaires périphériques, digestives, cardio-pulmonaires et rénales. Elle se caractérise par des anomalies de la microcirculation (plus rarement de la macrocirculation mais cet aspect est encore débattu) et des lésions de fibrose cutanée et/ou viscérale. Les lésions de sclérose cutanée peuvent être modestes, distales (doigts notamment) et au niveau du pourtour buccal ou étendues au-dessus des coudes et des genoux, ou plus rarement

atteindre le tronc, l'abdomen. Il existe des formes rares, sans atteinte cutanée, appelées « sine scleroderma».”[88]

Au plan épidémiologique, “La ScS touche avec prédilection les femmes (3 à 8 femmes pour 1 homme). Il existe un pic de fréquence entre 45 et 64 ans. La prévalence de la ScS est encore mal connue. En France, la prévalence de la sclérodermie systémique [...] est évaluée [...] entre 6000 et 9000 patients adultes.” [88].

Les critères diagnostics pour cette pathologie sont également fournis par la classification ACR/EULAR et sont présentés au tableau 2.

**Tableau 2**: Critères de classification ACR/EULAR de la sclérodermie systémique. Le critère est retenu s'il est présent au moins une fois dans l'histoire clinique du patient. Un score de 9 ou plus permet de poser le diagnostic. Abbréviations : ScS : Sclérodermie Systémique, MCP : Métacarpo-Phalangienne (articulation de la main), HTAP : hypertension artérielle pulmonaire.

| Domaine | Critères * | Score # |
|---|---|---|
| Épaississement cutané (ne tenir compte que du score le plus élevé) | Épaississement cutané des doigts des mains s'étendant au-delà des articulations MCP | 9 |
| | Doigts boudinés | 2 |
| | Atteinte des doigts ne dépassant pas les articulations MCP | 4 |
| Lésions pulpaires (ne tenir compte que du score le plus élevé) | Ulcères pulpaires digitaux | 2 |
| | Cicatrices déprimées | 3 |
| Télangiectasies | | 2 |
| Anomalies capillaroscopiques | | 2 |
| Atteinte pulmonaire | HTAP et/ou fibrose pulmonaire | 2 |
| Phénomène de Raynaud | | 3 |
| Anticorps spécifiques de la ScS | Anti-topoisomérase I | 3 |
| | Anticorps anticentromères | |
| | Anti-ARN polymérase de type III | |

#### 3.2.2.4. La maladie de Takayasu

La définition de la maladie de Takayasu fournie par le PNDS national en 2019 [89] est la suivante : “ L'artérite de Takayasu est une vascularite affectant les vaisseaux de gros calibre, notamment l'aorte et ses branches principales (artères sous-clavières, carotides, vertébrales, rénales, digestives, iliaques), mais aussi les coronaires et les artères pulmonaires. L'atteinte peut être segmentaire ou diffuse à l'ensemble de l'aorte thoracique et abdominale et ses branches.”

Concernant l'épidémiologie, "l'incidence annuelle est estimée entre 2 et 3 cas par million d'habitants et en France entre 1,2 à 2,6 cas/million/an". Par ailleurs, " La femme est majoritairement atteinte, avec un ratio femme/homme variable de 8/1 au Japon, 5,9/1 au Mexique, 4,8/1 en France, 1,2/1 en Inde"[89].

Le diagnostic pour cette pathologie est porté par la présence d'une "imagerie caractéristique de vascularite des artères de gros calibres" chez un patient de moins de 50 ans, en l'absence d'argument pour une autre cause vasculaire. Les signes cliniques: claudication d'un membre supérieur (douleur ou faiblesse survenant à l'effort et forçant son arrêt), abolition d'un pouls d'un membre supérieur, asymétrie tensionnelle, hypertension artérielle réno-vasculaire, souffle cervical, etc… "peuvent renforcer le diagnostic, mais peuvent être absents" [89].

Ces quatre pathologies ont été choisies pour leur rareté et la diversité des manifestations cliniques et paracliniques qu'elles représentent. Elles constituent des pathologies tests pour faire une preuve de concept de l'utilité d'un tel outil.

### 3.2.3. Description de la population

Concernant la population extraite pour le projet, les femmes représentent 68.3% des patients - ce qui est attendu étant donné la prévalence féminine principale des quatre pathologies d'intérêt étudiées (cf paragraphe précédent). La figure 15 met en évidence la répartition des sexes et des âges dans la population de l'étude.

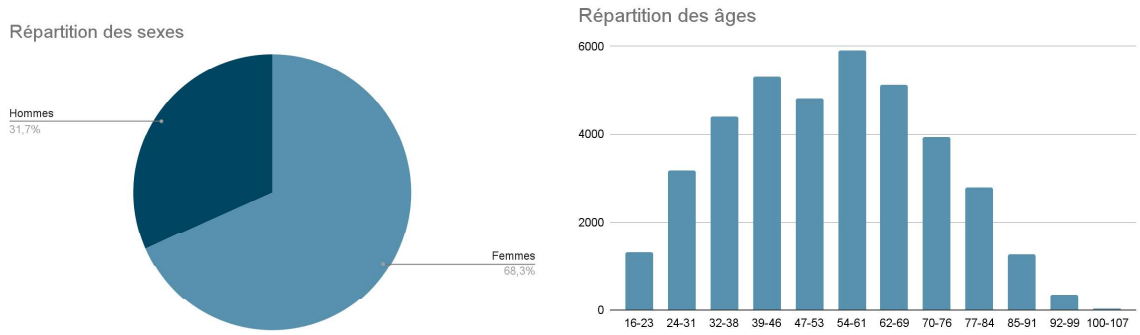



**Figure 15** : Répartition des sexes et des âges dans l'étude. L'âge est évalué à l'échelle de la visite du patient, l'âge médian est de 54 ans.

Par ailleurs, la répartition des GHM (Groupe Homogène de Maladies) pour les séjours était la suivante (figure 16).

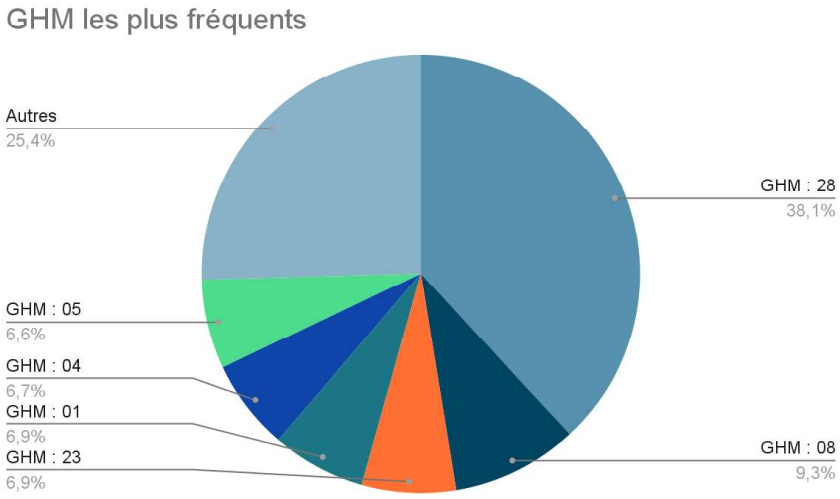



**Figure 16** : Répartition des codages GHM des séjours.

| | |
|---|---|
| GHM : 28 | Séances (dialyse, chimiothérapie tumorale, transfusion, radiothérapie etc..) |
| GHM : 08 | Affections et traumatismes de l'appareil musculosquelettique et du tissu conjonctif |
| GHM : 23 | Facteurs influant sur l'état de santé et autres motifs de recours aux services de santé (rééducation, chimiothérapie non tumorale, exploration etc..) |
| GHM : 01 | Affections du système nerveux |
| GHM : 04 | Affections de l'appareil respiratoire |
| GHM : 05 | Affections de l'appareil circulatoire |

Enfin, également à titre descriptif, la répartition des documents dans l'espace-projet est présentée figure 17.

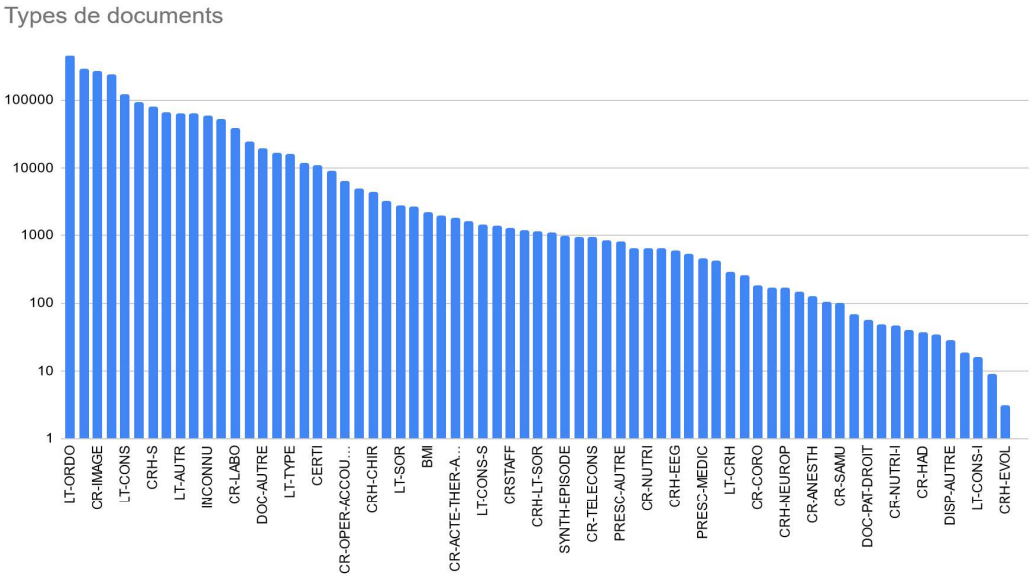



**Figure 17**: Répartition des types de documents, le volume total des documents dans l'espace étant de 2 millions de documents textuels. L'échelle présentée ici est logarithmique. Il y a donc plus de 100 000 documents d'ordonnance "LT-ORDO", de comptes-rendus d'imagerie "CR-IMAGE", et de lettres de consultations "LT-CONS", par exemple.

# 4. Les différentes étapes de l'algorithme de construction de cohorte de patients similaires

L'état de l'art et le contexte de l'infrastructure et du jeu de données ayant été présentés, il s'agit maintenant de décrire les étapes de l'algorithme d'extraction de patients similaires.

L'ordre de description proposé est celui du parcours de la donnée depuis le traitement automatique des documents PDF en amont de l'algorithme (partie A, article 1 soumis), suivi d'une partie détaillant l'algorithme de cohorte de patients similaires (partie B), une partie présentant la pré-sélection de patients (partie C, article 2 en révision), une partie présentation l'extraction des concepts médicaux en comparant les performances de deux méthodes, l'une à partir de modèle français, l'autre à partir de modèle anglais avec une étape de traduction (partie D, article 3 en révision), une partie développant la classification de ces concepts extraits (partie E, article 4 publié), une partie sur la construction des cohortes de patients similaires, testée sur quatre phénotypes d'intérêt (partie F, article 5 publié) et enfin une dernière partie sur les avancées des travaux actuels, prenant en compte le traitement des données de biologie et de médicaments issues du texte.

Conformément aux recommandations de rédaction de l'école doctorale, ces articles ne sont pas traduits, mais un résumé en français est proposé.

Dans tous ces travaux, j'ai été responsable de la conception des expériences, des analyses, des jeux d'annotation pour les étapes d'extraction et de normalisation des entités nommées (en utilisant les codes des concepts de l'UMLS) et de la rédaction de chacun des articles. Plus spécifiquement, en ce qui concerne la programmation, j'ai pu affiner le modèle camemBERT-large [24] sur l'espace de mon projet (2 millions de comptes-rendu), programmer le classifieur multi-label pour classer les pathologies en grands domaines médicaux, tester un grand nombre d'hyperparamètres et d'architectures de modèles pour l'extraction d'entités nommées et la normalisation [45,56,60]. Pour l'expérience de présélection de patients similaires (article 2, section 4.3.) et l'expérience de comparaison de performance entre les modèles français et anglais (article 3, section 4.4), j'ai co-supervisé une étudiante de Master 2 avec le Professeur Xavier Tannier pour toutes les étapes de programmation, d'analyse d'erreurs, etc. Pour le travail plus récent sur l'extraction d'informations sur la biologie et les médicaments, j'ai également co-encadré deux étudiants de Master 2 pour la programmation (avec l'équipe Data Science de l'AP-HP), et réalisé l'analyse, toutes les étapes

d'annotation pour l'extraction d'entités, des sections et la normalisation sur les codes UMLS et ATC, etc.....

Plus généralement, il me semble que la principale contribution de mon travail réside dans l'intégration des connaissances médicales et surtout du raisonnement clinique dans les différentes étapes de l'algorithme. Il s'agit notamment de pouvoir extraire de manière différenciée les signes physiologiques et pathologiques à partir de modèles de langage, dont j'ai rapidement compris qu'ils pouvaient prendre une dimension sémantique. De même, la classification de chaque symptôme et pathologie en spécialités médicales s'apparente au raisonnement clinique que nous adoptons dans la prise en charge diagnostique et thérapeutique des patients. Ces choix originaux ont été guidés par mon expérience clinique et d'ingénieur, mais aussi par les cours sur l'éthique de l'intelligence artificielle appliquée à la médecine que j'ai pu donner à des collègues médecins. Ces cours ont confirmé mon intuition du besoin d'interprétabilité, non pas des modèles, mais surtout des résultats fournis, notamment dans le cadre de l'aide à la décision diagnostique et thérapeutique.

## 4.1. Extraction des textes cliniques à partir des documents PdF (article 1 *soumis*)

Le travail sur les textes cliniques dès le début de ce projet a mis en évidence des défauts importants dans le rendu des textes cliniques bruts mis à disposition par l'EDS jusqu'alors dans les espaces de recherches. Ces défauts proviennent essentiellement du fait que les textes intégrés au sein de l'entrepôt de données sont tous issus de documents au format PDF auxquels un masque simple est appliqué pour pouvoir obtenir le texte brut. Cette étape est indispensable pour des raisons de pseudonymisation et également de mémoire de stockage. Malheureusement ce masque simple génère parfois la fusion du texte administratif (issu du papier à en-tête des services hospitaliers, information du service d'hospitalisation du patient, note de bas de page etc…) avec le texte clinique. Un autre défaut engendré par ce masque réside dans la suppression de l'architecture du texte avec une fusion des titres des sections au début du texte, comme présenté figure 18. Cette perte de structure est particulièrement préjudiciable dans l'analyse automatique du texte (par exemple pour délimiter l'histoire de la maladie - antérieure avec l'évolution clinique du patient au cours de l'hospitalisation, ou pour identifier les traitements à l'entrée par rapport aux traitements à la sortie, etc…).


**Figure 18**: Texte brut fourni par le masque simple à partir des PDF, avec fusion des titres de section au milieu du texte et note de pied de page “imprimé le…” dans le texte clinique.

Ce constat a été à l’origine d’un travail en collaboration avec l’équipe de Data Science de l’EDS et a fait l’objet d’une soumission en cours, présentée ici. L’algorithme présenté ici est celui qui est à présent déployé en production à l’EDS.

Résumé de l’article :

Cette étude présente un travail original de traitement des PdF des documents médicaux pour, d’une part extraire le texte clinique d’intérêt sans être parasité par toute la partie administrative (issue du papier à en-tête des services) et, d’autre part, préserver l’architecture du texte en maintenant l’ordre des sections avec les paragraphes respectifs. Une contrainte majeure de l’algorithme est qu’il doit être applicable sur un très grand nombre de documents (plus de 120 millions), et doit donc être particulièrement léger.

L’algorithme de détection automatique du texte clinique à partir des PDF des documents médicaux est un algorithme d’apprentissage profond, entraîné de façon supervisée sur un échantillon de 272 documents dans l’EDS, comprenant principalement, mais de manière non exclusive, des compte-rendus de consultation, d’hospitalisation, comptes-rendus des urgences et des ordonnances. L’annotation consistait à délimiter manuellement sur les PDF les notes de gauche, de droite, en-tête et pied de page, le titre et le corps du texte. Les étapes de l'algorithme sont les suivantes : 1. un analyseur PDF qui découpe le texte en ligne, 2. un classifieur des lignes en différents types : titre,

corps du texte, note de gauche etc…, 3. une agrégation des lignes par type. L'architecture du classifieur est basée sur un transformer non pré-entraîné avec un encodage à la fois de l'information de position de la ligne et de l'information du contenu textuel.
L'évaluation des performances est réalisée par ligne de texte bien classé ou non avec les métriques de précision, rappel, F1 score.

Par ailleurs, au-delà de la validation sur ce jeu d'entraînement, nous avons également évalué l'impact de ce nouveau traitement du PDF sur une tâche de reconnaissance d'entité nommée sur le texte brut en sortie. L'intérêt de l'extraction de texte "avancée" proposée étant de préserver l'architecture et notamment l'ordre des sections et des paragraphes dans le texte, l'évaluation de l'extraction d'entité nommée a été mesurée par section (les sections ayant été considérées comme des attributs pour ce traitement spécifique donc typiquement un médicaments apparaissant pour dans la section "traitement à l'entrée" aura pour attribut "traitement_entrée" etc…). Les sections étaient par ailleurs extraites à partir de règles.

Enfin l'algorithme bout-en-bout a été testé pour un cas d'usage précis de détection automatique d'infection aiguë dans les comptes rendus (infection mentionnée dans la partie "Evolution" du compte-rendu).

Pour les deux dernières évaluations, extraction d'entités et classification du document en infection aiguë, la comparaison est réalisée entre l'ancien masque, masque simple, et l'extraction avancée du texte clinique par le classifieur que nous avons développé. Les résultats montrent un F1 score à 0.84 [0.77; 0.90] pour la détection d'infection aiguë avec la nouvelle extraction versus 0.77 [0.70; 0.85] avec l'ancien masque. Les intervalles de confiance sont calculés par bootstrapping.

Nous démontrons que cette étape de détection du corps du texte améliore considérablement les performances des tâches d'extraction d'informations en aval. Comme preuve de concept, nous évaluons la performance d'un algorithme conçu pour identifier automatiquement les infections aiguës dans les documents cliniques, en utilisant un algorithme de reconnaissance d'entité nommée et une identification de section basée sur une classification d'entités à base de règles.

# Detecting automatically the layout of clinical documents to enhance the performances of downstream natural language processing

## Authors and affiliations:

**Christel Gérardin*, MD (ORCID: 0000-0002-9303-6349):** Sorbonne Université, Inserm, Institut Pierre-Louis d'Epidémiologie et de Santé Publique, Paris, France F7501, Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

**Perceval Wajsbürt*, PhD (ORCID: 0000-0002-9746-9993):** Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

**Basile Dura (ORCID: 0000-0002-8315-4050):** Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

**Alice Calliger (ORCID: 0000-0002-5767-313X):** Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

**Alexandre Mouchet (ORCID: 0009-0000-4360-542X):** Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

**Xavier Tannier**, PhD (ORCID: tbc.): LIMICS, Sorbonne Université

**Romain Bey, PhD (ORCID: 0000-0002-6413-5188):** Innovation and Data unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

***both authors contributed equally**

# Abstract

**Background:**

The use of clinically derived data in secondary use within health data warehouses for research or steering purposes can be complex, especially when analyzing textual documents from PDFs provided by the source software.

**Objective:**

Develop and validate an algorithm for analyzing the layout of PDF clinical documents to improve the performance of downstream natural language processing tasks.

**Materials and Methods:**

We designed an algorithm to process clinical PDF documents and extract only clinically relevant text. The algorithm consists of several steps: initial text extraction using a PDF parser, followed by classification into categories such as body text, left notes, and footers using a Transformer deep neural network architecture, and finally an aggregation step to compile the lines of a given label in the text. We evaluated the technical performance of the body text extraction algorithm by applying it to a random sample of documents that were annotated. Medical performance was evaluated by examining the extraction of medical concepts of interest from the text in their respective sections. Finally, we tested an end-to-end system on a medical use case of automatic detection of acute infection described in the hospital report.

**Results:**

Our algorithm achieved per-line precision, recall, and F1 score of 98.4, 97.0, and 97.7, respectively, for body line extraction. The precision, recall, and F1 score per document for the acute infection detection algorithm were 82.54 (95CI 72.86-91.60), 85.24 (95CI 76.61-93.70), 83.87 (95CI 76, 92-90.08) and 80.35 (95CI 70.08-89.75), 73.77 (95CI 62.71-83.61), 76.92 (95CI 69.09-84.8), with or without exploitation of the results of the advanced body extraction algorithm, respectively.

**Conclusion:**

We have developed and validated a system for extracting body text from clinical documents in PDF format by identifying their layout. We were able to demonstrate that this preprocessing allowed us to obtain better performances for a common downstream task, i.e., the extraction of medical concepts in their respective sections, thus proving the interest of this method on a clinical use case.

**Keywords: PDF extraction, Natural language processing, Phenotypes, Electronic Health Records, Machine Learning**

# 1. Introduction

## Background

In recent years, electronic health records (EHRs) stored in large clinical data warehouses have become widely available. Health databases, such as those of the Assistance Publique - Hôpitaux de Paris (AP-HP), have facilitated the secondary use of clinical notes for epidemiological research, pharmacovigilance, automatic detection of patient cohorts and the development of diagnostic or therapeutic prediction models. One of the challenges of these databases is to process a very large volume of documents: currently more than 120 million at the AP-HP. The automated analysis of these clinical notes has been made possible by natural language processing (NLP) algorithms, which are particularly adept at extracting named entities of interest - such as medications, symptoms, comorbidities and diagnostic procedures -, text classification, translation,etc.

However, NLP algorithms are often designed to be applied on plain text, but in many health databases, due to an imperfect interoperability of many clinical softwares, documents are primarily available only as PDFs whose layout depend on the clinical software from which they originate. Prior to the extraction of named entities of interest- or other NLP tasks, the plain text information is often derived from the direct capture of PDF documents using a simple mask, introducing noise and decreasing the performance of textual information extraction. This process often leads to the loss of the document structure, in particular regarding section layout. To address this issue, which is not specific to medical documents, several teams have proposed methods for joint analysis of document layout and corresponding text for enhanced comprehension.

Liu et al. [1] presented a graph convolution method for multimodal information extraction in visually rich documents, utilizing a combination of graph embeddings for layout encoding and text embeddings with a BiLSTM-CRF, outperforming models based solely on textual information. Xu et al [2, 3] and Huang et al. [4] introduced three successive methods - LayoutLM, LayoutLMv2, and LayoutLMv3 - for automatic document analysis that integrate both text and layout information into a single model. These templates have achieved state-of-the-art results, with LayoutLMv3 outperforming the others. LayoutLM is based on the BERT architecture, incorporating 2D layouts and embedded images, while the LayoutLMv2 and LayoutLMv3 versions use a multimodal Transformer architecture that incorporates text, layout and image information.

For all models, the initial extraction of text and layout is conducted using optical character recognition (OCR) or a PDF parser. Other methods have also been proposed [5, 6, 7].

One of the challenges of these information extractions is to respect the architecture of the text and therefore the titles of the sections, restored in the right order (usually corresponding to “reason of admission”, “medical history”, “usual treatment”, etc…). Several EHR text analysis can benefit from section identification: enabling a temporal relation extraction [8], abbreviation resolution [9], cohort retrieval [10]. Automatic or semi-automatic section identification in narrative clinical notes has been studied in the past, as shown by Pomares et al. in a recent review paper [11]. They define section identification as detecting the boundary of sections in the text. A section generally corresponds to a paragraph summarizing one dimension of the patient (medical history, allergies, physical exam, evolution during its hospital stay, usual treatments, discharge treatments, *etc.*). According to this review, the majority of the studied papers (59% for 39 studies analyzed) used rule-based methods, 22% machine learning methods and 19% both. Authors also highlight that very few studies presented results with a formal framework. Finally almost all the studies relied on a custom dictionary.

## Goal of the study

In this study, we propose an end-to-end algorithm that processes standard clinical documents in PDF format by extracting the body text separately from the left-hand notes, footnotes, signatures, and other elements, retaining only the clinically relevant content and preserving its original structure, thus correctly detecting the sections. The objective of the algorithm is to be applied to a very large number of documents (more than 120 million) and therefore must be very lightweight.
We demonstrate that this main body text detection step significantly enhances the performance of downstream information extraction tasks. As a proof of concept, we evaluate the performance of an algorithm designed to automatically identify acute infections in clinical documents, utilizing a named entity recognition algorithm and a section identification based on a rule-based entity classification.

# 2. Material and Methods

## 2.1. Datasets

For this study we had two datasets at our disposal described below:

- a first dataset of 272 annotated PDF documents for the development and technical validation of the text-extracting algorithm.

- a cohort of auto-immune diseases patients (with systemic lupus erythematosus, Takayasu disease, scleroderma and antiphospholipid syndrom) with 151 fully annotated document with medical concepts and 200 with or without a phenotype of interest, respectively for the medical validation and the illustrative use case.

This study was approved by the local institutional review board (IRB00011591, decision CSE22-18). Subjects that objected to the reuse of their data were excluded. For privacy protection reasons, researchers do not have direct access to the PDF documents. A text-extracting algorithm and a text-pseudonymization algorithm were consequently applied to all the documents before any delivery [Tannier2023].

1. Hospitals' clinical documents (development and technical validation)

The wide majority of clinical reports, including discharge summaries, imaging reports, emergency department reports, surgical reports, prescriptions, pathology reports, and more, are imported into the AP-HP clinical data warehouse as PDF files. A total of 272 randomly selected reports were sampled, ensuring a 3:1 weighting in favor of the following note types: consultation reports, hospitalization reports, operative reports, pathology reports, imaging reports, discharge letters, procedure reports, prescriptions, and emergency department visit reports. These reports were annotated by three independent annotators who segmented the PDFs by marking boxes of interest. These boxes encompassed body text, footers, headers, left notes, page indices, signatures, titles, and a "others" category for elements that did not fit the previous categories. The annotated bounding boxes were then aligned with the lines from the PyMuPdf v1.21.0 parser [12], as illustrated in Figure 1, to create a supervised line classification corpus. This dataset was subsequently divided into a training set consisting of 215 documents and a test set consisting of 57 documents, the number of corresponding line annotations can be found in Supplementary Material, Table 1.

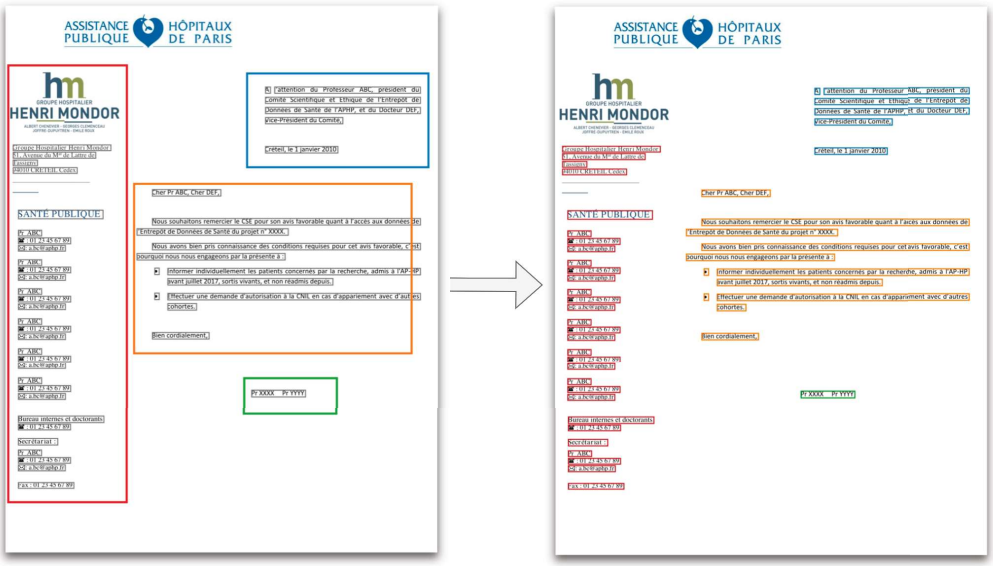



**Figure 1**: Alignment process between annotated bounding boxes and extracted lines using the PyMuPdf parser. The figure illustrates how the different categories of bounding boxes are aligned, enabling the creation of a supervised line classification corpus for further training and analysis.

2. Auto-immune diseases cohort's discharge summaries (medical validation and illustrative use case)

The auto-immune diseases cohort also comes from the AP-HP data warehouse and is restricted to four diseases: Systemic lupus erythematosus, Takayasu disease, scleroderma and antiphospholipid syndrome. This cohort was chosen because of its availability and the fact that an important dataset was already manually annotated.

We had two distinct datasets from this cohort:

- A first set of 151 hospitalization reports, with annotated medical concepts of four different UMLS semantic types: "*Chemicals and drugs*" (e.g., acetaminophen), "*Disorders*" (e.g., meningioma), "*Signs and symptoms*" (e.g., fever, headache), and "*Procedures*" (e.g., brain MRI), and also with vitals parameters present in the text (e.g. body temperature, blood pressure etc.). These reports come from We chose this dataset giving its availability and given the diversity of signs and symptoms. Within this document set, sections ("medical history," "medications," "conclusion," etc...) were also annotated.

  This dataset was used to train (80%) and test (20%) the named entity recognition algorithm

and section identification. The documents were annotated by a physician (C.G.), after a naive body text extraction described below.

- A second set of 200 hospital reports of other patients randomly selected in the auto-immune cohort dataset was annotated by a physician (C.G.) to indicate the presence or absence of an acute infection in the text. The definition of an "acute" infection was considered broadly: it could be bacterial, fungal, viral, or parasitic and could be an active or uncontrolled chronic infection (e.g., an acute complication of HIV disease) or a new infection (e.g., a pulmonary infection requiring hospitalization). In keeping with the usual clinical convention, an infection mentioned in the "clinical progress", "conclusion" or "reason for hospitalization" sections of the report was labeled as an acute infection, while those mentioned in the “medical history” section were considered as old, chronic controlled infections. (e.g., "had pertussis in childhood").

## 2.2. Algorithms’ architectures

The vast majority of clinical documents contained in the Health data Warehouse at the AP-HP are exports from one of the various software programs used by clinicians, and scanned documents make up a negligible fraction (less than 10%) of the total documents. Therefore, we do not resort to Optical Character Recognition (OCR) from rendered documents and focus on text-based document instead.

### Naive algorithm

Our baseline algorithm, which is the one previously used at the AP-HP CDW (Clinical Data Warehouse), generates plain texts by applying a simple mask to the patient documents in PDF format and a pseudonymisation model. This method, while requiring minimal training data (just enough to manually design the mask) and computational resources, effectively extracts the text body from simple documents that conform to the layout used when designing the mask. This method may work in the most common cases, but it does not accommodate different PDF layouts and sometimes produces a mix of administrative information (e.g. dates, hospital wards) and clinically relevant information. In addition, this mask sometimes results in the loss of the original structure of the document, with elements such as section headings being merged at the top of the extracted text. Figure 2 shows an example of the previous text extraction method.

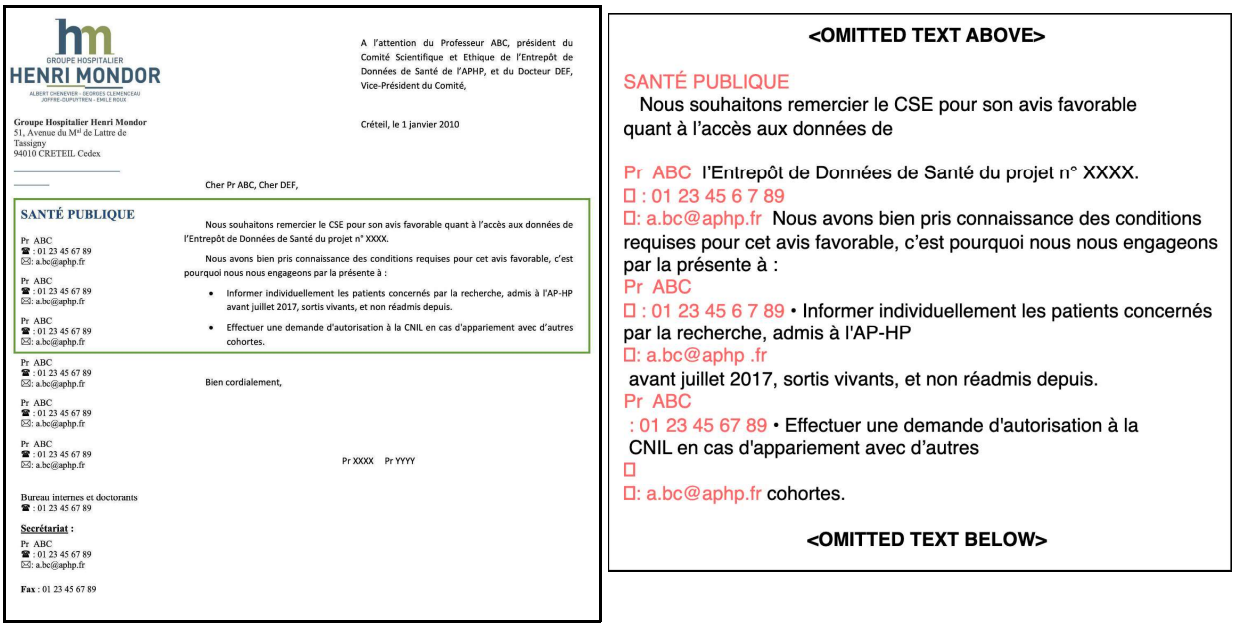



**Figure 2**: Illustration of a failure-case of the naive extraction algorithm. The text snippet shown on the right comes from the green rectangle superimposed on the PDF document on the left. The rules did not correctly identify the side note, resulting in a confusing blend of pertinent text and administrative details, marked in red.

## Advanced algorithm

The end-to-end system comprises a PDF parser to extract lines of text in a PDF, a classifier to infer the type of each line, and an aggregator to compile lines of given labels together to obtain the final textual output. The overall system is illustrated in Figure 3. We chose the PyMuPDF v1.21.0 library [12] to perform the line extraction from the PDF and PyTorch [13] v1.12.1 to implement the neural network. The aggregation is performed by sorting lines of a given label in a top-down left-right fashion and concatenating them.

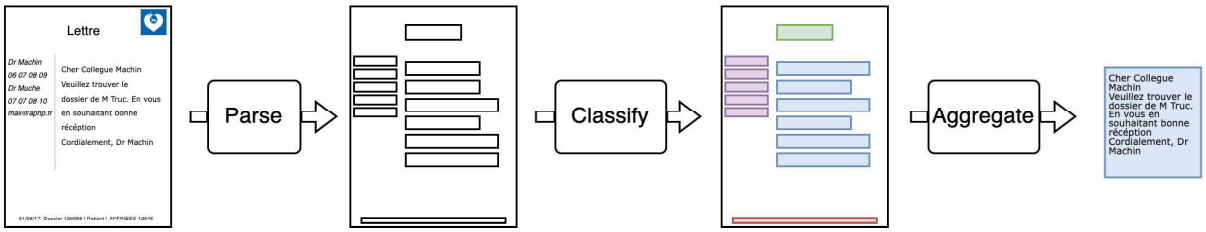



**Figure 3**: Overall system architecture.

## Deep-learning model

The architecture consists of a non-pretrained Transformer that performs classification on extracted lines from the PDF. Each line is represented by a sum of a textual and a layout embedding encoded with 96 dimensions. The textual embedding is a pooled convolution window of 3, 4 and 5 tokens over some embedded text features of the tokens in each line: their 3-letter prefix, 3-letter suffix, shape, and normalized variant, as illustrated by Figure 4. The layout embedding is a concatenation of sinusoidal encodings, introduced by Vaswani (2017) [14], of the left horizontal position, right horizontal position, top vertical bound, bottom vertical bound, width, and height.

Each line representation is then contextualized through a 4-layer Transformer with self-attention. Additionally, we inject the relative 2D distances between lines inside the attention mechanism using a similar mechanism to He et al. (2020) [15]. This attention is the sum of content-content attention (the standard dot product attention), content-position attention, and position-content attention:

$$
\begin{aligned}
\text{attention}(u, v) = & \frac{(W_1^c u) \cdot (W_2^c v)^T}{\sqrt{3d}} && \text{content to content} \\
& + \frac{([W_1^{dx} dx_{u \to v};\, W_1^{dy} dy_{u \to v}]) \cdot (W_2^c v)^T}{\sqrt{3d}} && \text{content to position} \\
& + \frac{(W_1^c u) \cdot ([W_2^{dx} dx_{u \to v};\, W_2^{dy} dy_{u \to v}])^T}{\sqrt{3d}} && \text{position to content}
\end{aligned}
$$

with $W_1^c, W_2^c, W_1^{dx}, W_1^{dy}, W_2^{dx}, W_2^{dy}$ six projection matrices, $dx_{u \to v}$ the embedding of the relative horizontal position of $v$ w.r.t. $u$ and $dy_{u \to v}$ the embedding of the relative vertical position of $v$ w.r.t. $u$.

Finally, each contextualized line embedding is forwarded through a linear layer followed by a softmax to compute the probabilities of each line's label. The complete architecture is described in Figure 4.

The model, containing 3 million parameters, was trained for 1,000 steps by minimizing the cross-entropy loss between the line's label logits and their gold-annotations using the Adam optimizer. The learning rate was scheduled with a warmup of 100 steps followed by a linear decay from 1e-4 to 0. The model architecture and hyperparameters were selected manually by maximizing the F1-score on a 90-10% train-dev split on the corpus full training set.

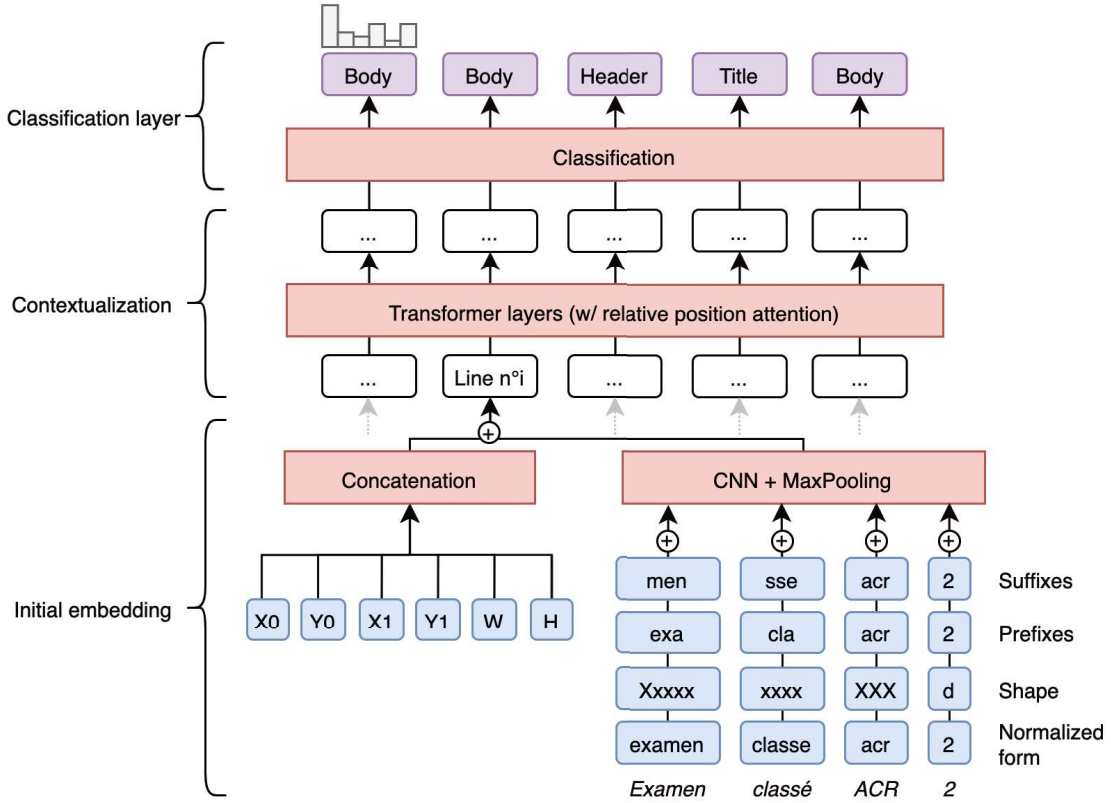



**Figure 4**: Architecture of the deep-learning line classification model. Textual and layout features of each line are embedded to obtain a single representation per line, and are then contextualized with a 4-layer Transformer using self-attention with relative position information. Lastly, they are classified using a linear layer and a softmax function to obtain the probability of each label.

Our implementation is available on github: https://github.com/aphp/edspdf (v0.5.3).

## 2.3. Algorithms' development and technical validation

To validate the performance of our PDF extraction algorithm, we conducted an evaluation using various performance metrics, including precision, recall, F1-score, with micro- and macro-averages. We assessed the performance of the algorithm on the test set of the annotated PDF dataset against the different labels, i.e. body, footer, header, left note, page, others, signature, and title.

Additionally, we performed ablation studies to understand the contribution of different components in the model. These studies involved the removal of the relative position information from the

attention and the removal of the whole Transformer layer, and their effects were observed on micro-and macro-averages of the F1-score, body F1-score, and body recall (resp. F1-score and recall for the *body* section identification). We report the results as the mean over 5 runs with different weight initialization seeds.

## 2.4. Medical validation : medical concepts-detection in their respective sections

Key information in clinical documents is conveyed by medical concepts directly present in the text. These medical concepts are present in their respective sections allowing to explain the clinical reasoning (entry treatments, discharge treatments, personal and family comorbidities, etc.). The hypothesis of our study is that a better extraction of the body text of the PDF - which separates the administrative and non-clinical information - can lead on the one hand to a better extraction performance of the clinical concepts directly and on the other hand to a better extraction of the sections to which these concepts belong.

### 1. Medical concept extraction and classification

Medical concepts can be classified on semantic types as follows as proposed by [16] : *Chemicals, Anatomical structures, Concept and ideas, etc*. These concepts retain the key information we want to extract and have been annotated with four semantic types : *Chemicals and drugs, Signs and symptoms, Diseases and Procedure* on the Autoimmune Disease Cohort dataset, as shown in Figure 5.


**Figure 5**: Example from the Autoimmune Disease Cohort, annotated with named entities that fall into one of the following four semantic types: *Chemicals and drugs, Signs and symptoms, Diseases and Procedure.*

The automatic extraction of these concepts – also called entity recognition – is performed by our previously described algorithm [15, 17]. It is based on an encoder and decoder architecture, similar to that of [18]. Word representations are computed by concatenating output embeddings from the FastText model [19], a BERT Transformer [20], and a character-based CNN (Convolutional Neural Network). These word representations are then re-contextualized using a multi-layer Bi-LSTM. Ultimately, named entities are identified by applying multiple conditional random fields (CRF, one for each label) on the text and disambiguating overlapping entities with the same label by matching the beginning and end boundaries using a biaffine matcher.
This algorithm and its performance at entity level is assessed on the 151 annotated documents with and without the advanced layout detection algorithm. The metrics used for performance evaluation are precision, recall and F1 score.

Finally, we used our multilabel medical concept classifier to classify all symptoms and disorders in the main medical domains (cardiology, neurology, etc..) -corresponding to the MeSH[21] -C headings- [17]. Specifically in our "Acute infection" use case, this algorithm predicts whether a disease is an infection or not.

## 2. Section identification

Like previous authors [11], we assumed that all sections are preceded by section titles (which was the case in 150 over 151 documents). For the detection of the section title, a custom dictionary was

created. The level of granularity of sections versus subsections was discussed collegially and, drawing on previous work [11], a set of 14 sections of interest was selected (see Table 1 and Supplementary data Table 2 for the synonyms dictionary).

| | |
|---|---|
| *Antécédents* | History |
| *Antécédents familiaux* | Family history |
| *Allergies* | Allergies |
| *Mode de vie* | Lifestyle |
| *Traitement* | Treatment |
| *Traitement entrée* | Treatment at admission |
| *Traitement sortie* | Discharge treatment |
| *Motif* | Motive (Reasons of admission) |
| *Histoire de la maladie* | History of the actual disease |
| *Evolution* | Clinical Progress |
| *Examen clinique* | Physical examination |
| *Constantes* | Vitals |
| *Examens complémentaires* | Complementary investigations |
| *Conclusion* | Conclusion |

**Table 1**: List of section types to be identified in a clinical report.

For simplicity and clarity, section identification was based on rules, taken from the custom dictionary, directly searching for exact mentions of the section title or synonyms. The method was tested on the auto-immune diseases cohort dataset with and without the advanced layout detection algorithm (described in Section II.A). The section types were all manually annotated by a clinician (C.G.) in the 151 documents extracted by the naive mask. When the structure was not explicit, the section starts

were inferred directly from the clinical information (e.g., family history for "Diabetes in brother and sister", etc.). We evaluate the ability of the system to extract the correct medical concepts in the correct sections of interest. The metrics used to assess performance were precision, recall and F1 score.

### 2.5. Illustrative use case: automatic detection of acute infections in the text.

Finally, we wanted to illustrate how our advanced body extraction algorithm, combined with section extraction and medical concept recognition and classification, could automatically detect acute infections in medical reports. Acute infection was chosen since it is a frequent disorder in auto-immune patients treated with corticosteroids or other immunosuppressive drugs. The global approach described above was used: first, the body text of the report was extracted, then, the sections were identified, and finally the concepts in the "Evolution" and "Conclusion" sections were extracted and classified. A concept classified as *infection* in the "Evolution" or "Conclusion" sections categorizes the patient as having an acute infection (including chronic decompensated infections requiring acute management).

## 3. Results

### 3.1. Text-extraction and line classification results (technical validation)

Table 2 displays the per-line micro-averaged precision, recall, and F1-score for our text-extraction and classification algorithm. The algorithm achieved a precision of 0.98, a recall of 0.97, and an F1-score of 0.98 for the “body” lines, and an overall micro-average of 0.96 for lines of all types.

Table 3 presents the findings of the ablation study, illustrating a decline in performance for both ablations. The performance degradation of the simplest, non-contextualized model ranges from

approximately 2 points in body recall, and body F1-score, 4.5-point decrease in micro-averaged F1-score, to an almost 6-point drop in macro-averaged score.

| **Label** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| body | 0.98 | 0.97 | 0.98 |
| footer | 0.84 | 0.88 | 0.85 |
| header | 0.90 | 0.95 | 0.93 |
| left_note | 0.98 | 0.99 | 0.99 |
| page | 0.96 | 0.91 | 0.94 |
| others | 0.97 | 0.92 | 0.94 |
| signature | 0.87 | 0.80 | 0.83 |
| title | 0.87 | 0.80 | 0.84 |
| ALL (macro-avg) | 0.92 | 0.90 | 0.91 |
| ALL (micro-avg) | 0.96 | | |

**Table 2**: Per-line precision, recall and F1-score of the body-extraction algorithm for the test set and for each line type (technical validation).

| Model | **Micro-avg F1** | **Macro-avg F1** | **Body F1** | **Body recall** |
|---|---|---|---|---|
| Full model | 0.96 | 0.91 | 0.98 | 0.97 |
| -Relative position attention | 0.95 | 0.89 | 0.97 | 0.96 |
| -Transformer | 0.92 | 0.85 | 0.95 | 0.95 |

**Table 3**: Ablation study of model architecture.

### 3.2. Section and medical concepts extraction results (medical validation)

Results on entity extraction in their respective sections are shown in Table 4. Only pairs (entities/sections) of medical interest were kept (i.e. Drugs in the sections “drugs at entry” and “drugs at discharge”, Symptoms in the section “physical examination” etc..) . The entities are annotated and the sections are extracted with a rule-based algorithm (and compared to a gold standard annotation). We compared both extractions: with the naive body extraction algorithm, a baseline mask, and with the advanced body extraction and noticed an overall improvement of 0.1 for the F1 score of the entities detection within their right respective sections.

| | **Naive body extraction** | | | **Advanced body extraction** | | |
|---|---|---|---|---|---|---|
| **Entity** ***Corresponding section*** | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| Drugs *at entry* | 0.68 | 0.99 | 0.81 | 0.91 | 0.98 | **0.94** |
| Drugs *at discharge* | 0.82 | 0.91 | 0.86 | 0.89 | 0.99 | **0.94** |
| Vital parameters *physical examination* | 0.84 | 0.93 | 0.88 | 0.94 | 0.99 | **0.96** |
| Procedure *investigations* | 0.83 | 0.98 | 0.90 | 0.87 | 0.98 | **0.92** |
| Signs and Symptoms *medical history* | 0.79 | 0.65 | 0.71 | 0.90 | 0.81 | **0.86** |
| Signs and Symptoms *family medical history* | 0.71 | 0.99 | 0.83 | 0.80 | 0.98 | **0.88** |
| Signs and Symptoms *conclusion* | 0.73 | 0.52 | 0.61 | 0.94 | 0.65 | **0.77** |
| Signs and Symptoms *clinical progress* | 0.81 | 0.83 | 0.82 | 0.85 | 0.84 | **0.84** |

| | | | | | | |
|---|---|---|---|---|---|---|
| Signs and Symptoms *clinical examination* | 0.83 | 0.80 | 0.82 | 0.95 | 0.97 | **0.96** |
| Signs and Symptoms *history of the actual condition* | 0.64 | 0.91 | 0.75 | 0.85 | 0.95 | **0.90** |
| Overall | 0.77 | 0.85 | 0.80 | 0.81 | 0.91 | **0.90** |

**Table 4** : Entity with section types as precision, recall and F1 score attributes, with naive (left) and advanced (right) body extraction. Only pairs (entities, sections) of medical interest were retained. "Vital parameters" correspond to: blood pressure, temperature, heart rate, etc. of the patient mentioned directly in the text.

### 3.3. Detection of acute infections results (illustrative use case)

The results of the overall pipeline on document classification for acute infection causing an hospitalization are presented in Table 5. We show that the advanced body detection algorithm enabled an improvement of +7 percentage points on the F1-score for detecting acute infections.

| | Naive body extraction | | | Advanced body extraction | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Acute infection (n = 200 documents) | 0.80 [0.71; 0.90] | 0.74 [0.63; 0.84] | 0.77 [0.70; 0.85] | **0.83** [0.73; 0.92] | **0.85** [0.77; 0.94] | **0.84** [0.77; 0.90] |

**Table 5** : Precision, recall, and F1 score per document of the acute infection detection algorithm with and without the advanced body detection algorithm (illustrative use case). Confidence intervals were obtained using the bootstrap method.

# 4. Discussion

Health data warehouses often provide researchers with data that has undergone several stages of preprocessing from the original medical record. Ensuring the reliability of this data is critical to obtaining meaningful analyses. In this study, we introduce a lightweight algorithm for extracting clinical text from clinical PDF documents using a transformer-based model. Our approach demonstrates promising results in various aspects of information extraction, ranging from line-by-line text extraction from the body of the document to named entity recognition tasks in specific sections of the document, and to impact on a specific medical use case.

## Model architecture

Our new advanced body text extraction method preserves the original structure of clinical documents, which is of key importance to researchers and clinicians, as it allows for better interpretability, close to clinical reasoning.

Our ablation studies highlight the importance of incorporating structural features, such as contextual embeddings and relative position information encoding. Despite having a limited number of parameters, our model achieves high scores for text-body extraction. These results suggest that both contextualization through the Transformer layer and the relative position encoding in the attention mechanism play important roles in effectively capturing the structure and layout of the PDF documents, ultimately leading to more accurate segmentation and extraction.

### Impact on downstream clinical tasks

The clinical case presented in this study is essentially a proof of concept that can be generalized to many other applications. Indeed, no supervised learning was required for the detection of the clinical cases of acute infection: the extraction of medical concepts in their respective sections is a completely generic method.

Acute infection was chosen because it is a common medical complication in our autoimmune diseases, with patients often on immunosuppressive therapy. But any other phenotype could have been detected as long as it was explicitly mentioned in the text. The interest of this end-to-end method could also have been illustrated to filter out, for example, patients under certain treatments on arrival at the hospital, patients with specific family comorbidities, patients coming for a particular reason ("fever", "relapse", etc.).

### Impact on pseudonymisation

An additional advantage of the PDF segmentation method proposed in this study is its potential impact on pseudonymization. In a previous article by Tannier et al. [Tannier2023], it was shown that employing this PDF segmentation method significantly reduces the number of identifying entities in

the extracted text, namely by 80%. This reduction is primarily due to the majority of identifying entities being located in side-notes, footers, and headers, which are effectively segmented and removed by our algorithm. As a result, the application of this segmentation method has been observed to increase the proportion of PDFs that have been thoroughly stripped of identifying words from 75.7% to 93.1%. This comparison was made by examining the same documents converted to text using both the naive text-extraction algorithm, i.e., a legacy rigid mask method, and the new advanced algorithm presented here. Consequently, our PDF segmentation approach not only improves the extraction of clinically relevant information but also contributes to better privacy protection through more effective pseudonymization.

## Limitations

We have not directly compared our algorithm to others like LayoutLM due to their significantly larger size and increased computational requirements. Our model, containing only 3M parameters, is fast enough to fit into the daily integration of 200,000 PDF files, the text extraction step taking approximately 15 minutes with the use of 2 GPUs and 16 cores. In contrast, the base version of LayoutLMv3, with its 133M parameters, demands considerably more computational resources and processing time. The efficiency of our algorithm in terms of time and computational resources makes it a more suitable option for large-scale PDF segmentation tasks, such as those encountered in the AP-HP's clinical data warehouse, which contains more than 120 million documents to process.

With regard to the classification of acute infection documents, we based our classification on the structure of the hospitalization reports and defined as "acute" all infectious diseases mentioned in the "clinical course" and "conclusion" section, which mainly corresponds to an actually acute infection during the hospitalization but can sometimes be simply old infections that the clinician may decide to mention.

### Perspectives

In this work, we mainly focused on body text extraction for better clinical analysis of documents; however, other applications can be found such as table and form detection, or metadata extraction

from header sections for administrative and data quality validation, which will be performed in future work.

# 5. Conclusion

In this study, we developed and validated a lightweight transformer-based algorithm for extracting clinically relevant text from PDF documents. The algorithm efficiently handles various layouts of clinical PDF documents and improves the performance of downstream natural language processing tasks. Our approach demonstrates its effectiveness in preserving the original structure of clinical documents, resulting in improved interpretability and alignment with clinical reasoning, which is particularly valuable for researchers and clinicians. In addition, the computational resource efficiency of the model makes it suitable for large-scale PDF segmentation tasks in environments such as the AP-HP clinical data warehouse. The performance evaluation showed promising results in extracting medical concepts in their respective sections and on an illustrative medical use case.

# Acknowledgment

We thank the clinical data warehouse (Entrepôt de Données de Santé, EDS) of the Greater Paris University Hospitals for its support and the realization of data management and data curation tasks. We thank Xavier Tannier and Fabrice Carrat for fruitful discussions.

# Authors contribution

P.W., A.C. and B.D. had full access to all the data in the study. They take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: C.G., P.W., B.D., A.C., A.M., R.B.
Annotation and interpretation of data: C.G., B.D., A.C., P.W.
Manuscript drafting: C.G., P.W., R.B.
Algorithm development: P.W., B.D., A.C.
Manuscript critical proofreading C.G., P.W., A.M., B.D., R.B

Supervision: R.B.

# Conflict of Interest Disclosures

None reported.

# Data sharing

Access to the clinical data warehouse's raw data can be granted following the process described on its website: eds.aphp.fr. A prior validation of the access by the local institutional review board is required. In the case of non-APHP researchers, the signature of a collaboration contract is moreover mandatory.

# Funding/Support

This study has been supported by grants from the AP-HP Foundation.

# Role of the Funder/Sponsor

The funder was involved neither during the design and conduct of the study nor during the preparation, submission or review of the manuscript.

# Bibliography

[1] Liu X, Gao F, Zhang Q, Zhao H. Graph convolution for multimodal information extraction from visually rich documents. arXiv preprint arXiv:1903.11279. 2019 Mar 27.

[2] Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. Layoutlm: Pre-training of text and layout for document image understanding. InProceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2020 Aug 23 (pp. 1192-1200).

[3] Xu Y, Xu Y, Lv T, Cui L, Wei F, Wang G, Lu Y, Florencio D, Zhang C, Che W, Zhang M. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740. 2020 Dec 29.

[4] Huang Y, Lv T, Cui L, Lu Y, Wei F. Layoutlmv3: Pre-training for document ai with unified text and image masking. InProceedings of the 30th ACM International Conference on Multimedia 2022 Oct 10 (pp. 4083-4091).

[5] Majumder BP, Potti N, Tata S, Wendt JB, Zhao Q, Najork M. Representation learning for information extraction from form-like documents. Inproceedings of the 58th annual meeting of the Association for Computational Linguistics 2020 Jul (pp. 6495-6504).

[6] Kim G, Hong T, Yim M, Nam J, Park J, Yim J, Hwang W, Yun S, Han D, Park S. Ocr-free document understanding transformer. InComputer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII 2022 Oct 20 (pp. 498-517). Cham: Springer Nature Switzerland.

[7] Yu W, Lu N, Qi X, Gong P, Xiao R. PICK: processing key information extraction from documents using improved graph learning-convolutional networks. In2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 4363-4370). IEEE.

[8] Kropf S, Krücken P, Mueller W, Denecke K. Structuring legacy pathology reports by openEHR archetypes to enable semantic querying. *Methods Inf Med*. 2017;56(3):230–7.

[9] Zweigenbaum P, Deléger L, Lavergne T, Névéol A, Bodnari A. A supervised abbreviation resolution system for medical text. *In: Working Notes for CLEF Conference, Valencia, Spain*, September 23-26, volume 1179 of CEUR Workshop Proceedings; 2013

[10] Edinger T, Demner-Fushman D, Cohen AM, Bedrick S, Hersh W. Evaluation of clinical text segmentation to facilitate cohort retrieval. *AMIA Annu Symp*. 2017;2017:660–9.

[11] Pomares-Quimbaya A, Kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. BMC medical research methodology. 2019 Dec;19(1):1-20.

[12] Liu, R., & McKie, J. X.., PyMuPDF. Available at: http://pymupdf.readthedocs.io/en/latest/ [Accessed april 25, 2023]

[13] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32 [Internet]. Curran Associates, Inc.; 2019. p. 8024–35.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009.

[He2020] He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. http://arxiv.org/abs/2006.03654

[15] Perceval Wajsbürt. Extraction and normalization of simple and structured entities in medical documents. Santé publique et épidémiologie. Sorbonne Université, 2021. English. ⟨NNT : 2021SORUS541⟩. ⟨tel-03624928v2⟩

[16] McCray, A. T.; Burgun, A. & Bodenreider, O.Aggregating UMLS semantic types for reducing conceptual complexity.Studies in health technology and informatics, 2001, 84, 216-220

[17] Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artificial Intelligence in Medicine. 2022 Jun 1;128:102311.

[18] Yu, Juntao, Bernd Bohnet, and Massimo Poesio. "Named entity recognition as dependency parsing." *arXiv preprint arXiv:2005.07150* (2020).

[19] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* 5 (2017): 135-146.

[20] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[21] https://www.ncbi.nlm.nih.gov/mesh/

[Tannier2023] Tannier, X., Wajsbürt, P., Calliger, A., Dura, B., Mouchet, A., Hilka, M., & Bey, R. (2023). *Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse*. *arXiv preprint arXiv:2303.13451 (2023)*

# Supplementary materials

Supplementary Table 1: Lines repartition for the Train and Test datasets.

| Line label | Train | Test |
|---|---|---|
| **Body** | 11,841 | 3,275 |
| **Footer** | 602 | 136 |
| **Header** | 3,748 | 927 |
| **Left note** | 4,903 | 1,325 |
| **Page index** | 232 | 67 |
| **others** | 608 | 185 |
| **Signature** | 245 | 74 |
| **Title** | 267 | 59 |
| **Total documents** | 215 | 57 |
| **Total pages** | 423 | 115 |
| **Total lines** | 22,446 | 6,048 |

Supplementary Table 2: Details of the section's name (In French) for the rule-based section detection.

| Section type | Term |
|---|---|
| Conclusion | conclusion |
| Conclusion | Au total |
| Conclusion | conclusion médicale |
| Conclusion | Synthèse du séjour |
| Conclusion | Synthèse |
| Conclusion | conclusion de sortie |
| Conclusion | syntese medicale / conclusion |
| Conclusion | synthese medicale |
| Conclusion | conclusion consultation |
| Conclusion | diagnostic retenu |
| | |
| Histoire de la maladie | histoire de la maladie |
| Histoire de la maladie | histoire recente |
| Histoire de la maladie | histoire recente de la maladie |
| Histoire de la maladie | rappel clinique |
| Histoire de la maladie | resume |
| Histoire de la maladie | resume clinique |

| | |
|---|---|
| Histoire de la maladie | histoire de la maladie - explorations |
| Histoire de la maladie | histoire de la maladie actuelle |
| Histoire de la maladie | évènements récents |
| Histoire de la maladie | evolution depuis la dernière consultation |
| Histoire de la maladie | Résumé clinique - Histoire de la maladie |
| Histoire de la maladie | histoire clinique |
| Histoire de la maladie | rappel |
| Histoire de la maladie | Rappel de la conclusion de la précédente consultation |
| Histoire de la maladie | Evaluation des effets secondaires en intercure |
| Histoire de la maladie | historique |
| Histoire de la maladie | Pour mémoire |
| Histoire de la maladie | Anamnèse |
| Histoire de la maladie | Résumé des séances |
| | |
| Indication | indication |
| Indication | Contexte clinique |
| Indication | renseignements cliniques |
| Indication | Motif de l'examen |
| | |
| Résultats | résultats |
| | |
| Motif | motif |
| Motif | motif médical |
| Motif | motif d'hospitalisation |
| Motif | motif de la consultation |
| Motif | motif de consultation |
| Motif | motif de la consultation recueilli par l'iao |
| Motif | motif de l'hospitalisation |
| Motif | Motif de prise en charge |
| Motif | Motif d'admission |
| Motif | Motif de présentation |
| | |
| Examens complémentaires | examens complémentaires |
| Examens complémentaires | examens complémentaires (résultats et commentaires) |
| Examens complémentaires | biologie |
| Examens complémentaires | ECG |
| Examens complémentaires | biochimie |
| Examens complémentaires | Scanner |
| Examens complémentaires | Radiographie |
| Examens complémentaires | résultats de biologie |

| | |
|---|---|
| Examens complémentaires | imagerie |
| Examens complémentaires | PET-scanner |
| Examens complémentaires | examen complémentaire réalisé à l'entrée |
| Examens complémentaires | echocardiographie |
| Examens complémentaires | interpretation des examens complementaires |
| Examens complémentaires | examen(s) complementaire(s) |
| Examens complémentaires | examens complementaires a l'entree |
| Examens complémentaires | examens complementaires realises pendant le sejour |
| Examens complémentaires | examens para-cliniques |
| Examens complémentaires | biologie à l'entrée |
| Examens complémentaires | examen histologique |
| Examens complémentaires | examens à l'entrée |
| Examens complémentaires | anatomo-pathologie |
| Examens complémentaires | Résultat du bilan |
| Examens complémentaires | Diagnostic histopathologique |
| Examens complémentaires | Examen cytologique |
| Examens complémentaires | Test au Synacthène |
| Examens complémentaires | Mammographie |
| Examens complémentaires | Etude immunohistochimique |
| Examens complémentaires | RADIOGRAPHIE THORACIQUE |
| Examens complémentaires | Examens d'imagerie |
| Examens complémentaires | Bilan génétique |
| Examens complémentaires | EXAMENS PARACLINIQUES |
| Examens complémentaires | Ionogramme |
| | |
| Autres | avis specialises |
| Autres | document remis au patient |
| Autres | dossier social |
| Autres | décision d'orientation |
| Autres | risque infectieux |
| Autres | liste des prescriptions demandées |
| Autres | action iao |
| Autres | actes ide réalisés à l'accueil |
| Autres | devenir réel du patient |
| Autres | dossier traumatologie |
| Autres | affection exonerante |
| Autres | modalites de sortie |
| Autres | documents remis au patient |
| Autres | rendez-vous pris |
| Autres | mode d'arrivée |

| | |
|---|---|
| Autres | planification des soins / suites à donner |
| Autres | CONSULTATION POST URGENCE |
| Autres | relecture du dossier |
| Autres | décision de la rcp |
| Autres | suivi post HdJ |
| Autres | codage |
| Autres | planification des soins |
| Autres | avis psychiatrique |
| Autres | Documents de sortie |
| Autres | Rendez-vous programmés |
| Autres | SYNTHESE INFIRMIERE |
| Autres | SYNTHESE INFIRMIERE A LA SORTIE |
| Autres | TRANSMISSION IDE |
| Autres | COMMENTAIRES |
| Autres | SCORES |
| Autres | Soins de rééducation |
| Autres | Soins infirmiers |
| | |
| Examen clinique | examen clinique initial |
| Examen clinique | examen clinique |
| Examen clinique | constantes initiales |
| Examen clinique | examen clinique à l'entrée |
| Examen clinique | examen dermatologique |
| Examen clinique | Evaluation de l'état général |
| Examen clinique | EXAMEN D'ENTRÉE |
| Examen clinique | Contexte clinique du patient |
| Examen clinique | CLINIQUE |
| Examen clinique | EXAMEN CLINIQUE A L'ADMISSION |
| | |
| Traitement | attitude therapeutique initiale |
| Traitement entrée | traitement à l'entrée |
| Traitement entrée | traitement en cours |
| Traitement entrée | traitement habituel |
| Traitement entrée | traitement actuel |
| Traitement de sortie | Traitement de sortie |
| Traitement de sortie | traitement et ordonnance de sortie |
| Traitement de sortie | ordonnance de sortie |
| Traitement | traitement |
| Traitement | prescriptions |
| Traitement | autres prescriptions |

| | |
|---|---|
| Traitement | prescriptions relatives au traitement de l'affection de longue durée reconnue |
| Traitement de sortie | prescriptions de sortie |
| Traitement de sortie | prescriptions medicales de sortie |
| Traitement | prescriptions sans rapport avec l'affection de longue durée |
| Traitement | TRAITEMENTS PRESCRITS |
| Traitement | traitements en cours ou d'administration recente |
| Traitement | médicaments |
| Traitement de sortie | prescriptions à l'issue de la consultation |
| Traitement | Traitement et surveillance |
| Traitement | Anti-infectieux |
| Traitement | Traitement spécifique |
| | |
| Antécédents | antécédents |
| Antécédents | atcd |
| Antécédents | antécédents médicaux |
| Antécédents | antécédents - allergies |
| Antécédents | antecedents medicaux et chirurgicaux |
| Antécédents | antecedents personnels |
| Antécédents familiaux | antécédents familiaux |
| Antécédents | antécédents chirurgicaux |
| Allergies | allergie |
| Allergies | Allergies connues |
| | |
| Antécédents | facteurs de risque cardiovasculaires |
| Antécédents | autres antecedents medicaux ou chirurgicaux |
| Antécédents | antécédents gynéco-obstétriques |
| Antécédents | sur le plan des facteurs de risque cardiovasculaires |
| Antécédents | Vaccinations |
| Antécédents | antécédents cardiaques |
| Antécédents | diagnostics associés |
| Antécédents | facteurs de risques |
| Antécédents | médicaux |
| Antécédents | chirurgicaux |
| Antécédents | Antécédents et mode de vie |
| Antécédents | Antécédents psychiatriques |
| Antécédents | Antécédent médico-chirurgical |
| Antécédents | vaccins |
| | |
| Evolution | Evolution clinique |
| Evolution | evolution |

| | |
|---|---|
| Evolution | évènements recensés au cours de ce séjour |
| Evolution | Evolution dans le service |
| Evolution | Conclusion à l'entrée |
| Evolution | Prise en charge |
| Evolution | Traitement / Evolution dans le service |
| Evolution | Evolution post greffe : |
| Evolution | HYPOTHESES DIAGNOSTIQUES : |
| Evolution | Suites opératoires : |
| | |
| Constantes | Paramètres vitaux à l'entrée |
| Constantes | parametres vitaux initiaux |
| Constantes | constantes initiales |
| Constantes | constantes |
| Constantes | dernières constantes |
| Constantes | donnees biometriques et parametres vitaux a l'entree |
| Constantes | parametres vitaux et donnees biometriques a l'entree |
| Constantes | ACCUEIL IAO |
| Constantes | Paramètres vitaux à l'accueil des urgences |
| Constantes | PARAMETRES DE SURVEILLANCE |
| | |
| Mode de vie | mode de vie |
| Mode de vie | habitus |
| Mode de vie | mode de vie - scolarite |
| Mode de vie | situation sociale, mode de vie |
| Mode de vie | contexte familial et social |
| Mode de vie | Environnement familial |
| Mode de vie | MDV |
| Mode de vie | Mode de vie et éléments biographiques |

## 4.2. Vue d'ensemble du prototype

Nous nous plaçons dans une situation clinique où un patient souffrant d'une maladie systémique nécessite une décision thérapeutique collégiale (par exemple : rechute d'une néphropathie lupique, récidive d'une thrombose chez un patient Takayasu, etc...). Pour ce patient à traiter, nous disposons d'une observation médicale textuelle d'hospitalisation, comprenant généralement le motif d'hospitalisation, les antécédents médicaux, l'histoire de la maladie et des traitements entrepris dans le passé, les symptômes cliniques (ex : arthralgies, etc.) et para-cliniques (ex : hypermétabolisme au PET scan, etc.), les traitements en cours et les résultats biologiques. Ces informations sur le patient à traiter sont automatiquement extraites par notre algorithme d'extraction d'entités nommées présenté dans les sections suivantes. La cohorte de patients similaires est construite au niveau du document, c'est-à-dire que l'algorithme extrait automatiquement de la base de données tous les patients dont les rapports médicaux mentionnent une histoire similaire, avec des symptômes cliniques et des résultats d'examens paracliniques communs.

La construction de cette cohorte est réalisée par un algorithme résumé dans la figure 19 :

1. En raison du volume important des données, une première étape consiste à sélectionner un premier sous-ensemble de patients d'intérêt, sur la base du diagnostic principal du patient (étape 1 Figure 19). Cette étape a été validée par un premier algorithme de recherche de mots-clés et de synonymes dans le texte (présenté dans la section 4.3. ci-dessous).

2. Pour les patients présélectionnés, une distance avec le patient index est calculée au niveau du document à partir de la similarité des termes (symptômes, comorbidités...) dans les rapports et les résultats biologiques (étape 2). Le calcul de la distance se fait en plusieurs étapes. Une étape d'extraction des termes médicaux (section 4.4.), puis, dans une deuxième étape, tous les symptômes et comorbidités sont classés dans un domaine médical principal, à savoir l'hématologie, la cardiologie, etc. (section 4.5.).

3. Cette distance peut ensuite être modulée, selon les souhaits du clinicien : soit en se basant uniquement sur certains domaines d'intérêt, soit en calculant une distance globale, pondérée sur tous les domaines médicaux. Le clinicien peut également décider d'ajouter à la distance entre les patients des informations biologiques quantitatives pour les tests biologiques fréquents, tels que les niveaux d'hémoglobine ou de créatinine. L'inclusion d'informations biologiques est présentée dans la section 4.7.

4. Un algorithme de regroupement hiérarchique ascendant est appliqué pour sélectionner le groupe de patients le plus proche avec cette distance multidimensionnelle, correspondant à la cohorte sélectionnée de patients similaires.

5. Pour ces patients similaires sélectionnés, les médicaments sont extraits, classés selon leurs classes thérapeutiques (voir section 4.7.), comme proposé par la classification ATC [15], quatrième niveau, c'est-à-dire les sous-groupes chimiques. Les complications mentionnées dans le texte (infection, insuffisance rénale ou hépatique, thromboembolique, etc.) sont extraites et leur risque relatif, associé à chaque classe thérapeutique, est calculé (étape 4 de la figure 19).

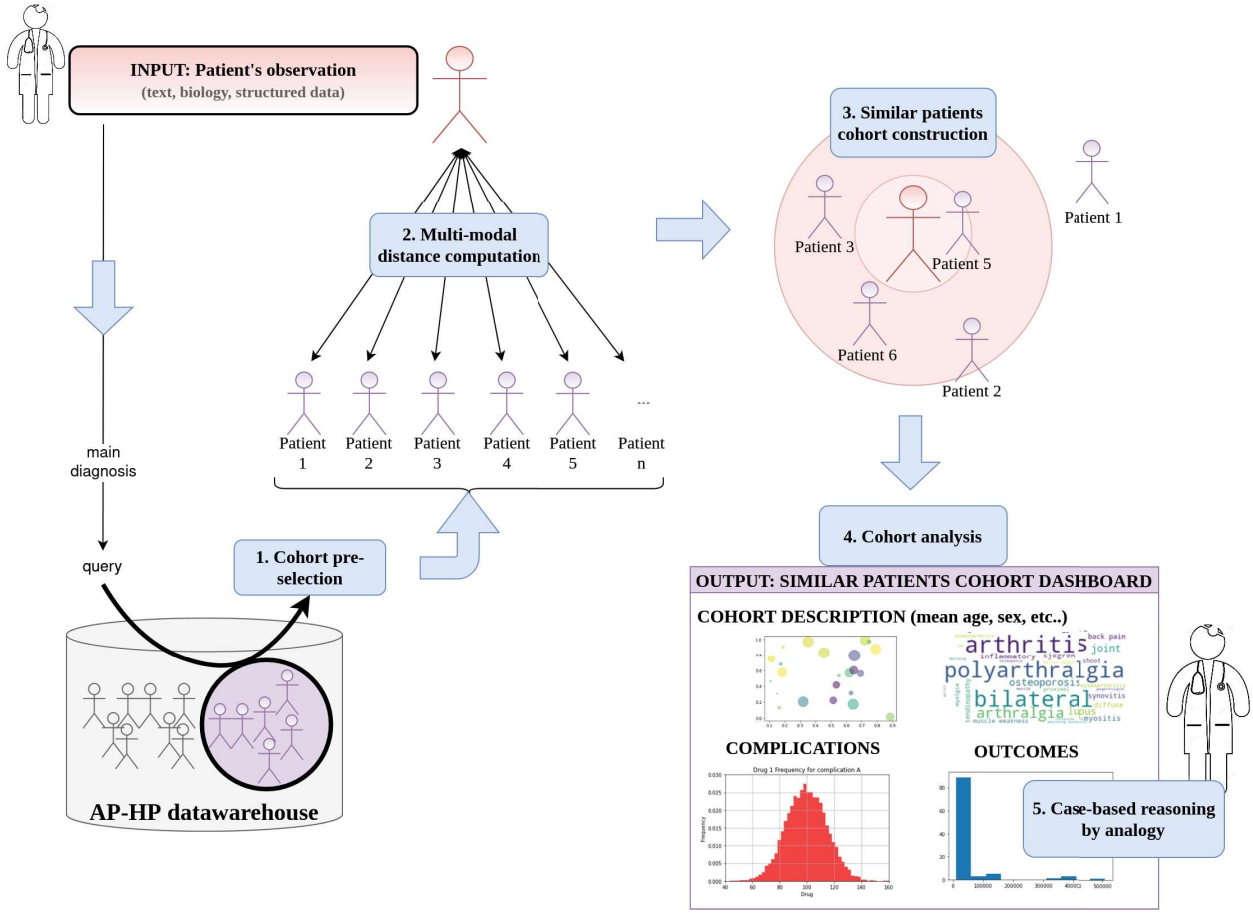



**Figure 19**: Schéma général du prototype.

6. Le résultat final pour les cliniciens est la description de la cohorte de patients similaires (âge moyen, sexe, concepts médicaux extraits des rapports sous forme de nuages de mots et de variance intra-groupe), les risques relatifs associés à chaque classe thérapeutique rencontrée pour les complications mentionnées et les résultats pour le patient (décès, réhospitalisation à 28 jours, etc.),

présentés par des outils de visualisation permettant une meilleure interprétation et une meilleure lisibilité pour permettre un raisonnement basé sur le cas (étape 5).

Ces dernières étapes 4, 5 et 6 sont en cours de développement et ne sont pas présentées dans ce manuscrit.

## 4.3. Extraction automatisée de documents d'intérêt sur une base de données large (article 2 *en review*)

Résumé :

Une première étape de présélection des patients d'intérêt est réalisée pour restreindre les documents candidats sur lesquels les algorithmes d'extraction d'entités nommées et de mesure de similarité seront ensuite appliqués. Cette étape a donc pour objectif d'optimiser l'extraction des documents textuels médicaux se rapportant à une caractéristique clinique spécifique (symptômes ou maladie) à partir de l'EDS.

Nous avons comparé deux types de méthodes. La première, centrée sur le document, est basée sur la représentation vectorielle des documents : tous les documents ayant une représentation vectorielle proche de celle d'un document index initial (c'est-à-dire les documents présentant les mêmes maladies ou symptômes mentionnés) sont sélectionnés. Cette méthode est directement pertinente dans notre contexte de création de cohorte de patients similaires, en effet, le cas d'usage ici est de retrouver automatiquement tous les comptes-rendus hospitalier potentiellement proches de l'observation initiale (ou bilan médical initial) du patient que l'on souhaite traiter. Les modèles testés ici doivent pouvoir représenter l'information du document qui reflète son phénotype (c'est-à-dire l'ensemble des maladies et symptômes observés et décrits dans le compte-rendu).

Trois modèles de représentation vectorielle de document ont été testés : TF-IDF [67], doc2vec [61] et docBERT [62].

La deuxième méthode est basée sur une expansion de requête : le clinicien saisit un diagnostic d'intérêt, et une recherche textuelle est effectuée en utilisant ces mots-clés, ses synonymes dans l'UMLS [16] (décrit section 2.1.2.) et les termes proches trouvés par similarité dans les embeddings CODER [60] (décrit section 2.3.).

Pour la recherche de documents similaires par représentation vectorielle, nous avions a disposition un jeu de 256 documents annotés avec 4 phénotypes d'intérêt par un clinicien extérieur à l'étude : la néphropathie lupique (48 documents), l'ostéoporose (23 documents), l'infection pulmonaire (33 documents) et la pneumopathie interstitielle secondaire à la sclérodermie (20 documents) ; ces documents correspondent à "l'observation initiale" de notre cas d'usage.

La difficulté de ce travail consiste essentiellement dans l'obtention de métriques fiables, en effet, du fait du grand nombre de documents sur lequel la requête est réalisée (en tout 2 000 000 de compte-rendus environs dont 10% de compte-rendus hospitaliers), il n'est pas humainement

raisonnable de lire et classer les documents selon les phénotypes d'intérêt, l'obtention d'un *gold standard* (annotation manuelle avec un degré de précision élevée) n'est donc pas envisageable. Deux *silver standards* (des annotations automatiques de précision moins élevée) ont donc été utilisés :

1. La présence d'une liste de mot clés dans les textes avec une liste de termes pré-établies manuellement pour tous les synonymes possibles rattachés au phénotypes d'intérêt (par exemple pour le phénotype "néphropathie lupique", les termes suivants étaient proposés : "glomérulonéphrite", "LGM"-en respectant les majuscules, "GEM", "néphrite" etc… associé au terme "lupique", "lupus" ou "LES" dans le texte). Cette liste de termes pré-établis a été réalisée par un clinicien interniste extérieur à l'étude.
2. Le codage CIM-10 des pathologies pour le séjour associé aux comptes-rendu hospitaliers, permettant d'évaluer au mieux la sensibilité de notre algorithme. Malheureusement, comme d'ailleurs constaté au cours de l'étude publié, le codage CIM10 en France est essentiellement utilisé pour valoriser financièrement les séjours des patients, ce qui induit des biais de surreprésentation des pathologies mieux valorisées et un "découplage" des pathologies. Par exemple, le codage pour "glomérulopathie au cours du lupus érythémateux disséminé" serait N085 mais il n'est pas utilisé (et d'ailleurs absent dans la base) au profit de deux codages : celui de l'insuffisance rénale et celui du lupus érythémateux disséminé. De même à la lecture des textes on constate par exemple que la "tuberculose ganglionnaire" est codée "tuberculose pulmonaire" le remboursement est le même mais le sens clinique est différent.

Les résultats obtenus par la méthode de représentation vectorielle des documents étaient décevant en particulier avec le docBERT [62], du fait notamment de la limitation de la longueur de représentation à 512 tokens.

La méthode d'expansion de requête à partir de l'UMLS [16] (et de l'algorithme du CODER [60]) permettait d'obtenir, sur 20 phénotypes (également proposés par un clinicien extérieur à l'étude), une précision moyenne (valeur prédictive positive) de 0,93 [0,90 ; 0,96] évaluée directement par un clinicien sur un ensemble de 50 textes cliniques aléatoires par phénotype et un rappel (ou sensibilité) de 0,78 [0,71 ; 0,85], évalué sur la base des codes CIM10.

Au-delà des résultats présentés dans l'article ci-dessous, en cours de review, nous souhaitons ajouter le tableau 3 suivant pour permettre une comparaison des deux méthodes.

| **modèle** | *Néphropathie lupique* | *ostéoporose* | *Infection Pulmonaire* | *Pneumopathie interstitielle diffuse de la sclérodermie* |
|---|---|---|---|---|
| TF_IDF | 0.21 [0.19; 0.23] | 0.21 [0.20; 0.22] | 0.29 [0.28; 0.31] | 0.58 [0.54; 0.61] |
| doc2vec | 0.41 [0.36; 0.45] | 0.37 [0.35; 0.40] | 0.32 [0.29; 0.36] | 0.58 [0.53; 0.63] |
| docBERT 512 first tokens | 0.02 [0.01;0.02] | 0.02 [0.01;0.02] | 0.02 [0.02;0.02] | 0.04 [0.03; 0.04] |
| docBERT 512 tokens random start in the text | 0.14 [0.13; 0.14] | 0.11 [0.10; 0.11] | 0.08 [0.08; 0.08] | 0.10 [0.10; 0.10] |
| Méthode d'expansion de requête | 0.99 [0.98; 1.0] | 0.98 [0.97; 0.99] | 0.99 [0.98; 1.0] | 0.99 [0.98;1.0] |

**Tableau 3** : Comparaison des précisions@100, la précision à 100 correspond au nombre de documents positifs pour le phénotype sur les 100 plus proches par similarité cosinus. La similarité cosinus est calculée à partir des documents du gold standard de 256 documents. Le document est défini comme positif s' il contient au moins un terme-clé du silver standard. Pour la méthode d'expansion de requête, il n'y a pas de distance, les résultats présentés ici sont calculés sur 100 candidats aléatoires. Cette analyse a été réalisée pour les quatre méthodes sur une même base restreinte de 10 000 documents (correspondant au silver standard contenant une liste de mots-clés pré-établie par un clinicien extérieur à l'étude) dont 664 candidats à trouver pour la néphropathie lupique, 731 pour l'ostéoporose, 943 pour l'infection pulmonaire et 886 avec une atteinte pulmonaire interstitielle de la sclérodermie.

Une analyse avec des résultats similaires a été réalisée directement sur le jeu de données gold standard de 256 documents : pour chaque phénotype, les documents étaient pris tour à tour comme des documents-index et la precision@10 étaient définis comme le nombre de documents positifs (au sens de l'annotation) dans les 10 plus proches. Pour le modèle TF-IDF, les precision@10 moyenne étaient de 0.26[0.18; 0.37] sur l'ensemble des quatre phénotypes, de 0.19 [0.12; 26] pour le modèle

doc2vec et 0.10 [0.08; 0.12] pour le modèle docBERT prenant une séquence de 512 tokens aléatoire, et 0.98 [0.96;1.0] pour le modèle d'expansion de requête.

Par ailleurs, en dehors des limites exposées dans l'article ci-dessous, la méthode utilisée est une expansion de requête classique et notamment non comparée à des méthodes plus récentes telles que Splade [104] qui permet une extension de requête parcimonieuse à partir du modèle BERT[4]. Néanmoins l'utilisation du modèle CODER n'avait pas été réalisée par le passé à notre connaissance, elle est certes moins sophistiquée sur le plan de modèle, mais nous paraît plus riche sur le plan de l'utilisation des connaissances médicales. Par ailleurs, cette étape de présélection de documents, destinée à être appliquée sur une vaste base de données devait impérativement être légère d'un point de vue computationnel. Des méthodes gourmandes de représentation vectorielle de document telles que le Longformer[32] n'ont par exemple pas pu être testées.

Contribution de l'auteur : conception et réalisation des analyses présentées, écriture de l'article, co-encadrement d'une étudiante en master 2 (réunion quotidienne) en cursus de double diplôme entre l'école Polytechnique et l'université de Shanghai : Yuhan Xiong. Parts égales de programmation et de testing des différents modèles.

# Document search in large clinical databases: from doc2vec to advanced keyword queries

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

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

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

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

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

*Corresponding Author :

Dr. Christel Gérardin,

Sorbonne Université, Inserm,

Institut Pierre-Louis d'Epidémiologie et de Santé Publique,

Paris, France F75012.

christel.ducroz-gerardin@iplesp.upmc.fr

# Abstract

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

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

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

**Results:** Our experience does not support the vector representation model of clinical notes for searching similar patients. However, searching with an advanced synonym search method can lead to very good results without additional burden for the clinician: we achieved a precision of 0.93 [0.90; 0.96] and a recall of 0.78 [0.71; 0.85] evaluated on the basis of the ICD10 codes of the retrieved patients, in a very reasonable time.



**keywords:** document retrieval, medical informatics, clinical phenotypes

# INTRODUCTION

## Background

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

Automatic retrieval of documents of interest is a major current concern, especially in the biomedical domain, from Pubmed article retrieval [9, 10] to clinical document retrieval [11, 12], and remains challenging in this domain due to the extensive use of acronyms, abbreviations, and complex and ambiguous terms.

The use of vector representation models such as doc2vec [13] and docBERT [14] for document retrieval has been used in the legal and business [15, 16, 17], and biomedical domains [9, 18].

In this study we used the UMLS® (for Unified Medical Language System) whose Metathesaurus englobes more than 30 vocabularies of many different languages including French. All medical concepts in the UMLS have a *unique concept identifier CUI* shared with its synonyms and sometimes abbreviations.

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

## Goal of this study

The objective of the study was to optimize the extraction of the clinical textual documents concerning a specific feature (symptoms or disease) from a large health data warehouse. We compared two

types of methods. The first one, document oriented, is based on the vector representation of documents: all documents having a vector representation close to that of an initial index document (i.e. documents with the same diseases or symptoms mentioned) are selected to be extracted. Three models have been tested: TF-IDF, doc2vec [13] and docBERT [14]. We compared document oriented methods with another approach - namely keyword oriented, based on an advanced keyword query: the clinician enters a diagnosis of interest, and a textual search is performed using these keywords, their synonyms in the UMLS and close terms found by similarity in CODER embeddings [21]. Figure 1 shows an overview of the two methods being compared.

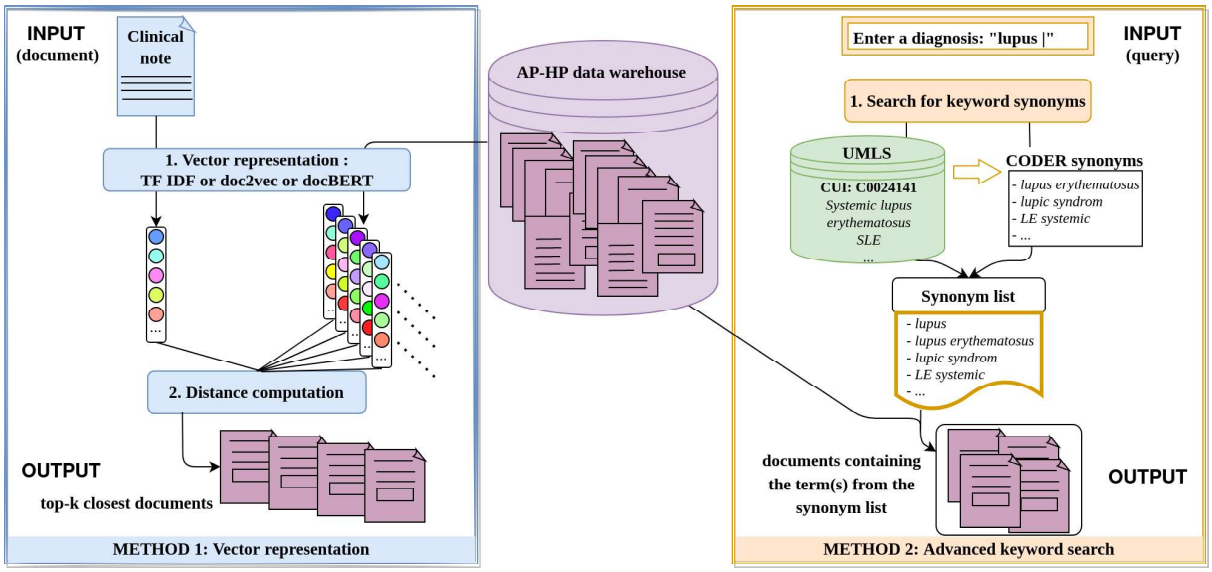



**Figure 1**: Overview of the two methods of searching for documents in our data warehouse. Method 1 is document oriented and method 2 is keyword oriented.

# MATERIAL AND METHODS

## 1. Dataset:

For this experiment, we used approximately 2 million documents associated with all patients aged 15 years or older with lupus, SAPL, scleroderma, and Takayasu disease who were seen at least once since 2017 in one of the 39 hospitals of the Assistance Publique Hôpitaux de Paris (AP-HP).

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

annotated with the corresponding phenotype. Among 256 randomly selected documents, 23 were annotated with "osteoporosis," 48 with "nephritis in SLE," 20 with "interstitial lung disease in scleroderma," and 33 with "pulmonary infection."

## 2. Vector document representation

Three main algorithms were tested to represent the clinical notes and find close documents.

A TF-IDF method, corresponding to a term frequency/inverse document frequency representation. It particularly captures relevant words and is of major interest for our autoimmune disease database. We used the TfIdfVectorizer from sklearn [22].

A doc2vec model [13] was also trained on our global dataset. This numerical representation of a document is based on the word2vec model [23]. In the doc2vec model, a document vector is trained in parallel with the training of the word vectors. Basically, this representation should reflect the "topic" of the document and two documents with the same topic should be close whereas two documents with different topics should be remote. We used the doc2vec model from gensim library [24]. We tried both DBOW and DM models with a dimension of 300.

Similarly, a docBERT model [14] from a camemBERT-large model [25, 26] was also used as a document representation. The BERT model is a transformer [8]: a deep neural network with attention layers that also represents words or paragraphs in their context. We used a model fine-tuned on our database.

## 3. Advanced keywords search:

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

CODER (Cross-lingual knowledge-infused medical term embedding) is an algorithm trained by contrastive learning on UMLS, based on a BERT model allowing term normalization (i.e. linking a term to its specific UMLS concept). The underlying principle is that UMLS synonyms have a close cosine similarity between their respective embeddings. We used the CODER_all version specifically on the French UMLS vocabularies and obtained the synonyms of the terms. An example of a query and its

synonyms is presented in Table 1, which allows us to obtain a significantly larger list and especially to access the acronyms of the terms. Once the list of synonyms was found, we performed a direct search of the original terms and synonyms on the documents.

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

| | | |
|---|---|---|
| **Query** | “Purpura thrombopénique idiopathique” | Idiopathic thrombocytopenic purpura |
| **Same UMLS CUI terms** | PTI,<br>purpura thrombocytopénique auto immun,<br>maladie de werlhof,<br>purpura thrombopénique immunologique,<br>purpura thrombopénique idiopathique,<br>purpura thrombocytopénique autoimmun,<br>purpura thrombopénique autoimmun,<br>purpura thrombopénique idiopathique de werlhof,<br>PTAI,<br>syndrome de werlhof,<br>purpura thrombopénique auto immun,<br>purpura idiopathique,<br>purpura thrombocytopénique idiopathique | ITP,<br>autoimmune thrombocytopenic purpura,<br>werlhof's disease,<br>immunological thrombocytopenic purpura,<br>idiopathic thrombocytopenic purpura,<br>autoimmune thrombocytopenic purpura,<br>autoimmune thrombocytopenic purpura,<br>idiopathic thrombocytopenic purpura of werlhof',<br>AITP,<br>werlhof syndrome,<br>autoimmune thrombocytopenic purpura,<br>idiopathic purpura,<br>idiopathic thrombocytopenic purpura |
| **CODER Synonyms (limit 0.75)** | purpura thrombocytopénique idiopathique,<br>purpura thrombopénique idiopathique,<br>purpura thrombopénique idiopathique de werlhof,<br>maladie de werlhof,<br>purpura thrombopénique à médiation immunitaire, | idiopathic thrombocytopenic purpura,<br>idiopathic thrombocytopenic purpura of werlhof,<br>disease of werlhof,<br>immune-mediated thrombocytopenic purpura,<br>autoimmune thrombocytopenic purpura,<br>idiopathic thrombocytopenia, |

| | |
|---|---|
| purpura thrombocytopénique autoimmun,<br>thrombocytopénie idiopathique,<br>purpura hyperglobulinémique,<br>purpura thrombopénique auto immun ,<br>purpura thrombocytopénique auto immun,<br>purpura idiopathique,<br>syndrome de werlhof,<br>purpura thrombopénique immunologique<br>thrombocytose idiopathique,<br>purpura thrombopénique autoimmun | hyperglobulinemic purpura,<br>autoimmune thrombocytopenic purpura ,<br>autoimmune thrombocytopenic purpura,<br>idiopathic purpura,<br>werlhof syndrome,<br>immunological thrombocytopenic purpura<br>idiopathic thrombocytosis, |

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

# 4. Evaluation

## 4.1. Vector document representation

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

For each phenotype of the 256 documents (Nephritis in SLE, etc...) we had a list of terms manually defined by a clinician as exhaustive as possible of synonyms of this particular phenotype, the list of documents containing these terms corresponding to our gold standard. Supplementary Table 1 summarizes the list of manual terms entered. We then calculate the accuracy of the top-k documents by checking whether we find any of these terms in the text documents. Among the 10,000

documents, the manual list method extracted 664 patients with nephritis in SLE, 731 with osteoporosis, 943 with lung infection, and 886 with ILD in scleroderma.

For the second assessment, we took a random sample of 100 documents from the overall database and considered ICD10 coding as another second gold standard. For these 100 random documents, we calculated whether the k closest documents had the same ICD10 coding for the principal diagnosis (main disease), for at least one associated diagnosis (comorbidities), and for at least one principal or associated diagnosis with the sampled document. We calculated the average accuracy with a 95% confidence interval on the 100 documents in the sample. It is important to note that ICD10 coding is not done at the document level but at the patient level (i.e. the aggregation of the coding of all the documents of the same patient). For example, if a patient comes for an endoscopy, or a specialist consultation, the corresponding report will not mention verbatim all the patient's previous diseases (e.g. "osteoporosis") but the coding will still contain the corresponding code. We therefore also considered a "patient-level" precision, by computing how many patients in the top k candidates had at least one ICD10 coding in common with the index patient. In this experiment, we considered "coarse-grained" ICD10 codings, i.e., one letter, two digits.

### 4.2. Advanced keywords search evaluation:

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

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

| Phenotype | Number of documents with manual list gold standard |
|---|---|
| Nephritis in SLE | 19248 |
| Osteoporosis | 20897 |
| Lung infection | 18355 |
| ILD in scleroderma | 23511 |
| Pulmonary hypertension | 28514 |
| Gastroesophageal reflux | 20648 |
| Gougerot Sjogren syndrome | 16420 |
| Pulmonary embolism | 39523 |

**Table 2**: Number of documents gold standard for the 8 training phenotypes.

For these training phenotypes, a list of corresponding ICD10 codes was also prepared in advance. For instance, pulmonary embolism is "I26". The recall was defined as the number of patients found by our advanced keyword document search method versus the number of patients to be found with the corresponding ICD10 of the phenotype.

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

For those test queries, a clinician assessed a random set of 50 documents, verifying if the document mentioned the disease or not. This additional verification allowed us to evaluate directly the accuracy at document scale. The recall is computed as previously with the ICD10 coding of the patients. For

more complex queries such as "*stroke in lupus*" or "*pericarditis in lupus*", we considered the patients with both "stroke" and "lupus" encodings (i.e. the intersection of the two).

# RESULTS

## 1. Vectorial document representation:

### 1.1. In comparison with the manual list gold standard:

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

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

**Table 3**: Precision results of the top 100 candidates for each phenotype filtered on discharge summary. Restricted database of 10 000 documents with respectively 664 patients for nephritis in SLE, 731 with osteoporosis, 943 with lung infection and 886 with ILD in scleroderma.

### 1.2. In comparison with IDC10 encodings:

As described in the "Method" section, we also evaluated the accuracy of document retrieval for the vector representation of documents with a set of 100 sample documents by comparing the common ICD10 encodings of the k (here 100) closest documents of the entire sample set, as shown in table 4. We checked whether the documents had the same ICD10 encoding for the principal diagnosis, or at least one associated diagnosis, or at least one principal or associated diagnosis. As in the previous experiment, we obtained unsatisfactory results for clinical practice, especially for the docBERT model.

| **Evaluation \ Model** | *At least one ICD10 diagnostic in common* | *Same principal ICD10 diagnostic* | *At least one associated ICD10* | *Patient level (at least one common ICD10)* |
|---|---|---|---|---|
| TF_IDF | 0.55 [0.49; 0.60] | 0.22 [0.16;0.27] | 0.39 [0.34;0.45] | 0.67 [0.61; 0.72] |
| doc2vec | 0.58 [0.52; 0.64] | 0.25 [0.20; 0.31] | 0.39 [0.33; 0.45] | 0.61 [0.55; 0.67] |
| docBERT | 0.20 [0.18; 0.22] | 0.07 [0.05; 0.09] | 0.08 [0.08; 0.12] | 0.27 [0.24; 0.30] |

**Table 4**: ICD10 precision results for the top 100 candidates based on a sample of 100 random documents from a restricted database of 10,000 documents. Here, precision is calculated by counting the number of the 100 closest documents having at least one ICD10 coding in common with the document tested.

Concerning the calculation time it took 9.8 minutes to infer the docBERT representation for 10 000 text documents, 21 seconds for doc2vec and 5.6 minutes for TF/IDF, on a CPU with 30G of RAM.

## 2. Advanced keywords search on the training set

We first conducted an experiment on training phenotypes to determine the best method for advanced keyword search. A summary of the accuracy and recall results is presented in supplementary Table 2. The gold standard for accuracy calculation corresponds to documents extracted *via* a manual synonym list, and recall is calculated at the patient level with the set of ICD10s expected for the corresponding phenotype. Given the results, in order to maximize the number of synonyms and documents without decreasing accuracy, we chose a threshold of 0.75 for the cosine similarity of CODER embeddings and a preprocessing with stemming.

## 3. Advanced keywords search on the test set:

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

| | Query | Precision (on 50 manually-annotated document per query) | Recall (comparison with respective CIM10) - 2 words query | Number of corresponding documents |
|---|---|---|---|---|
| 1 | “Rheumatoid Arthritis” | 0.98 | 0.73 | 15189 |
| 2 | “Takayasu” | 1 | 0.94 | 2459 |
| 3 | “Pericarditis in lupus” | 0.92 | 0.93 | 7490 |
| 4 | “Acute myocardial infarction” | 0.94 | 0.58 - 0.90 | 10583 - 24017 |
| 5 | “APS” | 1.0 | 0.48 - 0.90 | 4406 - 23181 |
| 6 | “Kidney transplantation” | 0.92 | 0.98 | 10716 |
| 7 | “Prostate cancer” | 1.0 | 0.83 | 2971 |
| 8 | “Lung tuberculosis” | 1.0 | 0.55 | 3586 |
| 9 | “Autoimmune hepatitis” | 0.8 | 0.85 | 2797 |
| 10 | “Dermatomyositis” | 1.0 | 0.77 | 3510 |
| 11 | “Idiopathic thrombocytopenic purpura” | 0.98 | 0.81 | 3749 |
| 12 | “Acute kidney injury” | 0.86 | 0.81 | 15775 |
| 13 | “Raynaud syndrome” | 0.98 | 0.98 | 31900 |
| 14 | “Pyelonephritis” | 1.0 | 0.77 | 7475 |
| 15 | “HIV” | 0.90 | 0.98 | 43582 |
| 16 | “Scleroderma” | 1.0 | 0.92 | 24199 |
| 17 | “Diabetes” | 0.96 | 0.96 | 51224 |
| 18 | “Stroke” | 0.64 | 0.63 | 28162 |
| 19 | “Stroke in lupus” | 0.72 | 0.52 | 5650 |

| | | | | |
|---|---|---|---|---|
| 20 | “spontaneous miscarriage” | 0.98 | 0.66 - 0.82 | 9071 - 14329 |
| | **Overall** | **0.93 [0.90; 0.96]** | **0.78 [0.71; 0.85] - 0.83 [0.77; 0.89]** | |

**Table 5:** Precision on a sample of 50 random documents (corresponding to the number of documents with the correct phenotype mention in their text evaluated by a clinician) and recall based on ICD10 coding results for each query, the "two word query" type corresponds to the case where the clinician enters two different terms for the same phenotype.

Some queries, however, had poorer recall results due to the absence of frequent French acronyms in the UMLS. This is the case of “*antiphospholipid syndrome”* which is very frequently mentioned in French only with its acronym “APS” (“SAPL” in French) which is not present as a distinct entity in the UMLS. Similarly, “spontaneous miscarriage” is frequently referred to as “FCS” in French (stands for “Fausse Couche Spontanée”), but the latter is not the corresponding entity in the UMLS. Concerning “acute myocardial infarction” it is now frequently mentioned as “ACS” (“SCA” in French for “Syndrome Coronarien Aigu”) with the same corresponding ICD10 encoding but is not present as the right synonym in the UMLS (since “SCA” stands for “Sudden Cardiac Arrest” in English). The addition of these acronyms in their respective queries has indeed led to a significant improvement in recall: from 0.48 to 0.9 for “APS”, from 0.58 to 0.9 for “acute myocardial infarction”, and from 0.66 to 0.82 for the “spontaneous miscarriage”.

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

For the “stroke” results, an error analysis showed that only 18% of 50 random documents not found were ultimately relevant, with “cerebral vasculitis” or “ cerebral thrombophlebitis” mentioned. Supplementary Table 3 shows the error analysis results.

It took 10 minutes to extract the corresponding documents for all the 20 queries.

# DISCUSSION

In this study, we show that the classical document representations by TF-IDF, doc2vec [13] and docBERT [14] are not efficient in retrieving patients with specific written clinical phenotypes, or the same ICD10 coding, even at a parent ICD10 level (one letter, two digits). This can be explained by the fact that these methods extract a vector summarizing the document, but similarity between patients can appear along many different dimensions. Distance metrics such as cosine are applied uniformly over the vectors and do not make the dimensions of interest explicit. The keyword search method makes this dimension of interest explicit, which explains the better performance.

Nevertheless, we propose a new interesting method, hybrid between the symbolic method (directly based on metathesaurus) and the deep learning method, corresponding to an advanced keyword search. We achieved very good precision and recall based on IDC10 coding comparison. Our method is fully replicable and all our code is available on github at https://github.com/ChristelDG/EHR_2_vec.

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

One of the main advantages of our method is that it takes clinical time into account and enables to rapidly look for patients with a single term or acronym. The use of CODER synonyms allows one to look for terms even if the input diagnosis is not directly in the UMLS or if it is misspelled and is one of the major contributions of our study. Furthermore we proposed a method, evaluated at document and patient level by several gold standards with clinical control of documents and error analysis.

One of the limitations of our study is that none of the proposed evaluation metrics is perfectly satisfactory, indeed, a manual term list may miss a part of the documents with misspelled terms, or other unreflected synonyms, so it cannot be considered for the recall calculation.
Moreover, as shown by the error analysis and as already mentioned several times in the literature [29, 30, 33], the IDC10 codings are not accurate enough. This can be explained by the fact that ICD10 coding is done to financially value patients' stays (and in particular the coding "pulmonary tuberculosis" corresponds globally to the same valuation as lymph node or meningeal tuberculosis).

However, these different metrics, taken as a whole and coupled with the analysis of errors, give a good idea of the performance and the problems encountered by users.

Another limitation is that our method is still sensitive to acronyms or abbreviations that are not all present in the UMLS. As an illustration of this problem, we show that adding no more than three usual French acronyms absent from the UMLS enabled to improve the overall recall from 0.78 to 0.83 [0.77; 0.89].

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

# CONCLUSION

We tested three state-of-the-art vector document representations to perform extraction of documents of interest from a large database. We show that none of these three methods was sufficiently effective for this task in the context of autoimmune patient hospitalization reports. We propose a new advanced keyword search method with automatic synonym search with very good accuracy and recall performance.

## ETHICS

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

## ACKNOWLEDGMENT

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

## AUTHOR CONTRIBUTIONS

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

## COMPETING INTERESTS

The authors declare no competing interests.

## DATA AVAILABILITY

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

## ABBREVIATIONS

AP-HP: assistance publique hôpitaux de Paris

APS: Antiphospholipid syndrome

CNIL: Commission Nationale de l'Informatique et des Libertés

EHR: Electronic health records

ILD: Interstitial Lung Disease

NER: named-entity-recognition

NLP: natural language processing

SLE: systemic lupus erythematosus

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Supplementary Table 1**: Corresponding queries for comparison with document extraction

| *Phenotypes* | *Manual list* | *Additional list* |
|---|---|---|
| Nephritis in SLE | ['lupus nephrotopathy', 'lupus glomerulonephritis', 'lupus with renal involvement', 'lupus renal involvement', 'renal failure secondary to lupus', 'lupus glomerulopathy', 'lupus gn', 'class IV renal involvement', 'class V renal involvement', 'class III renal involvement', 'class VI renal involvement', 'extra membranous glomerulonephritis class V'] | ['glomerulonephritis', 'chronic renal failure', 'chronic renal disease', 'GEM', 'HSF', 'segmental and focal hyalinosis', 'renal damage']<br><br>AND<br><br>['lupus'] |
| osteoporosis | ['osteoporosis', 'osteoporotic'] | |
| Lung infection | ['inhalation pneumopathy', 'pneumopathy to', 'legionellosis', 'pulmonary infection', 'infectious pneumopathy', 'mechanically ventilated pneumopathy', 'MVP', 'pneumonia', 'bilateral pneumopathy', 'basal pneumopathy', 'bi-basal pneumopathy', 'basal lobar pneumopathy', 'ALFP', 'acute frank lobar pneumopathy', community-acquired lung disease', 'acute lung disease', 'documented lung disease', 'ventilator-associated lung disease', 'pulmonary-acquired sepsis', 'pulmonary-acquired septic shock', 'upper lobar lung disease', 'necrotizing lung disease', 'bronchopneumonia', 'bronchopneumonia'] | |
| ILS in Scc | ['interstitial lung disease', 'interstitial lung disease', 'interstitial syndrome', 'lung damage', 'IPD', 'PINS', 'pulmonary fibrosis', 'interstitial fibrosis', 'IDF', 'interstitial damage', 'fibrosing | |

| *Phenotypes* | *Manual list* | *Additional list* |
| --- | --- | --- |
| | lung disease']<br>AND<br>['systemic scleroderma', 'Scc', 'diffuse cutaneous scleroderma', 'limited cutaneous scleroderma', 'CREST syndrome','CREST'] | |

**Supplementary Table 2**: Accuracy and recall results for each phenotype.

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

**Supplementary Table 3** : Error analysis on the three phenotypes with less performances : Analysis is performed by a clinician on a set of 50 random documents for each phenotype

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

## 4.4. Reconnaissance d'entité nommée et traduction (article 3 *en review*)

Résumé :

Une fois la cohorte présélectionnée, les concepts médicaux sont extraits de tous les documents textuels sélectionnés.

L'étude présentée ici par du constat que d'une part les performances des modèles de langage en anglais biomédicaux tel que le clinicalBERT [27] sont les meilleurs de la littérature, que d'autre part, les grandes bases de connaissances médicales telles que l'UMLS[16] possèdent beaucoup plus de concepts en anglais (environ 5 fois plus en 2020), et que par ailleurs les modèles de traduction tels que Google translate ou open-mt disponible dans la librairie Huggingface [90], (une plateforme en ligne qui héberge un très grand nombre de modèles transformers dans de très nombreuses langues).

L'expérimentation porte donc sur la comparaison de deux méthodes : L'une en langue anglaise qui comporte les étapes suivantes : traduction du texte clinique français (à partir d'un modèle opus-mt affiné sur des textes biomédicaux), puis extraction d'entités nommées et normalisation (avec un modèle fusionnant les deux étapes : MedCAT [47] et un modèle en deux temps. L'autre méthode, en langue française, réalise directement l'étape d'extraction d'entité nommée et de normalisation en français.

Pour les deux expérimentations anglaise et française en deux temps : reconnaissance d'entité nommée et normalisation, la même architecture de reconnaissance d'entités nommées, (présentée section 2.2.1.) a été utilisée avec deux modèles de langage différents : le clinical BERT pour la version traduite en anglais et le camemBERT [24], affiné sur l'espace projet. Deux modèles de normalisation (présentés sections 2.3.) ont également été comparé : le CODER [60] et le modèle multilingue développé par Wajsburt et al.[56].

Les golds standards utilisés pour pouvoir comparer les performances sont des textes annotés manuellement avec l'annotation du type sémantique et des CUI de l'UMLS. Deux jeux ont été testés avec ces annotations : le jeu de données disponible publiquement QUAERO [91] et un jeu de notes cliniques de notre espace annoté manuellement.

Les résultats montrent que le fait de passer par une étape de traduction induit une perte d'information trop importante, avec une perte de 12 points pour la f1 score en moyenne.

# Impact of translation on biomedical information extraction from real-life clinical notes

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

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

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

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

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

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

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

**Abstract**

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

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

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

In conclusion, despite the recent improvement of the translation models, there is a signifi- cant performance difference between the two approaches in favor of the native French method which is more efficient on French medical texts, even with few annotated documents.

*Keywords:* Concept Normalization, Named Entity Recognition, Natural Language Processing, Translation

## 1. Introduction

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

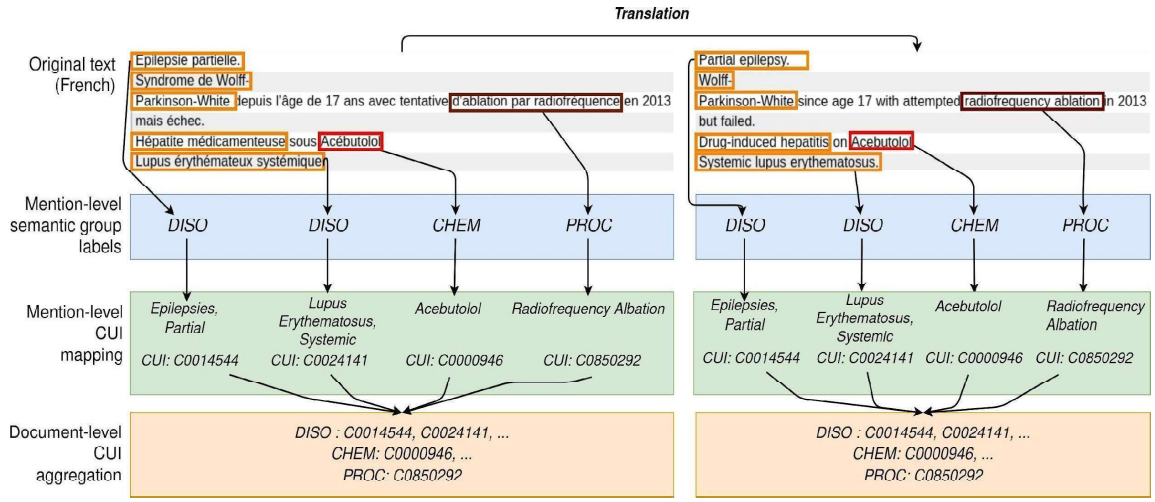



**Figure 1: Overall method objective: from raw text to CUI information document by document.** The "Mention level" denotes the analysis done at the word or group of words level: first, the NER step (in blue), then the normalization (in green), finally all mentions with normalized CUIs are aggregated at the document level (orange part). Both native French and translated English approaches are compared.

In many languages other than English, efforts still need to be made to obtain such interesting results, in particular due to a much smaller amount of accessible data [4].

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

We evaluate and discuss the results on several French biomedical corpora, including a new set of 42 hospitalization reports annotated with 4 entity groups. We evaluate the normalization task at the document-level, in order to avoid a cross-lingual alignment step at evaluation time, which would add a potential level of error and thus would make the results more difficult to interpret (see word alignment in [6, 7]). This normalization is performed by matching all the terms to their *Concept Unique Identifier* (CUI) in the UMLS® [1]. Figure 1 summarizes these different steps, from the raw French text and the translated English text to CUI aggregation and comparison at document-level. All our codes are available on github [8].

## 2. Background

The different steps of our algorithms rely heavily on Transformers language models [2]. These models are currently the state of the art for many natural language processing (NLP) tasks, such as machine translation, named entity recognition, classification, and text normalization (also known as "entity binding"). Once trained, these models can represent any specific language, such as biomedical language or legal language. The power of these models comes from their neural architecture but also depends largely on the amount of data on which they are trained. In the biomedical domain, two main types of data are available: public articles (e.g. PubMed ) and clinical electronic medical records databases (e.g. MIMIC III [14]), and the most powerful models are for example BioBERT [15] which has been trained on the whole of PubMed in English, and ClinicalBERT [16] trained on PubMed

and MIMIC III. In French language, the variety of models is less important, with the models CamemBERT [17] and FlauBERT [18] for the general domain, and no specific model available for the biomedical domain.

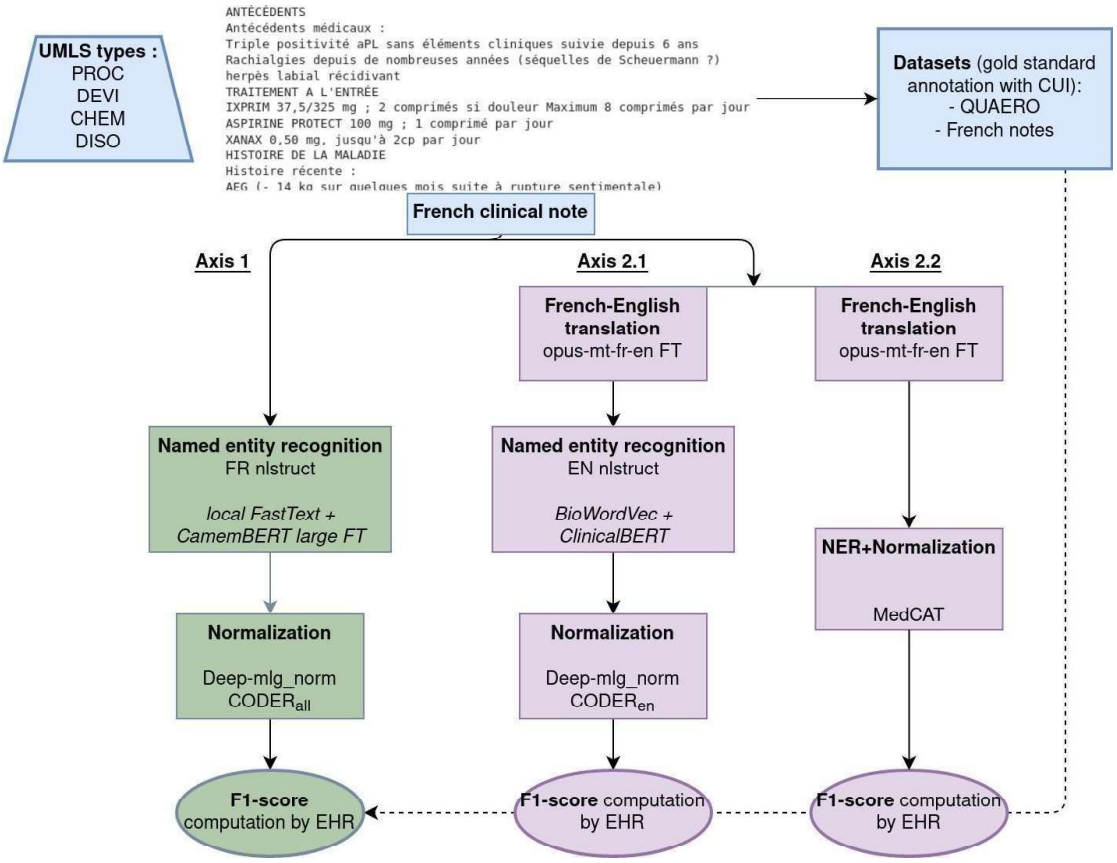



**Figure 2: Diagram of the different experiments comparing French and English language models without and with intermediate translation steps.** The Axis 1 (green axis on the left) corresponds to the native French branch with a NER step based on a FastText model trained from scratch on French clinical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

NER and normalization steps: MedCAT [13].

In addition to the particularly powerful English pre-trained models, universal biomedical terminologies (i.e., metathesaurus) also contain significantly more English terms than other languages. For example, the Unified Medical Language System (UMLS®[1]) contains at least ten times more English terms than French terms, which can enable ruled-based models to perform better in English. As mentioned above, each concept of reference in the UMLS®[1] is assigned a Concept Unique Identifier (CUI), associated with a set of synonyms eventually in several languages, and a semantic group -such as *Disorders, Chemical and Drugs, Procedure, Anatomy* and so on.

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

## 3. Materials and Methods

### *3.1. Overview*

Figure 2 presents the main approaches and models used in our study. We explored a “native French approach axis” (axis 1 in Figure 2) based on French language models learned on and applied to French annotated data; and two “translated English approach axes” (axes 2.1 and 2.2) based on a translation step and English concept extraction tools. We compare the performance of all axes with an average of the CUI predictions accuracies at document level, for all documents.

#### *3.1.1. Native French approach*

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

### *3.1.2. Translated-English approach*

Axes 2.1 and 2.2 first consist in a translation step, presented in section 3.3 below, operated by the state-of-the-art *opus-mt-fr-en* algorithm [11] or Google Translate®. Then, like axis 1, axis 2.1 is based on a NER and a Normalization step. The NER step is done by the same algorithm but trained on the n2c2 2019 dataset[24], for the normalization step with used the same deep multilingual algorithm[9] (section 3.5.1) and the English version of CODER [10] based on a BioBERT[15] model. This axis allows to compare two methods whose difference is only the translation step.

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

## *3.2. Datasets*

### *3.2.1. Overview*

For all our experiments, we chose to focus on only four UMLS®[1] group: *Chemical & Drugs* (CHEM) and *Devices* (DEVI) corresponding to medical devices such as pace-maker, catheter, etc., *Disorders* (DISO) corresponding to all signs, symptoms, findings (for instance positive or negative biological test results) and diseases, *Procedures* (PROC) correspond- ing to all diagnostic and therapeutic procedures such as imaging, biological tests, operative procedures, etc.

Table 1 presents the datasets used for all our experiments with the corresponding numbers of documents. First, two French datasets were used for the final evaluation, as well as for training the axis-1 models. QUAERO is a freely available corpus [25] based on pharmacological notes with two sub-corpora: Medline (PubMed abstract short sentences) and EMEA (drug notices). We also

annotated a new real-life clinical notes dataset from the Assistance-Publique Hôpitaux de Paris datawarehouse, described in Supplementary materials Section 1.1.1.

Second, we used the corpus n2c2 2019 [24] with annotated CUIs – on which we auto- matically added the UMLS®[1] semantic group information, to train the axis-2.1 system and to evaluate the English NER and normalization algorithms. We also used the Mantra dataset [26], corresponding to a multilingual gold standard corpus for biomedical concept recognition.

Finally, we fine-tuned and tested the translation algorithms on both 2016 [27] and 2019 [28] WMT biomedical corpora. Detailed description of the number of respective entities in the datasets can be found in the supplementary Table 1.

Table 1: **Datasets**. Presentation of all datasets used.

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

French corpus annotation methods are detailed in section 1.1.1 of supplementary materials with supplementary Figure 1. Entities repartition for this annotation is detailed in Supplementary Table 1.

### *3.3. Translation*

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

### *3.4. Named Entity Recognition*

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

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

### *3.5. Normalization algorithms*

This step of our experiments is essential in order to compare a native French and a translated English method and consists in mapping each mention extracted from the text to its associated CUI in the UMLS®[1]. We compare three models for this step, described below: the deep multilingual normalization algorithm developed by [9], the CODER[10] and the MedCAT model[13] that performs both NER and normalization at the same time.

All these three models do not need any training dataset other than the UMLS®.

#### *3.5.1. Deep multilingual normalization*

This algorithm from Wajsbürt et al. [9] considers the normalization task as a highly- multiclass classification problem with a cosine similarity and a softmax function as a last layer. The model is based on contextualized embedding, using the pre-trained multilingual BERT model [3] and works in two steps: during the first step, the BERT model is *fine-tuned* and French UMLS terms and their corresponding

English synonyms are learned. Then, in the second step, the BERT model is frozen and the representation of all English-only terms (i.e. only present in English in the UMLS®[1]) is learned. The same training is used for the native French and translated English approach. This model was trained with the 2021 UMLS®[1] version, corresponding to the version used for the annotation of the French corpus. This model was thus trained on more than 4 million concepts corresponding to 2 million CUIs.

### *3.5.2. CODER*

The CODER algorithm [10] is developed through contrastive learning based on the UMLS®[1] medical knowledge graph, concept similarities are computing from terms representation and relation of this knowledge graph. The contrastive learning is used to learn embeddings through a Multi-Similarity loss [37]. The authors have developed two versions: a multilingual based on multilingual BERT [3] and an English one based on BioBERT[15] pre-trained model. We used the multilingual version for axis 1 (native French approach), and the English version for axis 2.1. The two types of this model (CODER all and CODER en) were trained with the 2020 UMLS version (publicly available models). The CODER all [10] was thus trained on more than 4 million concepts corresponding to 2 million CUIs and the CODER en was trained on more than 3 millions terms and 2 millions CUI.

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

The MedCAT algorithm is detailed in Section 1.1.1 (supplementary materials).

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

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

## 4. Results

The sections below present the performance results for each step. The n2c2 2019 challenge corpus [24] allowed us to evaluate the performance of our English models on clinical data and the Biomedical Translation shared task 2016 [27] to evaluate our translation performance on biomedical data with a BLEU score [38].

### *4.1. NER performances*

To be able to compare our native French and translated English approaches, we used the same NER model (section 3.4), trained and tested on each respective datasets described above (section 3.2). Table 2 presents the corresponding results. The overall F1-scores are similar from one dataset to another: from 0.72 to 0.77.

### *4.2. Normalization performance*

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

Table 3: **Normalization performance.** Presentation of the accuracy results of the Normalization models computed from the annotated datasets, focusing on the four semantic groups of interest : CHEM, DEVI, DISO, PROC.

| Dataset<br>Models | EMEA test | French notes | n2c2 2019 test |
|---|---|---|---|
| deep mlg norm | **0.65** | **0.57** | **0.74** |
| CODER all CODER | 0.58 | 0.51 | – |
| en | – | – | 0.63 |

*4.3. Translation performances*

For the two translation models, the respective BLEU scores [38] are computed on the 2016 Biomedical Translation shared task [27]. A fine-tuned version of opus-mt-fr-en [11] on the 2016 and 2019 Biomedical Translation shared tasks was also tested. However, the Google translate model could not be used for our experiments involving clinical notes due to confidentiality reasons.

Table 4 shows the BLEU score results for the three models, showing that fine-tuning on the opus-mt-fr-en model [11] on biomedical datasets led to the best results, with a BLEU score[38] of 0.51. We will use this model for the overall performance of axes 2.1 and 2.2.

### *4.4. Overall performances from raw text to CUI predictions*

This section presents the overall performance of the 3 axes, in an end-to-end pipeline. For axis 2, the results are those obtained with the best normalization algorithm (presented in Table 3). The model used for translation was the opus-mt-fr-en [11] fine-tuned model. The results are presented in Table 5, the best results are obtained by the native French approach on the EMEA corpus [25] and the French clinical notes. The 95% confidence intervals were calculated using the empirical bootstrap method [39].

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

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

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

| | | EMEA test | | | French notes | | |
|---|---|---|---|---|---|---|---|
| | | precision | recall | f1-score | precision | recall | f1-score |
| Methods | Axis 1 (French NER+normalization) | 0.63 | 0.60 | **0.61 [0.53;0.65]** | 0.49 | 0.53 | **0.51 [0.47;0.55]** |
| | Axis 2.1 (Translation+NER+normalization) | 0.53 | 0.40 | 0.45 [0.38;0.51] | 0.41 | 0.38 | 0.39 [0.34;0.44] |
| | Axis 2.2 (Translation+MedCAT[13]) | 0.53 | 0.46 | 0.49 [0.38;0.54] | 0.38 | 0.38 | 0.38 [0.36;0.40] |

## 5. Discussion

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

### *5.1.1. The quality of the translation is not sufficient*

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

relatively low, although improved by a fine-tuning step.

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

Commercial API-based translation services could not be used for our task due to data privacy issues. However, the opus-mt model is considered state of the art, it is adjustable on domain specific data, and the translation results presented in Table 4 confirm the lack of performance difference between this model and the google translate model.

Even if our experiments were performed on only one language, the French-English pair is one of the best performing in recent translation benchmarks[43]. It is unlikely that other languages would lead to significantly better results.

*5.1.2. Error Analysis*

In these experiments, the overall results may appear low, but the task is still complex, especially because the UMLS® [1] contains many synonyms with different CUIs. To better understand, we performed an error analysis on the normalization task only, as shown in Supplementary Table 3, with a physician's evaluation, on a sample of 100 errors for both models. We calculated that 24% and 39% of the terms found by the deep normalization algorithm [9] and CODER [10] respectively were actually synonyms but with two different UMLS CUIs. For example, cardiac ultrasound has CUI C1655737 while echocardiography has another CUI C0013516, similarly H/O: thromboembolism has a CUI of C0455533 while history of thromboembolism has a CUI of C1997787 and so on. In addition, as shown in Supplementary Table 3, abbreviations and misspelled words also induce many errors and are difficult to manage, even though some abbreviations are already built into UMLS. Another limitation comes from the ever-changing versions of the UMLS®. In any case, it is the relative differences between the results that matter for our purposes, not the absolute values.

*5.1.3. Limitations*

This work has several limitations, first of all, the real-life French clinical notes had very few terms attached to the "Devices" semantic group, thus preventing the NER algorithm from finding them in the test dataset. However, this drawback, penalizing the native French approach, still allows us to conclude on the results. Moreover, in this study, we did not take into account the attributes of the extracted terms such as the negation, the hypothetical attribute or the belonging to another person than the patient, this for comparison purposes, indeed the datasets QUAERO [25] and n2c2 2019 [24] did not have this information labeled.

**Ethics**

The study and its experimental protocol was approved by the AP-HP Scientific and Ethical Committee (IRB00011591 decision number CSE 20-0093). Patients were informed that their EHR information could be reused after an anonymization process and those who objected to the reuse of their data were excluded. All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the CNIL: Commission Nationale de l'Informatique et des Libertés [44]).

**Data availability**

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

**Acknowledgments**

The authors would like to thank the AP-HP data warehouse, which provided the

data and the computing power to carry out this study under good conditions. We wish to thank all the medical colleges, including internal medicine, rheumatology, dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency and intensive care units, that gave their agreements for the use of the clinical data.

**Competing interest**

Authors declare no competing interest

**Consent for publication**

Not applicable

**Funding**

Not applicable

**Authors contribution**

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

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

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

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

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

**References**

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

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

(2017).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv, arXiv preprint arXiv:1810.04805 (2019).

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

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

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

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

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

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

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

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

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

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

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

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical

language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

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

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

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

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

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

[21] N. Perez, M. Cuadros, G. Rigau, Biomedical term normalization of EHRS with UMLS, arXiv preprint arXiv:1802.02870 (2018).

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

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

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

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

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

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

[28] R. Bawden, K. Bretonnel Cohen, C. Grozea, A. Jimeno Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, A. Siu, K. Verspoor, M. Vi- cente Navarro, Findings of the WMT 2019 biomedical translation shared task: Eval- uation for MEDLINE abstracts and biomedical terminologies, in: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Association for Computational Linguistics, Florence, Italy, 2019, pp. 29–53. doi:10.18653/v1/W19-5403.

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

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

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

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

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

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

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

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

[37] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.

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

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

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

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

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

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

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

## 4.5. Classification des concepts médicaux (article 4 publié à *AIIM*)

Résumé :

Afin d'améliorer l'interprétabilité clinique ultérieure et d'analyser les patients selon plusieurs dimensions médicales, tous les symptômes et pathologies extraits sont classés selon une classification multi étiquette.

Cet algorithme de classification est présenté dans l'article ci-dessous. Cette classification multi-étiquette des concepts médicaux a été inspiré du Défi Fouille de texte de 2021(DEFT) [92] auxquels nous avons participé, qui proposait de classer les notes cliniques des patients en fonction des spécialités de pathologie mentionnées positivement (c'est-à-dire ayant vraiment eu lieu) dans le texte. Par exemple le texte « Il s'agit d'une femme de 29 ans présentant des vomissements gravidiques sévères avec troubles hydro-électrolytique » devait être étiquetté : *digestif* (pour vomissement), *gynécologique* (pour gravidique) et *nutritionnel* (pour trouble hydro-électrolytique).

Les classes des textes proposées sont celles des têtes de chapitre de la branche C du MeSH [14] , correspondant à 22 domaines médicaux : infections, ophtalmologie, immunologie, dermatologie etc... La méthode proposée pour classer les textes ici correspond à une étape d'extraction d'entité nommée, avec le même modèle que présenté section 2.2.1.2. Puis une étape de classification des termes.

La classification des termes est réalisée par un modèle BERT [4] pour la classification des séquences (BERTforSequenceClassification) a été utilisé et entraîné sur un jeu de données hybride issus d'une part de l'ensemble des termes du MeSH-C [14] et de leurs synonymes dans l'UMLS[16] (partageant le même CUI) et d'autre part d'entités directement issus des notes cliniques. Obtenant ainsi un jeu de données de 42 000 couples termes/étiquettes. Une version augmentée à partir des synonymes anglais et français, correspondant à plus de 300 000 termes avec un BERT[4] multinlingue a permit d'obtenir le meilleur score F1 à 0.811 avec un score F1 à 0.809 pour la version française, ces résultats étaient très proches des gagnants du DEFT (F1 score à 0.814) qui avaient établis un grand nombre de règles manuelles [93]. Par exemple, un nouveau lexique a été créé pour l'ensemble des marqueurs de négation et d'incertitude à partir du corpus du DEFT, qui paraît peu généralisable [93].

Nous avons par la suite également évalué ce classifieur sur notre propre jeu de données annoté, avec les mêmes performances. Pour l'algorithme final de création de cohortes de patients similaires, nous avons retenu la version française.

Artificial Intelligence In Medicine 128 (2022) 102311

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

# Multilabel classification of medical concepts for patient clinical profile identification

Christel Gérardin [a,b,*], Perceval Wajsbürt [c], Pascal Vaillant [d], Ali Bellamine [b], Fabrice Carrat [a,e], Xavier Tannier [c]

[a] Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Inserm, 27 rue Chaligny, 75012 Paris, France
[b] Département de médecine interne, APHP. Sorbonne Université, France
[c] Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé (LIMICS), 75006 Paris, France
[d] Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en eSanté (LIMICS), Sorbonne Université, Inserm, F-93000 Bobigny, France
[e] Public Health Department, Hôpital St-Antoine, APHP. Sorbonne-Université, Paris, France

ARTICLE INFO

*Keywords:*
Biomedical concepts
Multilabel classification
NER
Transformers
Multilingual NLP

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.

## 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 [2–4]. 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),

* Corresponding author at: IPLESP, 27 rue de Chaligny, 75012 Paris, France.
*E-mail address:* christel.ducroz-gerardin@iplesp.upmc.fr (C. Gérardin).
[1] https://www.nlm.nih.gov/mesh/meshhome.html

https://doi.org/10.1016/j.artmed.2022.102311
Received 9 September 2021; Received in revised form 21 April 2022; Accepted 22 April 2022
Available online 26 April 2022

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 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).

Fig. 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.

Fig. 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 Fig. 3).

### 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

[2] http://mesh.inserm.fr/FrenchMesh/

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

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

| 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 |

[a] 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.

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 Fig. 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.

## 3. Methods

### 3.1. System overview

Fig. 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 Fig. 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).

**Table 2**
DEFT 2021 corpus statistics.

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

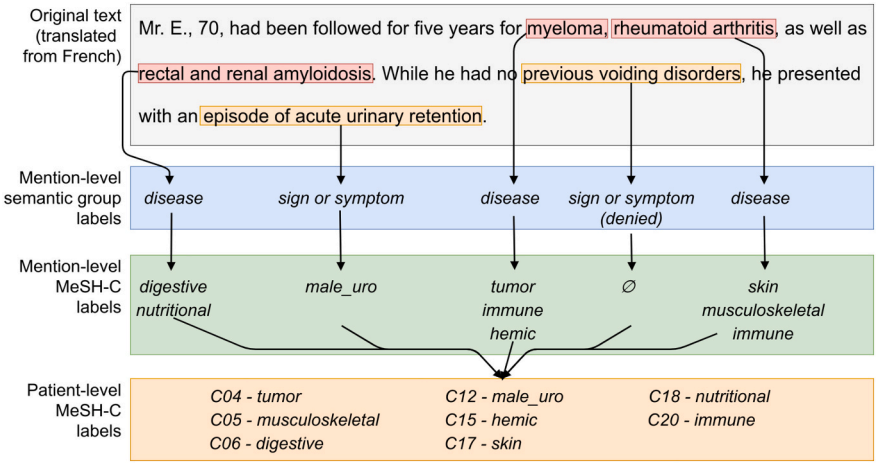



**Fig. 1.** Annotations provided in the DEFT 2021 corpus. For each medical concept of interest (highlighted), there is an entity label "disease", "sign or symptom" and the negation/hypothesis/link to someone else attribute. Each of the positive entities can be mapped to several MeSH-C chapter headings (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.

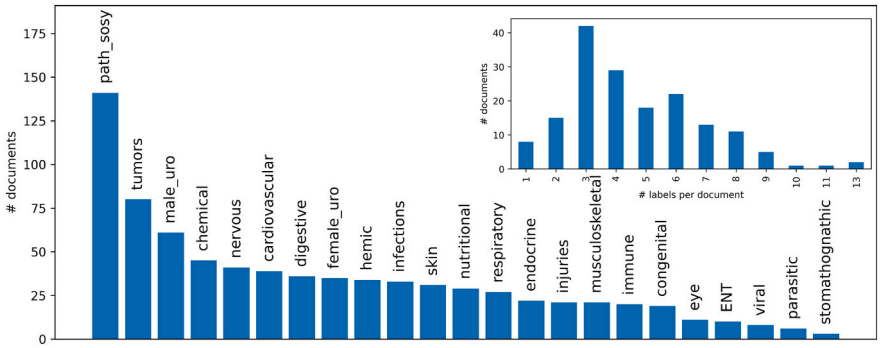



**Fig. 2.** Distribution of labels in the DEFT training dataset. The y-axis represents the number of documents, and all labels presented are listed in Table 1. The thumbnail represents the number of labels per document.

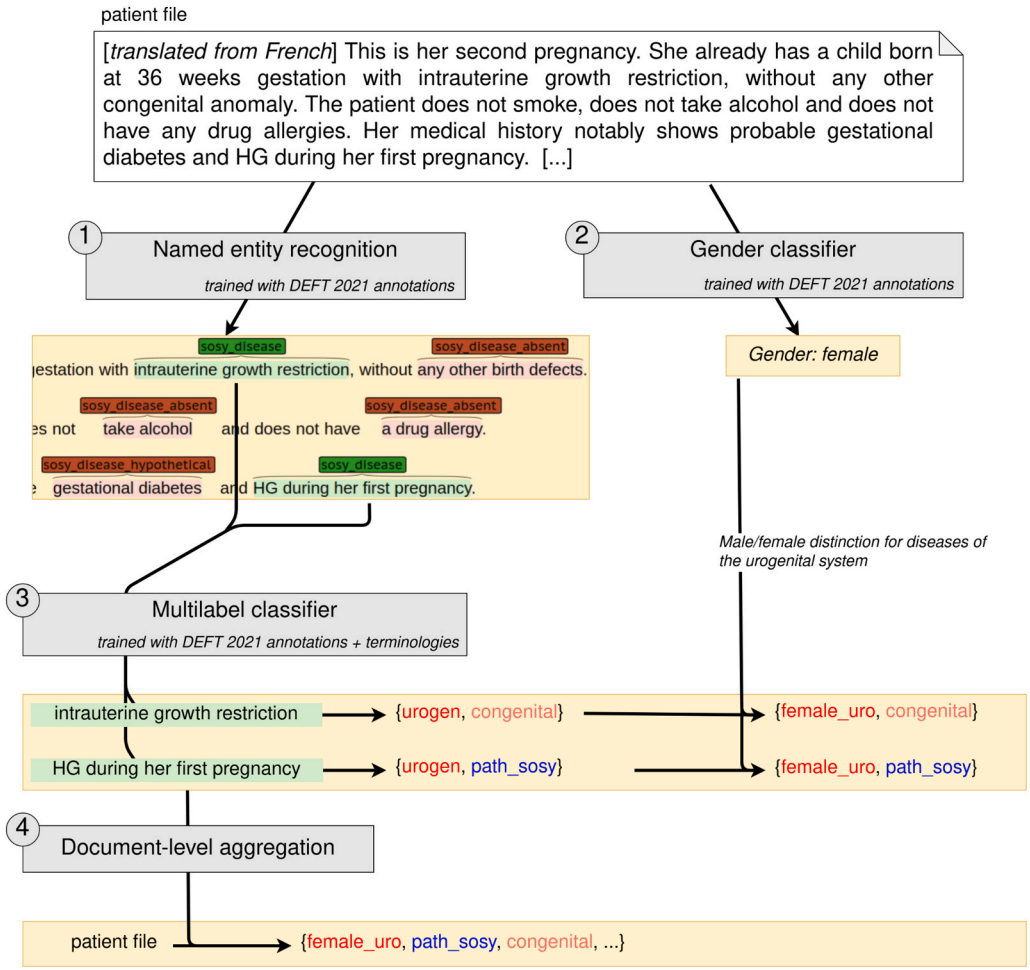



**Fig. 3.** General system architecture. First, named entity recognition is performed, trained on the annotated DEFT dataset to extract positive medical concepts (1). In parallel, the gender classifier, also trained on the DEFT dataset, determines the written gender of the patient (2). Then, a multilabel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

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 Fig. 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

**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.

| 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 |

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.

### 3.2. Named entity recognition (NER)

The named entity recognition model is illustrated in Fig. 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.

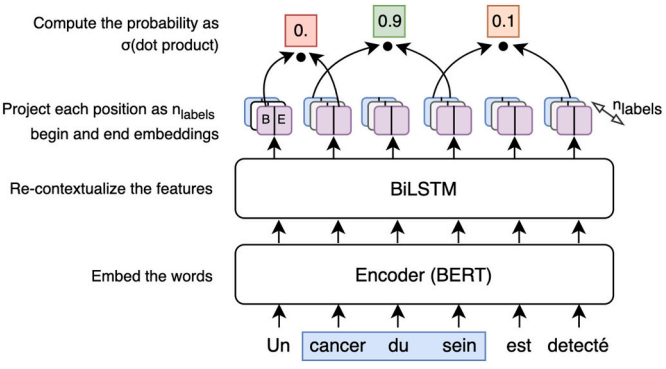



**Fig. 4.** NER system architecture. Each word is projected into $n_{labels}$ to begin representations and $n_{labels}$ to end representations. Finally, each triplet (B, E, L) is scored as a dot product between the begin representation of label L at position B and the end representation of label L at 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[4] 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[5]). 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.

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

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

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

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 Fig. 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 Fig. 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].[6]

## 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).

**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. The value in bold corresponds to our best result.

| | 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 | |

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).

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 9363 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

[6] 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.

though the number of training epochs required is higher for smaller term sets.

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 Fig. 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.

## 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.

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

| | 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 |

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],[7] 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 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 4000 French open access biomedical articles on EuropePMC,[8] 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 4000 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

[7] 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.

[8] https://europepmc.org/

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 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.

## Appendix 1. UMLS vocabularies used for the training set[9]

| 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 |
| 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 |

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

## 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 |

## References


[1] Grouin C, Grabar N, Illouz G. 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]. In: Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT); 2021. p. 1–13.

[2] Pennington J, Manning CD, Socher R. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–43.

[3] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv 2018;1802. preprint arXiv:1802.05365.

[4] Delvin J, Changm W, Toutanovak Leek. BERT : pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 conference of the North American Chapter of the Association for Computational Linguistics : human languagetechnologies. Proceedings of the Conference 1; 2019. p. 4171–86.

[5] Zhang D, Thadajarassiri J, Sen C, Rundensteiner E. Time-aware transformer-based network for clinical notes series prediction. In: Machine learning for healthcare conference; 2020. p. 566–88.

[6] Soni S, Roberts K. Patient cohort retrieval using transformer language models. In: AMIA annual symposium proceedings. 1150; 2020.

[7] Feng J, Shaib C, Rudzicz F. Explainable clinical decision support from text. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP); 2020. p. 1478–89.

[8] Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. arXiv 2019;1906. preprint arXiv:1906.00346.

[9] Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, Gao Y, Huang Y. Clinical trial cohort selection based on multi-level rule-based natural language processing system. J Am Med Inform Assoc 2019;11:1218–26.

[10] Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, Hirschman L, Ball R, Eval ADE. An evaluation of text processing systems for Adverse Event Extraction from drug labels for pharmacovigilance. Drug Saf 2021;1:83–94.

[11] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than english: opportunities and challenges. J Biomed Semant 2018;9(1):1–13.

[12] Hiot N, Minard AL, Badin F. DOING@DEFT : utilisation de lexiques pour Une classification efficace de cas cliniques. In: Traitement Automatique des Langues Naturelles ATALA; 2021. p. 41–53.

[13] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;1(Database issue):267–70. 32.

[14] Lamy JB, Venot A, Duclos C. PyMedTermino: an open-source generic API for advanced terminology services. Stud Health Technol Inform 2015;210:924–8.

[15] Tiedemann J, Thottingal S. OPUS-MT–building open translation services for the world. In: Proceedings of the 22nd annual conference of the european association for machine translation; 2020. p. 479–80.

[16] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rush AM. In: Transformers : state-of-the-art natural language processing. Online: Association for Computational Linguistics; 2020. p. 38–45.

[17] Hochreiter S, Schmidhuberj J. Long short-term memory. Neural Comput 1997;9(8): 1735–80.

[18] Yu J, Bohnet B, Poesio M. Named entity recognition as dependency parsing. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 470–6.

[19] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics : human languagetechnologies. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016. p. 260–70.

[20] Jaeyoung K, Mostafa EK, Jungwon L. Residual LSTM: design of a deep recurrent architecture for distant speech recognition. arXiv 2017. preprint arXiv: 1701.03360.

[21] Kingma DP, Ba J, Bengio Y, LeCun Y. Adam: a method for stochastic optimization. In: 3rd international conference on learning representations. ICLR; 2015.

[22] Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD, Stanza. A python natural language processing toolkit for many human languages. In: Proceedings of the 58th annual meetingof the Association for Computational Linguistics : system demonstrations; 2020.

[23] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I. Attention is all you need. Advances in neural information processing systems 2017: 5998–6008.

[25] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, De La Clergerie E, Seddah D, Sagot B. In: CamemBERT : a tasty french language model. Online: Association for Computational Linguistics; 2020. p. 7203–19.

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

[27] Anthony LFWolff, Kanding B, Selvan R. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. In: CML workshop on challenges in deploying and monitoring machine learning systems; 2020.

[28] Hamidi F, Scheuerman MK, Branham SM. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In: Proceedings of the 2018 CHI conference on human factors in computing systems. 8. Association for Computing Machinery; 2018. p. 1–13.

[29] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.

[30] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly Available Clinical BERT Embeddings. arXiv preprint; 2019. arXiv: 1904.03323.

[31] Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6 (1):1–10.

[32] Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, Celi LA. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PloS one 2018;13(2):e0192360.

[33] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

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

[35] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the North American

Chapter of the Association for Computational Linguistics: human language technologies; 2013. p. 746–51.
[36] Weng WH, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017;17(1):1–13.
[37] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information, transactions of the association for. Comput Linguist 2017;5:135–46.
[38] Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in french using multilingual terminologies and contextual embeddings. J Biomed Inform 2021; 114:103684.

## 4.6. Construction de cohorte de patients similaires (article 5 publié à *JMIR medical informatics*)

Résumé :

Enfin, nous avons cherché à prouver que l'extraction et la classification des concepts médicaux permettaient effectivement d'extraire des phénotypes d'intérêt par similarité.

L'étude proposée ci-dessous détaille donc l'étape d'extraction d'entité nommée utilisée pour création de cohorte, la classification des concepts médicaux extraits et enfin, le calcul de la distance entre deux notes cliniques extraites.

Grâce à l'étape de classification, la distance entre deux compte-rendus peut être focalisée sur certains domaines médicaux seulement en fonction de la question clinique. Par exemple, si l'on cherche à analyser des patientes lupiques ayant présenté un infarctus du myocarde, on calculera la distance (ou la similarité) entre deux documents sur les axes cliniques « immunologique » ou « cardiovasculaire » uniquement.

La reconnaissance d'entité nommée a été réalisée à partir d'un modèle entraîné sur 152 documents annotés (avec en tout environ 11 000 symptômes et maladies annotées). Le modèle est celui dont l'architecture est développée section 2.2.1.2. avec le modèle de langage camemBERT [24] affiné sur les deux millions de comptes-rendu (correspondant à environ 5G de texte) dont nous disposons dans l'espace pour encoder le texte.

Un point qu'il nous paraît important de souligner est que l'annotation réalisée ici pour les symptômes et les maladies a été très orientée par la clinique. Le parti pris ici n'était en effet pas d'annoter le caractère nié des termes à l'échelle syntaxique, mais plutôt à l'échelle sémantique. Concrètement, les symptômes ont été annotés positifs s'ils étaient présents, et négatifs s'ils étaient absents ou s'il s'agissait d'un signe physiologique. Par exemple : lorsque l'on écrit « le patient n'est pas *fébrile* » ou « le patient est *apyrétique* » le sens des deux phrases est le même, mais la syntaxe diffère. Dans notre cadre, l'annotation sera la même : le terme *fébrile* et *apyrétique* sont tous les deux annotés négativement comme une *absence de symptôme.* De même la description « pouls périphériques perçus » n'est pas un symptôme, tandis que « pouls périphériques absents » en est un.

Cette extraction fine des symptômes permet d'avoir deux représentations du patient : l'ensemble des *signes pathologiques* (symptômes) et l'ensemble des *signes physiologiques*. Le modèle de transformers BERT étant capable de capturer l'information syntaxique et sémantique, nous obtenions une F1 score de 0,81 [0,75 ; 0,88] à l'époque de l'article.

L'étape suivante est la classification multi-étiquette des concepts extraits. Comme présenté à la partie 4.5. précédente.

Une fois les termes extraits et classés, nous obtenons une représentation finale de la note clinique avec l'ensemble des mentions textuelles (par exemple « *lupus érythémateux disséminé »),* leurs classes (pour le lupus : « *maladies immunitaires* » et « *maladies de la peau* ») et la représentation vectorielle associé à la mention textuelle [1.2223, 5.344…] pour permettre un calcul de similarité.

La représentation vectorielle choisie pour chaque concept de cette représentation patient avant calcul de la similarité est celle issu du modèle FastText [22] entraîné *from scratch* (depuis une initialisation aléatoire) sur les 2 Giga Octets de texte de notre espace-projet.

Le calcul de la distance, y compris par classe, n'est ensuite pas direct puisque deux comptes-rendus de patients peuvent contenir des nombres différents de termes (c'est-à-dire de vecteurs) par classe. Par exemple, le compte-rendu d'un patient index à traiter peut ne présenter qu'un seul terme pour l'étiquette cardiovasculaire ("péricardite lupique") alors qu'un autre patient peut présenter de nombreux termes cardiovasculaires tels que "syndrome coronarien", "hypertension", "accident vasculaire cérébral", etc.

La distance utilisée ici est donc une distance de deux distributions de termes. Suivant l'idée de Kusner et al. [94], nous utilisons la « Earth mover's distance »(EMD), une distance qui minimise le coût à payer pour transformer une distribution en une autre. Nous calculons cette distance pour chaque étiquette. Dans notre cas, les distributions correspondent à l'ensemble des termes par étiquette, et chaque terme correspond à un point. La taille du point correspond à la fréquence d'occurrence du terme, et la distance entre les points correspond à la distance en cosinus entre les représentations vectorielles Fasttext [22] des termes. En disposant d'une distance, nous sommes maintenant en mesure de comparer les notes cliniques des patients sur chaque étiquette (à condition que le dossier du patient contienne au moins un terme présent pour cette étiquette) ou globalement (c'est-à-dire sur toutes les étiquettes communes).

Pour valider cet algorithme bout-en-bout, un clinicien interniste extérieur au projet a annoté 256 documents aléatoires dans notre base de données avec 4 phénotypes d'intérêt à rechercher : « ostéoporose », « pneumopathie infectieuse », « néphropathie lupique » et « pneumopathie interstitielle diffuse secondaire à une sclérodermie ». La règle d'annotation était la présence positive de ces diagnostics dans les textes (les diagnostics ne pouvaient par exemple pas être déduits des symptômes).

Une fois le jeu annoté, nous avons considéré, pour chaque note clinique présentant un phénotype donné comme un cas index à traiter et avons extrait les patients les plus proches sur les axes associés à ce phénotype (donc « nutritionnel » et « ostéomusculaire » pour ostéoporose, « pulmonaire » et « infection » pour « pneumopathie pulmonaire infectieuse », etc..).

Les métriques de performance sont ensuite empruntées aux évaluations des moteurs de recherche : pertinence (ici précision) des 3 notes cliniques les plus proches, des 10 les plus proches et précision moyennées sur tous les documents classés (permettant d'évaluer si tous les documents présentant le phénotype sont bien tous les plus proches).

Nous obtenons une précision des 3 premiers documents entre 0.85 et 0.99 en fonction des phénotypes et une précision moyenne entre 0.58 et 0.88.

Nous avons également développé une interface avec indice colorimétrique où le clinicien peut choisir en direct les axes de similarité sur lequel il veut comparer le patient qu'il souhaite traiter. Cette interface renvoie également un nuage de mots pour les mots les plus fréquents ayant conduit à la similarité et est disponible au lien suivant :

http://xavier.tannier.free.fr/misc/patient_similarity/demo.html

Original Paper

# Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study

Christel Gérardin[1], MA, MD; Arthur Mageau[2], MD; Arsène Mékinian[3], MD, PhD; Xavier Tannier[4], PhD; Fabrice Carrat[1,5], MD, PhD

[1]Institute Pierre Louis Epidemiology and Public Health, Institut National de la Santé et de la Recherche Médicale, Sorbonne Université, Paris, France

[2]Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 1137 Infection Antimicrobials Modelling Evolution, Team Decision Sciences in Infectious Diseases, Université Paris Cité, Paris, France

[3]Service de Médecine Interne, Inflammation-Immunopathology-Biotherapy Department, Hôpital Saint-Antoine, Sorbonne Université, Assistance Publique–Hôpitaux de Paris, Paris, France

[4]Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, Institut National de la Santé et de la Recherche Médicale, Université Sorbonne, Paris, France

[5]Public Health Department, Hopital Saint-Antoine, Assistance Publique–Hôpitaux de Paris, Paris, France

**Corresponding Author:**
Christel Gérardin, MA, MD
Institute Pierre Louis Epidemiology and Public Health
Institut National de la Santé et de la Recherche Médicale, Sorbonne Université
27 rue de Chaligny
Paris, 75012
France
Phone: 33 678148466
Email: christel.ducroz-gerardin@iplesp.upmc.fr

## Abstract

**Background:** Reliable and interpretable automatic extraction of clinical phenotypes from large electronic medical record databases remains a challenge, especially in a language other than English.

**Objective:** We aimed to provide an automated end-to-end extraction of cohorts of similar patients from electronic health records for systemic diseases.

**Methods:** Our multistep algorithm includes a named-entity recognition step, a multilabel classification using medical subject headings ontology, and the computation of patient similarity. A selection of cohorts of similar patients on a priori annotated phenotypes was performed. Six phenotypes were selected for their clinical significance: P1, osteoporosis; P2, nephritis in systemic erythematosus lupus; P3, interstitial lung disease in systemic sclerosis; P4, lung infection; P5, obstetric antiphospholipid syndrome; and P6, Takayasu arteritis. We used a training set of 151 clinical notes and an independent validation set of 256 clinical notes, with annotated phenotypes, both extracted from the Assistance Publique-Hôpitaux de Paris data warehouse. We evaluated the precision of the 3 patients closest to the index patient for each phenotype with precision-at-3 and recall and average precision.

**Results:** For P1-P4, the precision-at-3 ranged from 0.85 (95% CI 0.75-0.95) to 0.99 (95% CI 0.98-1), the recall ranged from 0.53 (95% CI 0.50-0.55) to 0.83 (95% CI 0.81-0.84), and the average precision ranged from 0.58 (95% CI 0.54-0.62) to 0.88 (95% CI 0.85-0.90). P5-P6 phenotypes could not be analyzed due to the limited number of phenotypes.

**Conclusions:** Using a method close to clinical reasoning, we built a scalable and interpretable end-to-end algorithm for extracting cohorts of similar patients.

***(JMIR Med Inform 2022;10(12):e42379)*** doi: 10.2196/42379

**KEYWORDS**

natural language processing; similar patient cohort; phenotype; systemic disease; NLP; algorithm; automatic extraction; automated extraction; named entity; MeSH; medical subject heading; data extraction; text extraction

# Introduction

## Background

Extracting clinical phenotypes from large electronic health record (EHR) databases, also known as clinical data warehouses, is a key step for several medical applications from epidemiological research [1] to prognosis prediction [2,3] and therapeutic decision support [4,5]. Reliable automatic extraction of patient phenotypes from large EHR databases remains a challenge, especially in languages other than English [6]. The actual identification of patients' phenotypes is still largely done via the International Classification of Diseases, Ninth/Tenth Revision (ICD-9/ICD-10) code extraction, reading of clinical notes, or extraction of entities via regular expressions. However, as shown by Farzandipour et al [7] on more than 300 EHR ICD-10 codes, 22.7% presented errors in principal diagnosis codes, of which 33.3% were major errors. Benkhaial et al [8] also showed in a study of 200 patients, ICD allergy codes were present for 18 patients, while 51 had allergy information in a written note, indicating that only 35% of the allergies were correctly coded. These identification methods thus lack precision and require important human control.

With the improvement of natural language processing over the last 10 years, new language models such as Word2vec [9], GloVe [10], FastText [11] and, more recently, Bidirectional Encoder Representations from Transformers (BERT) [12] have allowed significant progress for various natural language processing tasks such as translation, question-answering, and named-entity recognition via an efficient word representation. Named-entity recognition corresponds to the extraction of certain classes of entities in a raw text. In the medical domain, it can be "signs and symptoms," "disorders," "chemicals and drugs," etc.

Many research teams have developed new algorithms based on these word models to allow automatic patient phenotyping. De Freitas et al [13] proposed Phe2vec, a data-driven, unsupervised disease phenotyping algorithm. In their study, disease phenotypes correspond to the word representation of ICD-10 core concepts (or seed concepts) and their closest neighbors. A patient's clinical history is summarized by aggregating all the word vector representations of the medical concepts. Mapping a patient to a disease is then done by computing a cosine distance between the patient with each disease phenotype. In their method, the medical concept extraction step from clinical notes is performed based on 1 ontology [14]. Ferté et al [15] also proposed an algorithm for automatic phenotyping of EHRs by using ICD-10 codes and a dictionary-based entity recognition tool to extract interesting terms from clinical notes. Extracted terms were then mapped to their unified medical language system concept unique identifier as a feature for classification to provide an interpretable parametric predictor. Their work showed particularly interesting results for chronic conditions.

In this work, we extracted similar patients by focusing on 4 systemic diseases as a proof of concept: systemic lupus erythematosus (SLE), systemic sclerosis, antiphospholipid syndrome (APS), and Takayasu arteritis. SLE is an autoimmune disease that can affect a large number of organs: the skin (specific malar rash, photosensitivity, etc), kidneys (nephrotic syndrome and glomerular nephropathy), joints (most often without deformation), brain (with neuropsychiatric forms), etc. It is a rare disease that affects 41 in 100,000 people in France [16], and 9 women for 1 man in generally young (18-30 years old) adults. Systemic sclerosis can also involve various organs: the skin (sclerosis leading to significant functional impotence), the lungs (interstitial lung disease [ILD], fibrosis, and hypertension), the digestive system (reflux and chronic intestinal obstruction), etc. Its frequency is 1/5000 in France, and it preferentially affects women (4 women for 1 man) aged between 40 and 50 years. APS is a disease that causes venous and arterial thrombosis as well as obstetrical complications. Approximately 20%-30% of patients with lupus develop APS. Its frequency is approximately 1 in 12,000 [16]. Takayasu arteritis is an inflammatory disease that affects large vessels in young people. It is a very rare disease affecting 1.2 to 2.6 cases/million/year in France. It affects 4.8 women for 1 man between 20 and 40 years of age [17]. These 4 diseases were chosen because of their large spectrum of signs and symptoms and their similarity (especially for lupus and APS in terms of apparition frequency and APS and Takayasu for their arterial manifestations).

## Goal of This Study

In this study, we aimed to develop an automated end-to-end extraction of similar patient cohorts from electronic medical records. Specifically, we place ourselves in the following use case: we have a patient to treat with clinical information in a text document (mentioned as index patient in this paper), and we automatically search for the set of patients with similar symptoms and diseases mentioned in their hospitalization reports. To evaluate our method, we extracted cohorts of similar patients from index patients with certain phenotypes described in their textual reports, arbitrarily selected, and manually annotated by a clinician. Our main contribution in this paper is the development of an algorithm for the automatic construction of similar patient cohorts by a method close to clinical reasoning, as we argue in the Discussion section.

# Methods

## Algorithm Steps

In this section, we detail the main steps of our algorithm. Similarity is defined here as a patient with identical or closely related signs, symptoms, and disorders. The key steps for extracting these events from the text are a named-entity recognition step to extract medical concepts, a multilabel classification on each extracted term, and an average distance computation on an appropriate representation of all the terms on each label. We validated our interpatient distance by clustering 6 a priori defined phenotypes of interest: osteoporosis, nephritis in SLE, ILD in systemic sclerosis, lung infection, obstetric APS, and Takayasu arteritis. With the same interpatient distance, we then constructed similarity cohorts from index patients for each of these phenotypes.

## Overview of the Algorithm

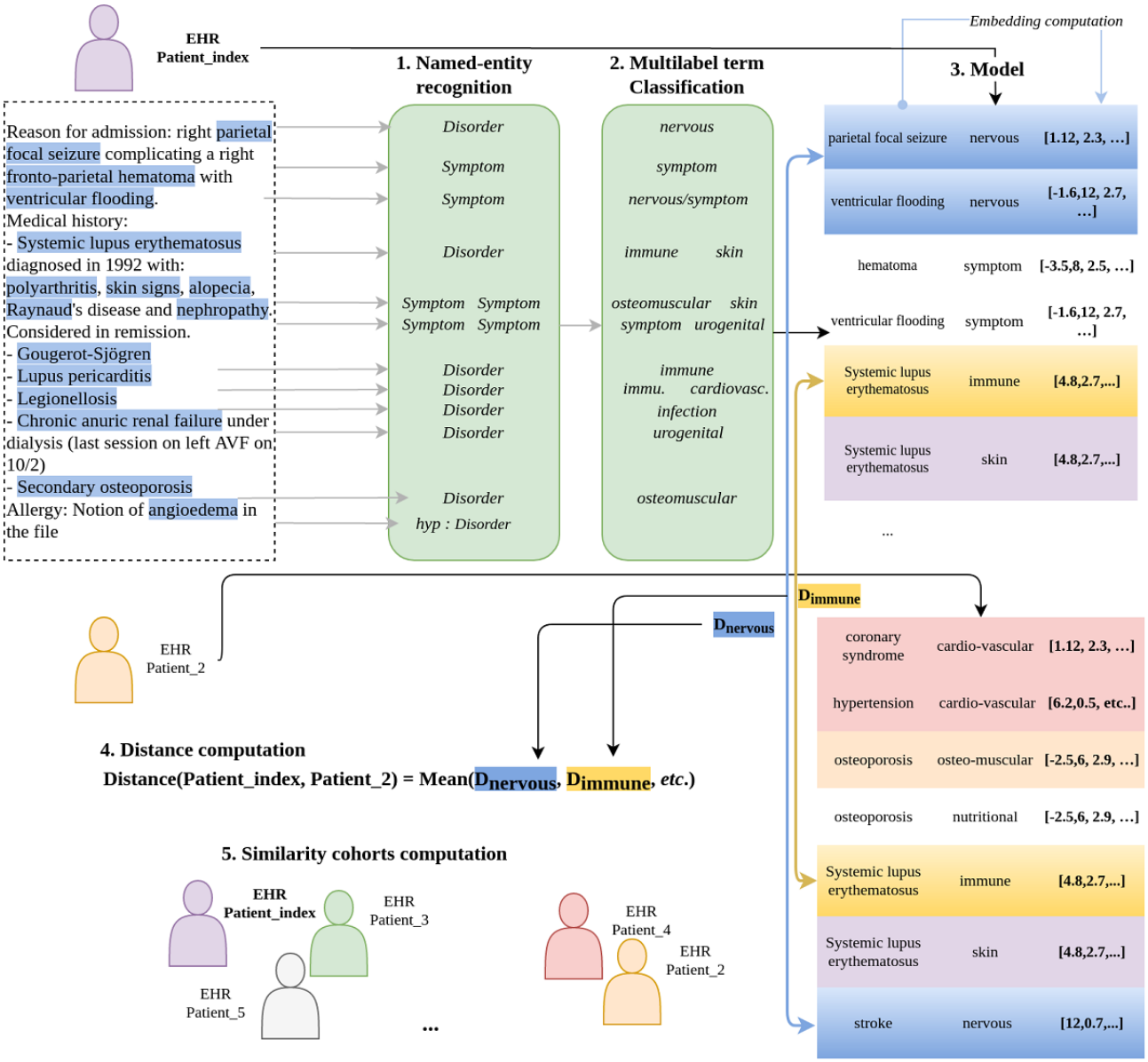
For readability, in the remainder of this paper, we use the term "patient" to refer to the "hospitalization report related to the patient."

The main steps of the algorithms are shown in Figure 1, considering an index patient:

1. Symptoms and diseases were extracted from a raw text while filtering out all negated, hypothetical, and belonging to family terms.
2. All extracted terms were classified into broad organ categories, that is, cardiovascular, immune, ophthalmologic, digestive, etc, by a multilabel classification step using our previously developed algorithm [18].
3. A vector (embedding) representation for all extracted terms was obtained leading to the index patient representation.
4. From this representation for other patients, the distance for each label of the index patient to the other patients was computed. Then, the average of the distances of all the labels was determined.
5. A cohort of similar patients was built from the patients closest to the index patient for each annotated phenotype.

We will refer to this patient's hospitalization report (Figure 1, index_patient) as a running example throughout the steps described below.

**Figure 1.** Overview of the algorithm to obtain a representation of the patients' electronic health records and to compute a distance from other patients' electronic health records. First, a named-entity recognition step is performed on a patient's electronic health record (to extract symptoms and disorders and filter all negated, hypothetical, and someone else's terms). Second, a multilabel classification step is performed for each extracted term to allow more clinical interpretation. Third, this leads to an electronic health record model containing all the extracted terms with their respective labels and embedding representations (last column of the model). Fourth, this allows a distance computation on each of the 22 labels (Dnervous corresponds to the distance between embeddings of all terms labelled nervous, Dimmune on the immune label, etc). Fifth, a similarity cohort computation is performed. EHR: Electronic Health Record.



### Data Sets and Annotation Rules

The data set of this study was obtained from the Assistance Publique-Hôpitaux de Paris (AP-HP) data warehouse. Patients were informed that their EHR information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the Commission Nationale de l'Informatique et des Libertés [19]).

The data set contained all hospitalization reports, consultation reports, test results, prescriptions, etc of all patients older than 15 years with lupus, scleroderma, APS, and Takayasu arteritis who made at least one visit to AP-HP hospitals since 2017. The data set constitutes a set of 2 million pseudonymized clinical records. It was extracted using only the ICD-10 codes of the principal diagnosis for lupus (M320, M321, M328, M329, L930, L931, corresponding to 5176 patients), systemic sclerosis (M340, M341, M348, M349, corresponding to 2833 patients), APS (D686 corresponding to 1250 patients), and Takayasu arteritis (M314, corresponding to 287 patients).

An internist physician annotated a training subset of 151 clinical notes (40 lupus, 35 APS, 37 systemic sclerosis, and 39 Takayasu) with symptoms or disorders by using specific attributes "negated," "hypothetical," and "belonging to family" when relevant. Guided by a clinical logic, we chose not only to annotate the negated terms as negation (eg, no fever, no diabetes) but also all the physiological descriptions (eg, peripheral pulse present, vesicular breath sounds present and symmetrical, regular heart sounds). All of these physiological findings were annotated as negative, because in clinical reasoning, we focus primarily on pathological signs. We adopted this approach also because the language models we use are able to capture both the syntactic and semantic levels of language. The medical subject heading (MeSH) category C [20] head chapters (eg, cardiovascular, immune, digestive) were also annotated at the entity level. This annotated data set was used to train both the named-entity recognition step with the symptoms and disorders labels and the multilabel classification step with MeSH [20] category C chapter head labels. Another test set of 256 hospitalization reports was annotated with one or more of the 6 phenotypes of interest, that is, osteoporosis, nephritis in SLE, ILD in systemic sclerosis, lung infection, obstetric APS, and Takayasu arteritis by another internist physician with no common patients between the training and testing data sets.

The annotation rules were defined before starting. First, a phenotype was only positively annotated if it was explicitly written, and no interpretation was made of signs and test results to guess the phenotype. For example, for osteoporosis, neither bone mineral density nor the number of vertebral fractures was interpreted, and the only terms retained positively were osteoporosis and corticosteroid-induced osteoporosis. Detailed examples can be found in Figure S1 of Multimedia Appendix 1. We selected these phenotypes for their clinical significance both in the 4 pathologies of interest studied and globally in terms of osteoporosis and lung infection phenotypes. These phenotypes were selected as an example, but our algorithm can be generalized to handle very different phenotypes.

### Word Representations

Two word representation models were used for this work. First, a French BERT model [12], camemBERT, trained by Martin et al [21] on a wide variety of French documents was used for the named-entity recognition and multilabel classification steps. Second, a FastText model developed by Bojanowski et al [11] was used for the patient model to calculate the interpatient distance. Both methods convert words into vectors of real numbers (called embeddings). BERT produces embeddings that take into account the context (other words in the phase), while FastText produces fixed embeddings (a word corresponds to a vector independently of the surrounding text). For our study, we had 2 million documents of all types (consultation records, hospitalization records, discharge summaries, etc), which correspond to a volume of 5 gigabytes of text. These data allowed us to train the FastText model from scratch. The camemBERT model was too large to train from scratch, but we fine-tuned it on our data, that is, we retrained its final layers. As a result, it was able to learn a context-appropriate vector representation (particularly effective for the feature extraction step 1); nevertheless, its initial vocabulary did not contain all the medical concepts, unlike the FastText model, which we used for the patient representation for the interpatient distance calculation.

### Named-entity Recognition

This first step enables us to extract positive symptoms (pathologic signs) and disorders, filtering all terms corresponding to hypothetical, negated, and family-related elements. For instance, in Figure 1 (index_patient), the extracted terms were "parietal focal status epilepticus," "frontoparietal hematoma," and "systemic lupus erythematosus," whereas "angioedema" was not kept since it was only hypothetical. The algorithm used for this first step is based on an encoder (with BERT layers) and a bidirectional long short-term memory decoder. This neural named-entity recognition model, described in [18], obtains an exact F-measurement of 0.931 on the English CoNLL data set [22], using the BERT-large embeddings [12], and 0.784 on GENIA [23], using the BioBERT-large model [24].

### Multilabel Classification

To improve clinical interpretability and to analyze patients along several medical dimensions (ie, labels), we chose to perform a multilabel classification of all the terms. The corresponding class is all the MeSH-C head chapters, corresponding to 22 medical fields: infections, ophthalmologic, stomatology, cardiovascular, digestive, respiratory, nervous, etc. A BERT model for the sequence classification was used and trained on all annotated entities and all MeSH terms and their synonyms. Synonyms of MeSH terms were obtained by extracting all the French terms sharing the same code unique identifier in the unified medical language system defined by their authors as a "set of files and software that brings together many health and biomedical vocabularies to enable interoperability between computer systems" [25]. This multilabel classifier has been

described in our previous study and evaluated on an external challenge with an F1-score from 0.809 to 0.811 depending on the model used [18]. For instance, for our index_patient in Figure 1, parietal focal status epilepticus is labelled as nervous, and systemic lupus erythematosus is labelled as immune and skin.

### Distance Computation

We used FastText to obtain an embedding representation of each extracted term. With all the patients represented as a list of embeddings for each label, the distance between the patients can be computed based on one particular label of interest (cardiovascular, urogenital, etc), or several, or all. However, 2 patient records may contain different numbers of terms (ie, vectors) per label. For example, index_patient on Figure 1 only presents 1 term on the cardiovascular label (lupus pericarditis), whereas patient_2 may present many cardiovascular terms such as coronary syndrome, hypertension, and stroke.

Following Kusner et al's [26] idea, we decided to use the earth mover's distance, a distance that minimizes the cost to be paid to transform one distribution into another. We compute this distance for each label. In our case, the distributions correspond to the set of terms per label, and each term corresponds to a point. The size of the point corresponds to the frequency of occurrence of the term, and the distance between the points corresponds to the cosine distance between the FastText embeddings of the terms. In our example, the immune label for index_patient is made of the terms SLE (1 occurrence), Raynaud (1 occurrence), Gougerot-Sjögren (1 occurrence), and lupus pericarditis (1 occurrence).

Having a distance, we are now able to compare patients' clinical notes on each label (provided that the patient's record has at least one term present for this label) or globally. To compare 2 patients globally, we summed the earth mover's distances of the 2 patients across each label and weighted them with the corresponding number of terms for each label. Equations (1) and (2) below specify the weighting term, where $HR_1$ and $HR_2$ denote 2 different hospitalization reports, and EMD() denotes the earth mover's distance between the 2 notes for a specific label i.

$$D(HR_1, HR_2) = (1/nlabels)*\Sigma\ (\lambda_i\ EMD(HR_1(label_i), HR_2\ (label_i)) \quad \textbf{(1)}$$

$$\text{with } \lambda_i = (nHR_1(label_i) + nHR_2(label_j)) / (nHR_1 + nHR_2) \quad \textbf{(2)}$$

where $HR_j(label_j)$ is the list of terms from $HR_i$ involving $label_j$ and nHR is the number of terms in the term subset HR.

### Evaluation

We evaluate our approach with the 6 use cases described earlier, each being associated with specific MeSH-C labels. For example, to obtain similar patients for the osteoporosis phenotype (labelled musculoskeletal and nutritional according to MeSH classification), we computed the earth mover's distance of the hospitalization reports only on these 2 labels. Similarly, for ILD in systemic sclerosis, we focused on the respiratory and immune labels. For lung infection, we focused on the respiratory and infections labels, and so on. However, our algorithms can be applied to any new use case and to any set of MeSH-C labels.

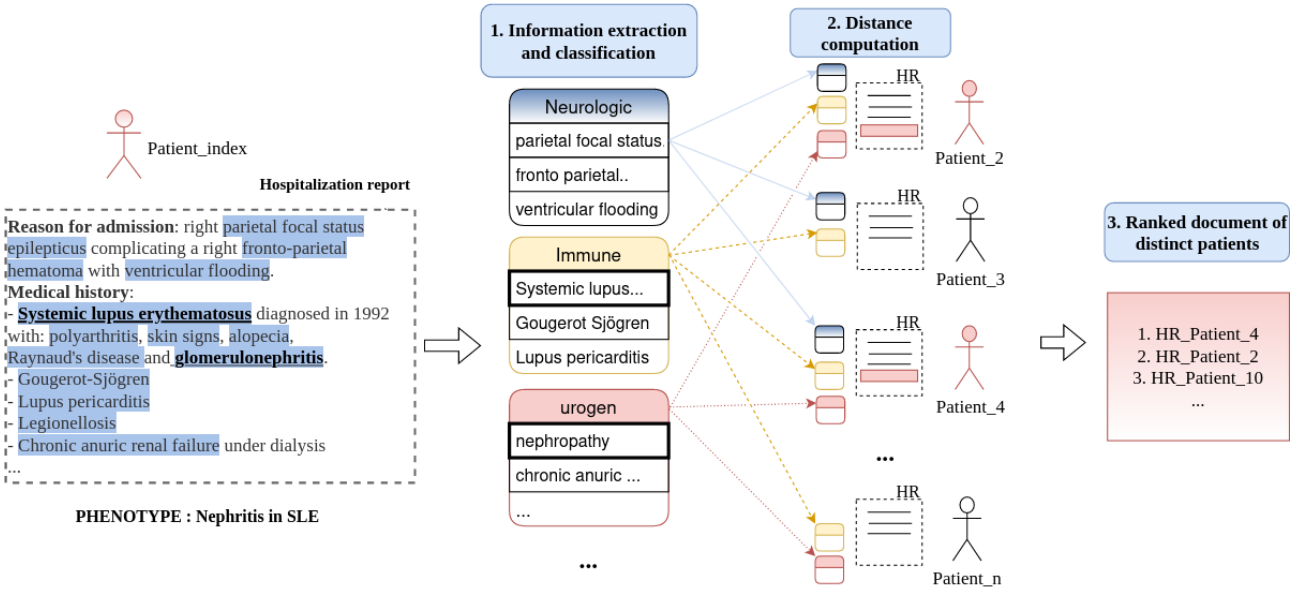
#### *Clustering*

To visualize our results and to confirm the relevance of our approaches, we performed an unsupervised hierarchical clustering of all patients in the training data set on each label and globally, checking if patients with similar phenotypes belonged to the same clusters. We used agglomerative hierarchical clustering (each hospitalization report is initialized as a singleton cluster, and clusters are merged two-by-two) with Ward's criterion, which minimizes the variance of the clusters. The same method was used for our 6 use case phenotypes. We used the SciPy library [27].

#### *Selection of a Cohort of Similar Patients From an Index Patient*

We approach the problem of building a cohort of similar patients as an information retrieval problem, where the patient's document (index patient) is a query. We then evaluate the ability of the system to return a ranked list of documents, with the most relevant/similar at the top of the list. Figure 2 gives an overview of this selection on the example of a patient with the phenotype "Nephritis in SLE." We evaluate the precision-at-k (percentage of correct phenotype prediction in the first k closest documents of distinct patients), the recall (percentage of all correct phenotypes that are selected in the first n closest patients, n being the number of patients in each phenotype), and the average precision. The average precision computes the average value of the precision for recall values over 0 to 1. It considers the order in which the patients are selected and corresponds to an estimate of the area under the precision-recall curve. For each phenotype, each patient from the test set is chosen in turn as an index patient, and the final results are an average over all patients. Confidence intervals were calculated using the normal distribution approximation.

**Figure 2.** Example of document selection for the phenotype "Nephritis in systemic lupus erythematosus." First, from the clinical observation of the index patient, symptoms and diseases are extracted and classified according to medical subject heading-C chapter headings (step 1). Then, the distance is calculated on the UroGen and immune classes (specifically for this phenotype, step 2). Finally, the closest documents are those with the same written phenotype, corresponding to the patients in red in the figure, leading to a ranked list of the closest documents of distinct patients (step 3). SLE: Systemic lupus erythematosus; HR: Hospitalization report.




### *Visualization*

A distance-based search result was also constructed to select the most similar patient to an index patient, with clickable labels where clinicians can choose any labels of interest they want to select (as in our phenotype examples). This search result returns the most similar patients on the selected labels in the descending order of similarity. A demonstration can be found in this following link [28], with 4 use cases with word clouds of medical terms enabling the similarity decision. All our codes are available on GitHub [29].

### Ethics Approval

The results shown in this study are derived from the analysis of the AP-HP data warehouse. This study and its experimental protocol was approved by the AP-HP Scientific and Ethical Committee (IRB00011591 decision CSE 20-0093). All methods were carried out in accordance with relevant guidelines (reference methodology MR-004 of the Commission Nationale de l'Informatique et des Libertés [19]). All medical records have been pseudonymized. Patients are informed by the AP-HP data warehouse that the data are pseudonymized and that they can object to their sharing. Their consent was therefore collected prior to our study.

## Results

### Clustering

The results of the unsupervised hierarchical clustering on our training data set of 151 EHRs are shown in Figure 3, Figure 4, and Figure 5. Each cluster is enhanced with its corresponding word cloud (highlighting the frequencies of occurrence of terms within each cluster). Interestingly, on the immune label (Figure 3), we were able to properly separate patients with scleroderma (left, orange cluster) from patients with lupus or lupus with APS (green clusters). As mentioned earlier, 30% of APS is secondary to systemic lupus, and indeed, several patients with APS in our data set also had lupus. Similarly, on the digestive label (Figure 4), we were able to separate upper digestive manifestations (left cluster) from liver issues (left clusters). With regard to the global clustering (using equations 1 and 2 above), we obtained 4 different clusters, as shown in Figure 5. Scleroderma is clustered separately with forms of cutaneous lupus (right, purple cluster) from lupus with thromboembolic manifestations and APS (middle, red cluster) from Takayasu (second left, green cluster). Interestingly, scleroderma with pulmonary arterial hypertension (left, little orange cluster) is close to the Takayasu cluster with arterial complications. The test set included 100 patients with lupus, 87 with scleroderma, 51 with APS, and 18 with Takayasu arteritis. Only 4 Takayasu stroke were labelled and 7 obstetrical APS, which did not allow us to perform clustering or other performance computations. The clustering results for phenotypes osteoporosis and lung infection with ground truth labelled documents are shown as examples in Figure 6 and Figure 7, respectively.

**Figure 3.** Unsupervised hierarchical clustering based on electronic health record earth mover's distance on the “immune” label. Word clouds of electronic health records words are plotted on each respective cluster. Interestingly, patients with systemic scleroderma all belong to the same cluster (orange). Only patients who were labelled “immune” are clustered; we thus represent 129 patients out of 151.

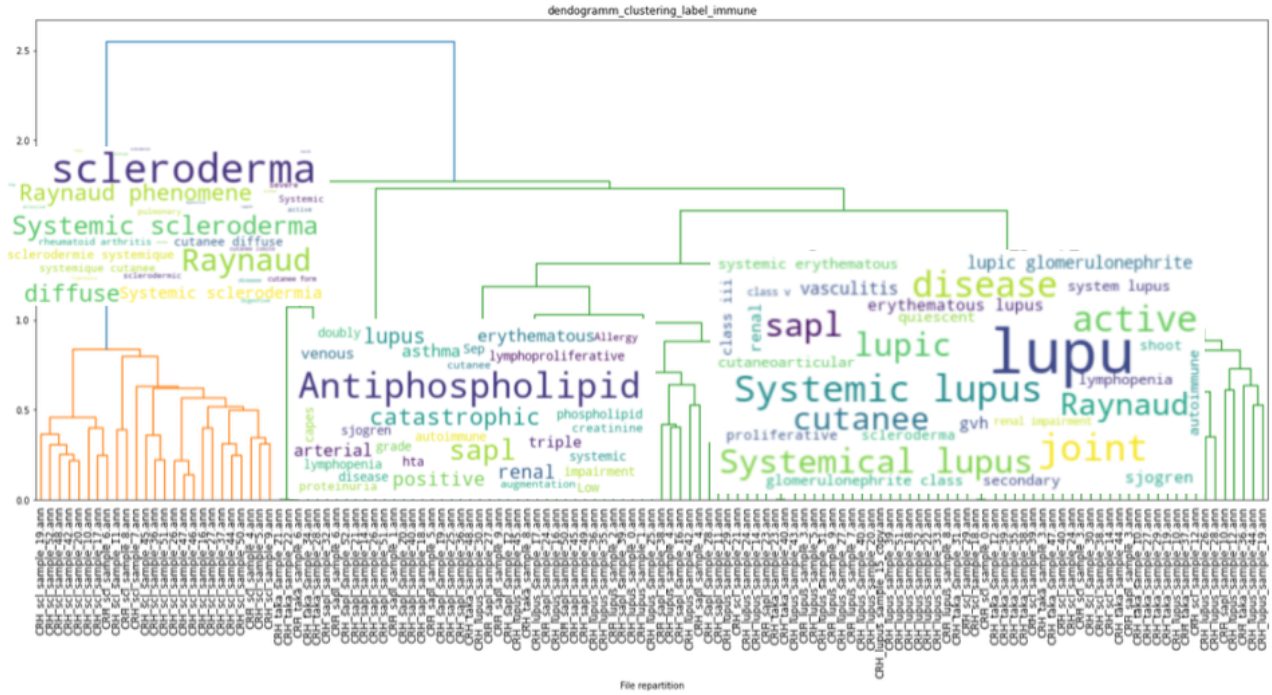



**Figure 4.** Unsupervised hierarchical clustering based on earth mover's distance of electronic health records on the label “digestive.” The word cloud of the electronic health records is shown on each respective cluster. Interestingly, the left cluster reports upper digestive manifestations (oesophagitis, gastroesophageal reflux or RGO in French), and the rightmost cluster represents patients with liver diseases (brown cluster: cytolysis, hepatitis, hepatic), whereas the middle cluster represents patients with both conditions. Only patients who were labelled digestive are clustered; we thus represent 89 patients out of 151.

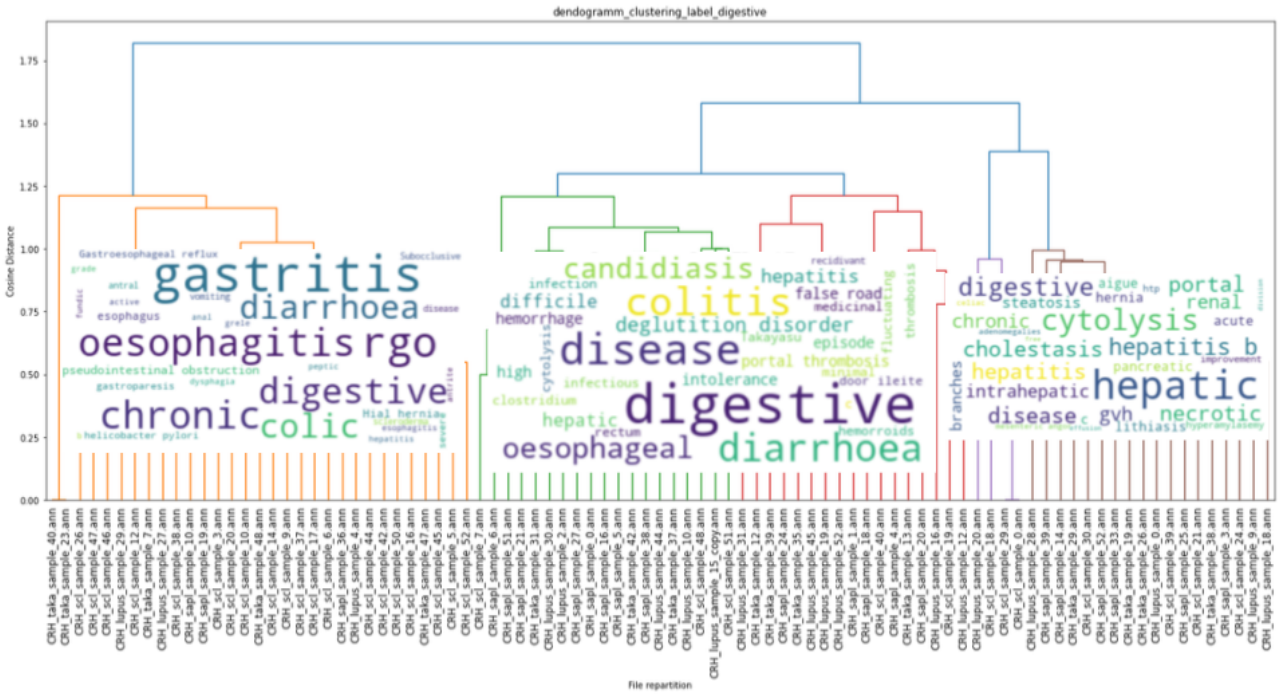


**Figure 5.** Unsupervised ascending hierarchical clustering based on the overall earth mover's distance of the electronic health records from equations (1) and (2). Word clouds of term frequency in the electronic health records are plotted on each respective cluster.

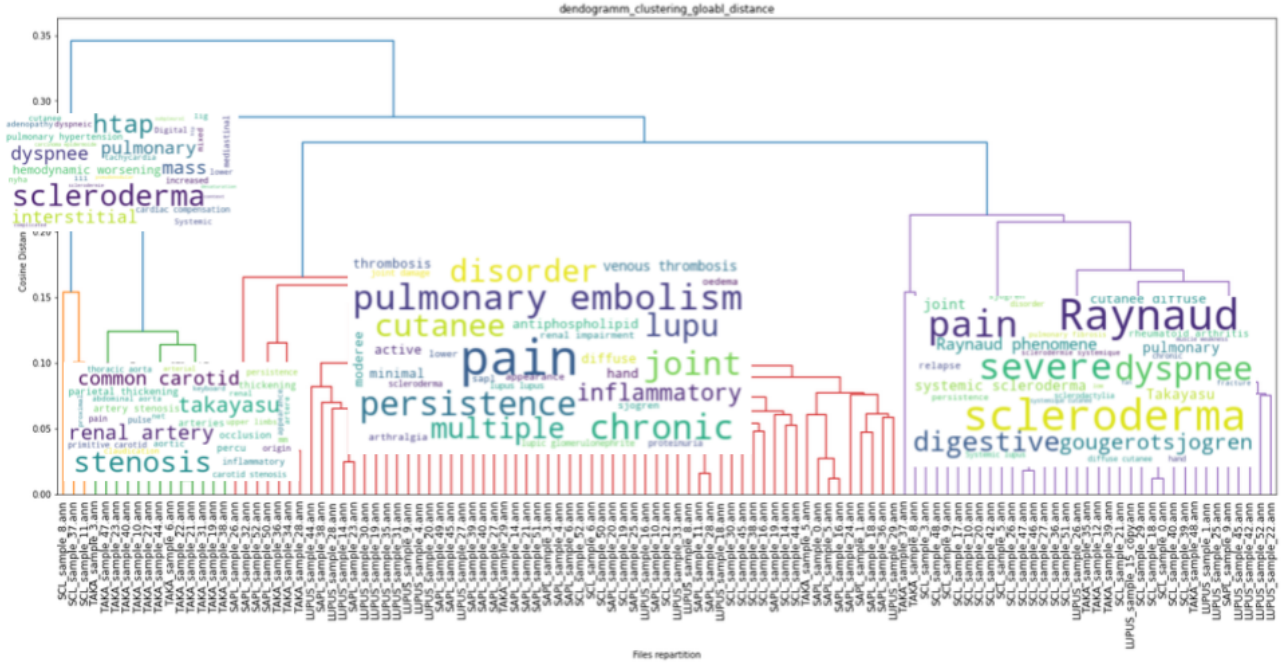



**Figure 6.** Unsupervised ascending hierarchical clustering based on earth mover's distance of electronic health records on the “osteomuscular” and “nutritional” labels (derived from the medical subject heading classification); only patients having the labels “osteomuscular” and “nutritional” are represented here (corresponding to 119 patients, not 256). All patients with osteoporosis were labelled “OSTEO” in the orange cluster. Other patients present in this cluster without explicitly written osteoporosis present “osteopenia” (all 4 first patients) of several vertebral fractures.

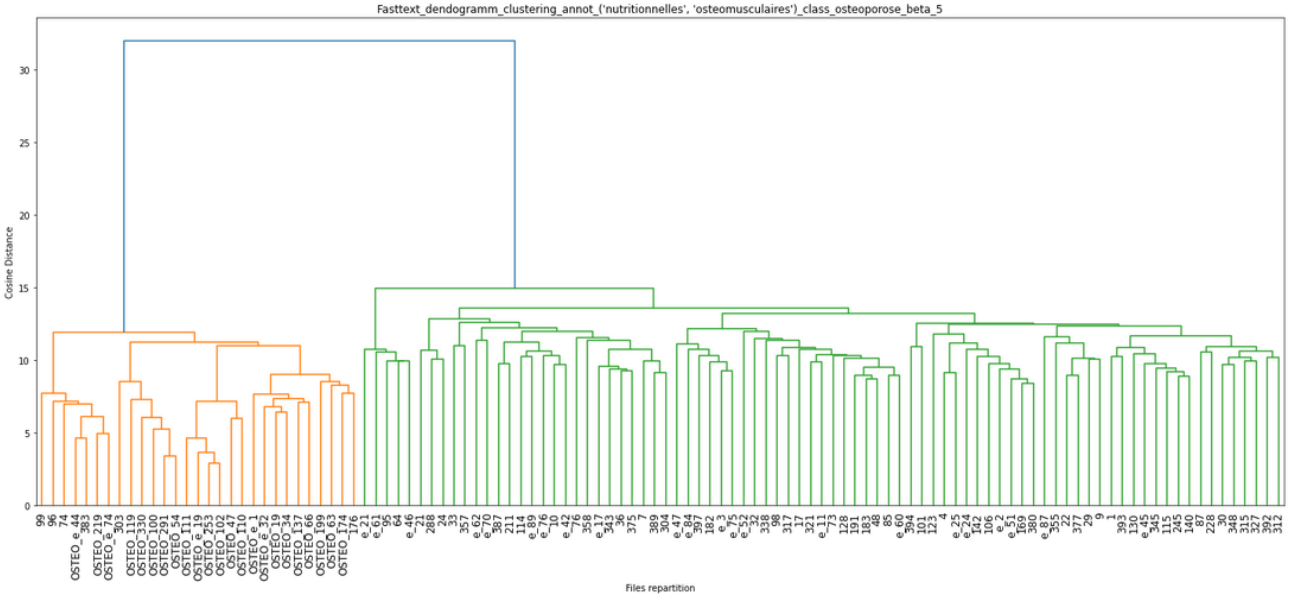
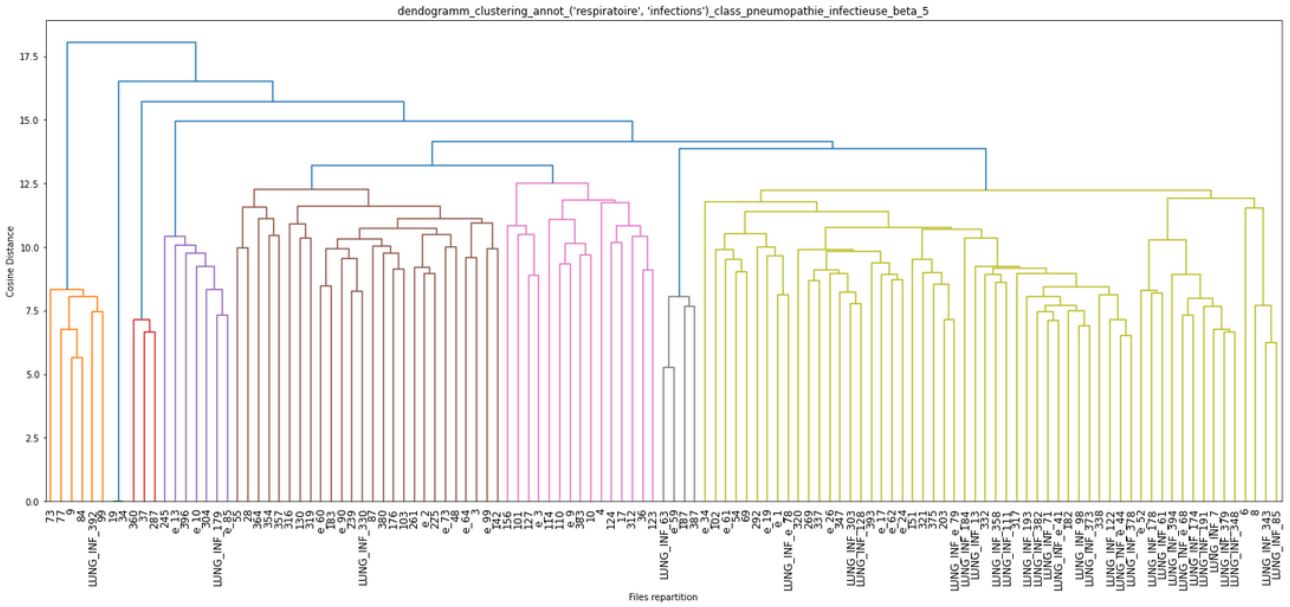


**Figure 7.** Unsupervised ascending hierarchical clustering based on earth mover's distance of electronic health records on the respiratory and infection axes (derived from the medical subject heading classification). All patients with lung infections were labelled "LUNG_INF" in the green cluster. Some outliers may be noticed; on the very left, the patient had purulent pleurisy, and one had pulmonary tuberculosis. The remaining patients on the left of the green cluster all had other linked manifestations such as bronchitis, parainfluenza infection, and bronchoalveolar lavage positive for *Klebsiella pneumoniae* and oropharyngeal flora.




## Selection of a Cohort of Similar Patients From an Index Patient

The performance of cohort construction for the first 4 phenotypes is presented in Table 1. The last 2 phenotypes (P5-P6) could not be analyzed due to a limited number of phenotypes at the annotation stage (7 and 4, respectively).

Overall, we obtained an average precision ranging from 0.58 to 0.88, precision@10 from 0.65 to 0.98, and recall from 0.53 to 0.83. However, the average precision was lower for P3 (ILD in systemic sclerosis) owing to the higher diversity of terms used to describe the lung condition, that is, fibrosis, ILD, scleroderma with pulmonary involvement, etc, and to the fact that the phenotype annotations were very specific. As an example, sclerodermatomyositis or mixed connective tissue disease with lung involvement, which are very close to this phenotype were not annotated positively. An error analysis with mention encountered on close patients can be found in Table S1 of Multimedia Appendix 1. For the 4 phenotypes P1-P4, the precision-recall curves (means for all patients within each phenotype) were computed and are shown in Figure S1 of Multimedia Appendix 1, which is another way of showing the average precision performances. We showed very good results for the P1-P2 and P4 phenotypes and satisfactory results for the P3 phenotype since the patients had to present exactly the same disease.

**Table 1.** Performance results for phenotype similarity (mean and 95% CI) for all patients of a phenotype. For each phenotype, each patient in the test set is chosen in turn as an index patient, and the final results are an average of all patients.

| | P1, osteoporosis (n=23) | P2, nephritis in systemic lupus erythematosus (n=48) | P3, interstitial lung disease in systemic sclerosis (n=20) | P4, lung infections (n=33) |
|---|---|---|---|---|
| Precision@3[a] | 0.97 (0.91-1.0) | 0.99 (0.98-1.0) | 0.85 (0.75-0.95) | 0.92 (0.84-0.99) |
| Precision@10 | 0.95 (0.91-0.99) | 0.98 (0.97-0.99) | 0.65 (0.58-0.72) | 0.86 (0.81-0.92) |
| Average precision | 0.88 (0.85-0.90) | 0.85 (0.83-0.87) | 0.58 (0.54-0.62) | 0.72 (0.69-0.75) |
| Recall[b] | 0.83 (0.81-0.84) | 0.79 (0.77-0.80) | 0.53 (0.50-0.55) | 0.66 (0.64-0.68) |

[a]Precision@3 patients (precision@10) is presented, which represents the obtained precision calculated on the 3 (or 10) patients closest to the index patient (ie, with the minimum distance).

[b]Recall is the recall calculated for all patients to be found with the same phenotype (ie, recall calculated on the 23 closest patients for osteoporosis, the 48 closest patients for nephritis in systemic lupus erythematosus, etc). Precision-recall curves for the 4 phenotypes are shown in Figure S1 of Multimedia Appendix 1.

## Visualization

As an illustration, Figures 8 and 9 below show the search results described earlier for a patient with ILD in systemic sclerosis and nephritis in SLE, respectively. We see that for an index patient with ILD in systemic sclerosis (Figure 8), choosing the immune and respiratory labels led to the finding of 10 patients out of the 15 first, having the same condition. Interestingly, among these 15 samples, the 5 unlabeled patients had a disease very close to the expected one: "ILD evolving to fibrosis" and

a “mixed connective tissue disease” for the first one (note_98, rank 4) and “sclerodermatomyositis” and “interstitial lung disease” for the second (note_182, rank 5). Further analysis of the errors is presented in Table S1 of Multimedia Appendix 1. A more extensive error analysis can be found in Table S1 of Multimedia Appendix 1. Figure 9 shows the search results for an index patient with nephritis in SLE. All the 21st closest patients on labels “immune” and “urogenital” showed nephritis in SLE.

**Figure 8.** Search results of an index patient with interstitial lung disease; the darker the color is, the closer the patients are to that particular label. Here, the selected labels “immune” and “respiratory” in 8 of the 10 first patients are labelled with “PINS_Sclerodermie” (in French, ie, interstitial lung disease in systemic sclerosis).

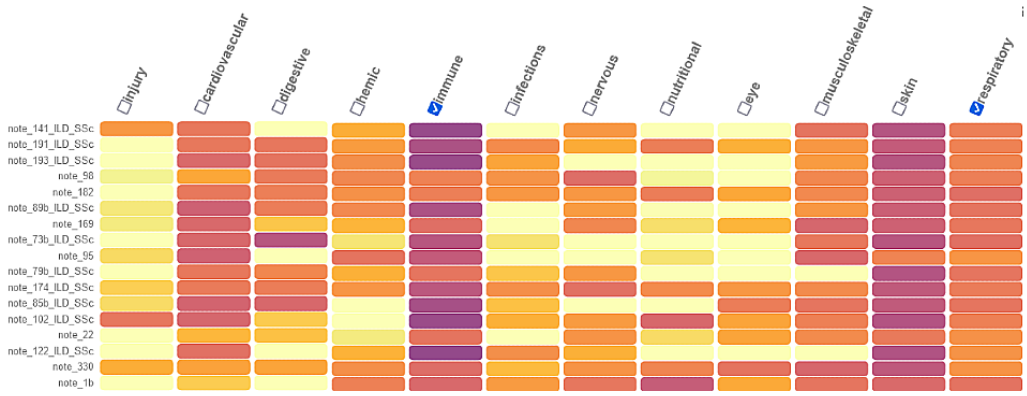
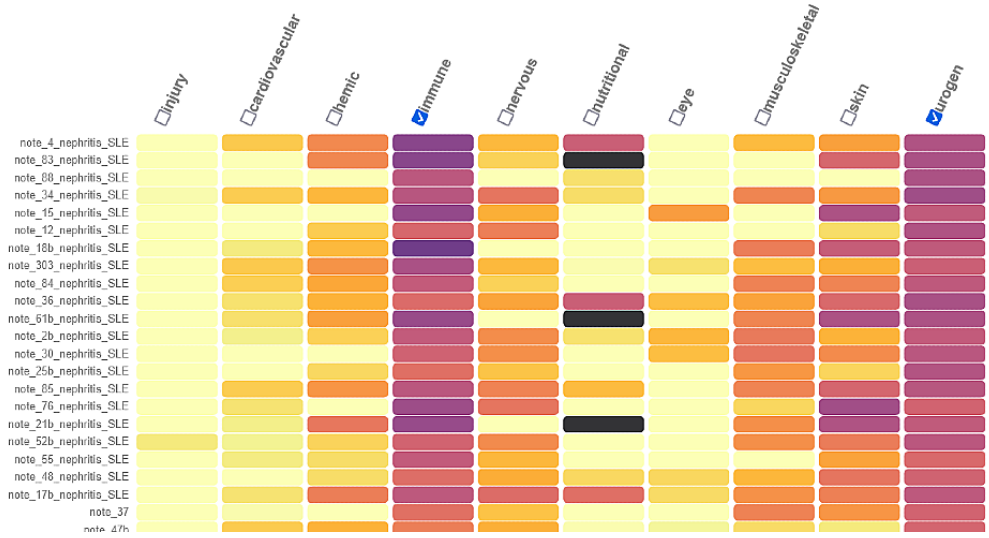



**Figure 9.** Search results of a patient with nephritis in systemic lupus erythematosus. The darker the color is, the closer the patients are to that particular label. Here, the selected labels “immune” and “urogenital” in all the 20 first closest patients are labelled with the right phenotype nephro_lupus.




## Discussion

### Summary

In this study, we developed a novel end-to-end algorithm from raw clinical notes to cohort similarity extraction. We have shown that we can cluster very specific phenotypes on an annotated data set and build similarity cohorts with good mean average precision results. These phenotypes and diseases were chosen as a proof of concept, with 2 general phenotypes such as osteoporosis and lung infection and 2 very specific phenotypes with nephritis in SLE and ILD in scleroderma. However, our algorithm can be applied to other phenotypes or diseases as well. Furthermore, our system can be applied to any other data warehouse and does not contain any handcrafted rules. An interactive demo is available online [28], and all our codes are available on GitHub [29].

### Advantages of Our Approach

The main advantage of our approach is the proximity to clinical reasoning—the named-entity recognition step focusing on the distinction between physiological and pathological signs and the observations of the patients on the 22 main medical domains (cardiovascular, pulmonary, hemic, immune)—thereby allowing clinicians to choose on which aspect patients should be similar. This analysis provides interpretable results to clinicians as well as high modularity, which is essential in the field of therapeutic decision support. In clinical practice, this algorithm would enable the physician to automatically extract similar patients, evaluate their clinical evolution, and extrapolate them to the patient they want to treat. Our algorithm focuses on 1 patient’s hospitalization report rather than on the entire patient’s record (EHR), as we want to extract patients with similar conditions and similar acute complications at a time. This algorithm is also able to compare along very fine-grained characteristics. For example, 2 patients with osteoporosis complicated by a bone

fracture will be closer than 2 patients with osteoporosis without a fracture. In addition, although our algorithm does not directly consider biological results in a quantitative manner, the clinician's interpretation of these results in the text is systematically integrated and analyzed as a symptom, for example, anemia, hypoalbuminemia, and positive antibodies. Similarly, the pathological description of imaging reports, such as an alveolar condensation in radiology images or an abnormal left ventricular ejection fraction in echocardiograms will be taken into account in our algorithm. We show very good results in terms of precision and average precision for selecting similar patient cohorts. The robustness of the algorithm is demonstrated on the one hand by the evaluation of the precision-to-3, which is calculated here not for the construction of the cohort but rather to show that there is, as expected, a gradient of similarity from the closest to the most distant patients, and on the other hand, as shown in the error analysis, patients close to a given index patient had very similar disease, even if the exact phenotype was not encountered.

## Comparison With Previous Work

Other studies have focused on patient similarity cohorts; for instance, in the French language, Garcelon et al [30] used a patient representation and a similarity measure to try to find patients with rare diseases in the Dr Warehouse database [31]. Although their objective is quite similar to ours, they used a different representation based on the term frequency–inverse document frequency weights of the extracted concept in each clinical note, and the concept extraction is based on handcrafted rules. They obtained a percentage of 71%-99% of indexed patients returning at least one similar true-positive patient within the first 30 similar patients, and the average number of patients with exactly the same disease among the 30 patients was 51%. In a second study based on the same term frequency–inverse document frequency similarity metric, they evaluated the association between clinical phenotypes and rare disease and measured the relevance of the first 50 similar patients by a domain expert a posteriori; they obtained average precision from 0.55 to 0.91 on 6 phenotypes with mean average precision of 0.79 [32]. The main differences from our method are that we focus on clinical interpretability, and our metric computation is based on one of the most recent and performant language models [12]. Moreover, in our case, the test set was annotated a priori. Jia et al [33] also proposed an interesting algorithm for diagnostic prediction based on patient similarity, but unlike our method, their named-entity recognition step is based on a dictionary of symptoms, while disorders are extracted from ICD-10 coding. The similarity regarding symptoms is binary: 1 if the symptom is shared by both patients and 0 if otherwise. The similarity of diseases is based on their respective ICD-10 similarity (using the ICD-10 coding tree structure).

Ng et al [34] presented an insightful method based on a precision cohort (ie, patient-similarity cohorts) to help clinicians make treatment decisions for chronic diseases. They trained a global similarity model on a set of thousands of predefined variables (disease variables were constructed using their ICD-9 and ICD-10 codes, laboratory variables with their Logical Observation Identifiers Names and Codes, etc) that learns a disease-specific distance (for the 3 chronic diseases presented: hypertension, type 2 diabetes mellitus, and hyperlipidemia), with significant manual work to build the training data set. The authors did not compute direct measures of similarity cohorts but the direct impact of their method, with 75%, 74%, and 85% of decision points in hypertension, diabetes, and hyperlipidemia, respectively, and with at least one significantly better treatment. In contrast, our method focused on the performance of the similarity cohorts with metrics used in the information retrieval field, does not rely on manual variable definition, and does not learn disease-specific distance but a completely generic distance. One of the main advantages of our work is the original calculation of distance per class between patients; to the best of our knowledge, there is no similar work in the literature to compare our work to. However, we show that the named-entity recognition algorithm obtained state-of-the-art results, and the multilabel classification obtained the same performance as the best team of a French national challenge [18].

## Limitations

Our work has several limitations. First, it does not cover mental health diseases, which are a completely different branch of the MeSH classification. However, training the multilabel classifier with a new label for mental health diseases with MeSH terms and synonyms can be done fairly directly based on our framework. In addition, due to time constraints, the data used in this paper were labeled by only 1 internist, and the quality of the data labeling cannot be assessed. In addition, one could argue that we did not compare our clustering and cohort similarity extraction with an ICD-10 extraction. However, because we built our initial data set with ICD-10 codes for our 4 main pathologies, we had an initial bias that we could not overcome for fair comparison. In addition, nephritis in SLE, ILD in systemic sclerosis, and lung infections do not have direct ICD-10 codes used in clinical practice. For example, "glomerular disease with SLE" has the ICD-10 "M3214" but in the entire database of 39 different hospitals, no patient had this particular code. This is because the coding is primarily done to describe the severity of the patient being managed, and this last code, in particular, does not reflect the severity of the renal involvement (in our case, codes for nephritis usually used would be N03, N04, or N05 and M320, M321, M328, and M329 for SLE). Similarly, scleroderma with pulmonary involvement has an ICD-10 code M348 that also does not appear in our database.

Assuming that an important clinical fact is repeated several times in a clinical report (eg, a patient hospitalized for acute coronary syndrome will have many cardiovascular terms linked to his/her cardiac condition), our distance computation from equations 1 and 2 depends on the number of terms in the document. Hence, 2 patients with the same major (repeated) problem would be relatively close. However, sometimes, repeated terms are not directly derived from a major clinical fact (for instance, medical history may be repeated several times without clinical relevance).

## Conclusion

In this work, we have presented a novel end-to-end interpretable algorithm to automatically extract similar patients from an index patient based on clinical note analysis. Our algorithm shows good performance results for 4 specific phenotypes in the

context of 4 systemic diseases. In this work, we focused only on pathological signs, but in clinical practice, one could also be interested in negative signs (for instance, the absence of Raynaud syndrome is very atypical in systemic sclerosis). This will be added in our future work, thereby adding a new physiological dimension to patients. In future work, the drug information will also be added for patient comparison, and similar to our presented approach, the clinician will then be able to focus only on treatments or on treatments and signs and symptoms. Finally, we will consider patients as a set of multiple longitudinal hospitalization reports (EHRs). An important perspective of this work is also to evaluate this tool in clinical practice.

## Acknowledgments

The authors would like to thank the Assistance Publique-Hôpitaux de Paris data warehouse, which provided the data and the computing power to carry out this study under good conditions. We would like to thank all the medical colleges, including the college of internal medicine, especially Prof Jean-Emmanuel Kahn, Dr Guillaume Bussone, Prof Sébastien Abad, Dr Virginie Zarrouk, Dr Noémie Chanson, Dr Antoine Dossier, Prof Luc Mouthon, and Dr Geoffrey Cheminet from the department of rheumatology. We would also like to thank Dr Augustin Latourte, Dr Florent Eymard, Prof Xavier Mariette, Dr Gaétane Nocturne, Prof Raphaele Serror, Prof Sébastien Ottaviani, Prof Francis Berenbaum, Prof Jérémie Sellam, Prof Yannick Allanore, Prof Jérôme Avouac, Prof Maxime Breban, Dr Félicie Costantino, and doctors from the dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency, and intensive care units, who gave their agreements for the use of the clinical data.

## Data Availability

The data sets generated during this study (anonymized similarity measures between patients for the 4 use cases described in this paper) are available in the data repository at this link [35]. The data sets analyzed in this study are not publicly available due the confidentiality of data from patient records, even after deidentification. However, access to the Assistance Publique-Hôpitaux de Paris data warehouse's raw data can be granted following the process described on its website [36] by contacting the Ethical and Scientific Community at secretariat.cse@aphp.fr. A prior validation of the access by the local institutional review board is required. In the case of researchers who are not from the Assistance Publique-Hôpitaux de Paris, the signature of a collaboration contract is mandatory.

## Authors' Contributions

CG was involved in conceptualization, data curation, formal analysis, investigation, methodology, software validation, writing the original draft, reviewing, and editing. Arthur M was involved in data curation, methodology, annotation, and writing the original draft. Arsène M was involved in designing the methodology and writing the original draft. XT was involved in conceptualization, formal analysis, methodology design, writing the original draft, reviewing, and editing. FC was involved in conceptualization, methodology, project administration, supervision, writing the original draft, reviewing, and editing.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Examples of terms extracted from a hospitalization report close to an index patient with interstitial lung disease.
[DOCX File , 63 KB-Multimedia Appendix 1]

## References

1. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. Cancer Res 2019 Nov 01;79(21):5463-5470 [FREE Full text] [doi: 10.1158/0008-5472.CAN-19-0579] [Medline: 31395609]
2. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A Database-driven Decision Support System: Customized Mortality Prediction. J Pers Med 2012 Sep 27;2(4):138-148 [FREE Full text] [doi: 10.3390/jpm2040138] [Medline: 23766893]
3. Lieu TA, Herrinton LJ, Buzkov DE, Liu L, Lyons D, Neugebauer R, et al. Developing a Prognostic Information System for Personalized Care in Real Time. EGEMS (Wash DC) 2019 Mar 25;7(1):2 [FREE Full text] [doi: 10.5334/egems.266] [Medline: 30937324]
4. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011 Nov 10;365(19):1758-1759. [doi: 10.1056/NEJMp1108726] [Medline: 22047518]
5. Callahan A, Polony V, Posada JD, Banda JM, Gombar S, Shah NH. ACE: the Advanced Cohort Engine for searching longitudinal patient records. J Am Med Inform Assoc 2021 Jul 14;28(7):1468-1479 [FREE Full text] [doi: 10.1093/jamia/ocab027] [Medline: 33712854]

6. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
7. Farzandipour M, Sheikhtaheri A, Sadoughi F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). International Journal of Information Management 2010 Feb;30(1):78-84 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2009.07.002]
8. Benkhaial A, Kaltschmidt J, Weisshaar E, Diepgen TL, Haefeli WE. Prescribing errors in patients with documented drug allergies: comparison of ICD-10 coding and written patient notes. Pharm World Sci 2009 Aug;31(4):464-472. [doi: 10.1007/s11096-009-9300-5] [Medline: 19412703]
9. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv 2013:1-12 [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
10. Pennington J, Socher. Global vectors for word representation. Glove; 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October; Doha, Qatar URL: http://www.aclweb.org/anthology/D14-1162 [doi: 10.3115/v1/d14-1162]
11. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. TACL 2017 Dec;5:135-146 [FREE Full text] [doi: 10.1162/tacl_a_00051]
12. Devlin J, Chang M, Lee. BERT: Pre-training of deep bidirectional transformers for language understanding. ACL Anthology. URL: https://aclanthology.org/N19-1423.pdf [accessed 2018-10-11]
13. De Freitas JK, Johnson KW, Golden E, Nadkarni GN, Dudley JT, Bottinger EP, et al. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns (N Y) 2021 Sep 10;2(9):100337 [FREE Full text] [doi: 10.1016/j.patter.2021.100337] [Medline: 34553174]
14. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinform 2009 Mar 01;2009:56-60 [FREE Full text] [Medline: 21347171]
15. Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform 2021 May;117:103746 [FREE Full text] [doi: 10.1016/j.jbi.2021.103746] [Medline: 33746080]
16. FAI2R. URL: https://www.fai2r.org/ [accessed 2022-11-25]
17. Takayasu Arteritis. URL: https://www.has-sante.fr/upload/docs/application/pdf/2020-01/pnds_takayasu_fair_-_favamulti.pdf [accessed 2022-11-25]
18. Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med 2022 Jun;128:102311. [doi: 10.1016/j.artmed.2022.102311] [Medline: 35534148]
19. CNIL. URL: https://www.cnil.fr/en/home [accessed 2018-05-10]
20. MeSH. National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/mesh/ [accessed 2017-02-10]
21. Martin L, Muller B, Suárez P, Dupont Y, Romary L, de La Clergerie E. CamemBERT: a tasty French language model. ACL Anthology. URL: https://aclanthology.org/2020.acl-main.645.pdf [accessed 2020-07-01]
22. Sang, Erik F, Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. ACL Anthology. 2003. URL: https://aclanthology.org/W03-0419.pdf [accessed 2003-06-12]
23. Kim J, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics 2003;19 Suppl 1:i180-i182. [doi: 10.1093/bioinformatics/btg1023] [Medline: 12855455]
24. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
25. Unified medical language system. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/index.html [accessed 2022-11-25]
26. Kusner M, Sun Y, Kolkin. From word embeddings to document distances. 2015 Presented at: ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning; July 6-11; Lille, France p. 957-966 URL: https://dl.acm.org/doi/10.5555/3045118.3045221
27. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):352 [FREE Full text] [doi: 10.1038/s41592-020-0772-5] [Medline: 32094914]
28. Patient similarity demo. Xavier Tannier. 2022. URL: http://xavier.tannier.free.fr/misc/patient_similarity/demo.html [accessed 2022-05-20]
29. Gérardin C. Cohort similarity. GitHub. 2022. URL: https://github.com/ChristelDG/cohort-similarity [accessed 2022-05-20]
30. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform 2017 Sep;73:51-61 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.016] [Medline: 28754522]

31. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform 2018 Apr;80:52-63 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.019] [Medline: 29501921]
32. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis 2018 May 31;13(1):85 [FREE Full text] [doi: 10.1186/s13023-018-0830-6] [Medline: 29855327]
33. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. Int J Med Inform 2020 Mar;135:104073 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104073] [Medline: 31923816]
34. Ng K, Kartoun U, Stavropoulos H, Zambrano JA, Tang PC. Personalized treatment options for chronic diseases using precision cohort analytics. Sci Rep 2021 Jan 13;11(1):1139 [FREE Full text] [doi: 10.1038/s41598-021-80967-5] [Medline: 33441956]
35. Gérardin C. Cohort similarity main data. GitHub. 2022. URL: https://github.com/ChristelDG/cohort-similarity/tree/main/data [accessed 2022-05-23]
36. Entrepot de données de Santé de l'AP-HP. Citrix Gateway. URL: https://www.eds.aphp.fr [accessed 2022-05-20]


## Abbreviations

**AP-HP:** Assistance Publique-Hôpitaux de Paris
**APS:** antiphospholipid syndrome
**BERT:** Bidirectional Encoder Representations from Transformers
**EHR:** electronic health record
**ICD-9/ICD-10:** International Classification of Diseases, Ninth/Tenth Revision
**ILD:** interstitial lung disease
**MeSH:** medical subject heading
**SLE:** systemic lupus erythematosus


*Edited by C Lovis; submitted 01.09.22; peer-reviewed by Y Xiong, J Candeliere, C Gaudet-Blavignac; comments to author 09.10.22; revised version received 17.10.22; accepted 22.10.22; published 19.12.22*

*Please cite as:*
*Gérardin C, Mageau A, Mékinian A, Tannier X, Carrat F*
*Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study*
*JMIR Med Inform 2022;10(12):e42379*
*URL: https://medinform.jmir.org/2022/12/e42379*
*doi: 10.2196/42379*
*PMID:*

## 4.7. Prochaines étapes de l'algorithme

Plusieurs branches de calcul de la cohorte de patients similaires sont actuellement en cours de développement. Il s'agit :

- de l'extraction et de la normalisation des données de biologies,
- de l'extraction et de la normalisation des documents,
- du regroupement (ou "clustering") hiérarchique ascendant pour déterminer la cohorte d'intérêt,
- de l'analyse de cette cohorte en termes de rapport de risque sur les devenirs du patient (décès, réhospitalisation etc..) en fonction des classes thérapeutiques reçues,
- de l'analyse des complications (infectieuses etc ..) directement issues du texte et de leur rapport de risque en fonction des traitements reçus.

Cette partie présente les résultats préliminaires des travaux actuels sur les données de biologies et de médicaments issues des textes.

### 4.7.1. Intégration des données de biologie

Les informations de biologie issues des données structurées (table de biologie) dans l'EDS ne reflètent que les examens réalisés à l'hôpital et n'intègrent donc pas les examens antérieurs anciens, notamment ceux qui ont parfois permis de poser le diagnostic (Anti-corps anti-nucléaires pour les patients lupiques, HbA1c pour les patients diabétiques, hémoglobine de base pour un patient drépanocytaire) ni les examens faits en ville, rapportés par le clinicien en consultation ou en hospitalisation de jour.

L'extraction des données de biologie des textes reste donc primordiale. Par ailleurs, pour une comparaison exacte des valeurs de biologie d'un patient à l'autre, la mention exacte du test biologique est extraite puis normalisée sur une base de connaissance médicale, et la valeur numérique avec les unités est également extraite.

Pour l'étape de reconnaissance d'entités nommées, un jeu de 100 comptes-rendus a été annoté avec annotation de la mention du test de biologie, par exemple « hémoglobine », « calcémie » etc. avec le label « BIO » et une autre annotation comprenait la mention complète de biologie « hémoglobine = 12g/dL », « calcémie corrigée = 2,4 mmol/L » etc. avec le label « BIO_Complète ». Ce type d'annotation permet un post-traitement simplifié pour l'extraction de la valeur numérique associée.

Pour s'assurer de la qualité de l'extraction de la valeur numérique et de l'unité, ces informations étaient également annotées manuellement.

Pour l'extraction de la biologie et l'extraction des médicaments, en vue de développer des modèles qui seront réutilisable pour l'EDS, en collaboration avec l'équipe Data science, nous avons utilisé une architecture de reconnaissance d'entité nommée un peu différente : l'encodage du texte est toujours réalisé par un camemBERT affiné sur l'EDS mais le décodage est réalisé par un CNN et un CRF en sortie pour la prédiction de l'entité. Cette architecture a également été utilisée pour extraire les mentions des médicaments et leurs attributs via une étape décrite ci-dessous. L'extraction des attributs a été réalisée par un deuxième décodeur CNN. Les performances globales ont été comparées à l'ancienne architecture et étaient similaires. Le modèle a également été testé avec le modèle de langage camemBERT-bio[29]. L'ensemble des résultats avec les intervalles de confiance calculés par bootstrapping est présenté tableau 5.

Les résultats pour la reconnaissance d'entité nommée « BIO » et « BIO_complète » sont un F1-score à 0.72 [0.64 ; 0.80] et 0.86 [0.84 ;0.89] respectivement. Le F1 score évaluant l'extraction de la mention textuelle et l'extraction de la valeur numérique est de 0.86 [0.81 ; 0.93].

Pour s'assurer de l'absence de doublon, nous avons effectué la normalisation des termes sur les CUI du groupe sémantique « Procedure » de l'UMLS [16] , type sémantique « Laboratory procedure » et ayant au moins un synonyme en code SNOMED_CT[12] en se restreignant aux termes anglais et français. Le contrôle de la normalisation a été réalisé sur un jeu annoté manuellement avec les CUI.

Un exemple de l'ensemble des synonymes considérés pour le CUI « C0201480 », correspondant au terme « bilan hépatique » est présenté tableau 4.

**Tableau 4** : Ensemble des synonymes considérés pour le terme « bilan hépatique ».

| | |
|---|---|
| CUI UMLS | C0201480 |
| Synonymes SNOMED_CT [12] | “Hepatic function panel”, “Hepatic function panel (procedure)”, “LFT - Liver function test”, “Liver function test” |
| Synonymes MeSH[REF] Français | “Tests de la fonction hépatique”, “Exploration fonctionnelle hépatique”, “Tests d'exploration de la fonction hépatique”, “Tests fonctionnels hépatiques”, “Tests hépatiques”, |
| Synonymes Meddra[REF] Français | “Test hépatique”, “Tests hépatiques SAI”, “Tests hépatiques” |

Pour réaliser la normalisation, plusieurs algorithmes ont actuellement été testés :

- Un algorithme de *fuzzy matching* : comparaison inexacte par méthode de Sørensen-Dice [95] qui consiste à découper le mot en digramme (chaîne de caractère de deux lettres) et à calculer le rapport entre le nombre de digrammes communs (intersection entre deux mots) sur la somme des digrammes de chaque mot, dont la formule est rapportée ci-dessous.

$$s = \frac{2|X \cap Y|}{|X| + |Y|}.$$ Formule de Sorensen (wikipédia), pour deux ensemble finis quelconques

  Cette méthode obtient une précision de 0.63 [0.58 – 0.67]

- Une méthode basée sur l’algorithme mesurant la distance sinusienne de l’embedding CODER [60] du terme extrait, comparé avec tous les embeddings CODER [60] de l’UMLS [16], restreint aux synonymes SNOMED_CT [12] et aux type sémantiques « laboratory procedure », avec une précision de 0.70 [0.63 - 0.77].

- Une méthode basée sur un réentraînement de l’algorithme CODER [60] avec les poids du modèle camemBERT entraîné sur camemBERT-EDS [31]. Les résultats préliminaires montrent une précision de 0.68 [0.61 ; 0.75], après un entraînement de 300 000 itérations. L’article initial de CODER[60] proposant 1 000 000 d’itérations et notre courbe de la fonction d’erreur poursuivant sa décroissance après les 300 000 itérations, nous comptons reprendre

l'entraînement dès que les ressources informatiques seront disponibles sur une durée de plusieurs jours consécutifs.

Une fois extraite, et normalisée, l'intégration des données de biologie dans le calcul multimodal de la distance entre les comptes-rendu médicaux sera une option.

### 4.7.2. Intégration des données de médicaments

**Tableau 5**: Résultats préliminaires de l'extraction des entités de médicaments et de biologies et des attributs des données de médicaments et de biologie le modèle de langage camemBERT entrainé sur l'EDS [31] à gauche et le camemBERT-bio [29] à droite.

Résultats classifications entités et attributs

| | | Camembert Finetune | | Camembert Bio | |
|---|---|---|---|---|---|
| Catégorie | Entité | F1 | IC 95% | F1 | IC 95% |
| Entités | | | | | |
| Classiques | Chemical and drugs | 0.95 | 0.92-0.97 | 0.86 | 0.83-0.88 |
| | Bio_comp | 0.86 | 0.84-0.89 | 0.83 | 0.80-0.85 |
| | Bio | 0.72 | 0.64-0.80 | 0.56 | 0.49-0.63 |
| | Diso | 0.84 | 0.82-0.87 | 0.75 | 0.73-0.78 |
| Tech | Dosage | 0.91 | 0.86-0.95 | 0.81 | 0.77-0.85 |
| | Strength | 0.95 | 0.91-0.99 | 0.88 | 0.81-0.93 |
| | Route | 0.71 | 0.46-0.86 | 0.57 | 0.00-0-89 |
| | Forme | 0.68 | 0.56 - 0,80 | 0.69 | 0.55-0.81 |
| Attributs | | | | | |
| Action | Start | 0.57 | | 0.56 | |
| | Stop | 0.75 | | 0.72 | |
| | UniqueDose | 0.61 | | 0.54 | |
| Certainty | Certain | 0.97 | | 0.96 | |
| | Conditional | 0.83 | | 0.91 | |
| Temporality | Past | 0.80 | | 0.71 | |
| | Present | 0.83 | | 0.79 | |
| Negation | neg | 0.77 | | 0.73 | |

Une fois l'extraction des médicaments réalisée, chaque mention textuelle est normalisée sur un code ATC (Classification hiérarchique commune internationale des médicaments). Les premiers résultats de normalisation indiquent 73% de précision avec la distance de Sorensen dice. Les abréviations (MTX pour méthotrexate, RTX pour Rituximab, etc.) restent à traiter et seront normalisées par l'algorithme du CODER [60], car existantes dans l'UMLS [16].

Un exemple du résultat de cette normalisation est fourni figure 20.

| | drug | ATC | score | pred_atc | pred_string |
|---|---|---|---|---|---|
| 22 | propanolol | C07AA05 | 0.952381 | [C07AA05] | [propranolol] |
| 28 | kayexalate | V03AE01 | 0.952381 | [V03AE01] | [kayexalate,] |
| 71 | prevenar13 | J07AL02 | 0.888889 | [J07AL02] | [prevenar] |
| 106 | prevenar13 | J07AL02 | 0.888889 | [J07AL02] | [prevenar] |
| 150 | clpidogrel | B01AC04 | 0.952381 | [B01AC04] | [clopidogrel] |

**Figure 20** : Etape de normalisation des mentions textuelles des médicaments issus du texte. La colonne “drug” correspond à la mention issue du texte et la colonne “pred_string” à sa prédiction.

Au cours de ces dernières expérimentations, nous avons également considéré la taille du jeu de données annotées comme un hyperparamètre. Cet hyperparamètre peut être comparé par exemple au nombre de sujets nécessaires dans les études épidémiologiques. Pour déterminer cet hyperparamètre, deux méthodes peuvent être utilisées : un calcul a priori comme proposé par Liu et al. [96], mais les valeurs prédictives positives et négatives ne sont pas évidentes à anticiper dans notre contexte, ou bien une évaluation a posteriori en calculant la courbe d’évolution des métriques de performance avec différentes tailles du jeu d'entraînement. Pour les différentes expériences, nous avons montré que ces courbes atteignent un plateau à partir de 70 documents annotés et 30 pour l’extraction des médicaments, figure 21.

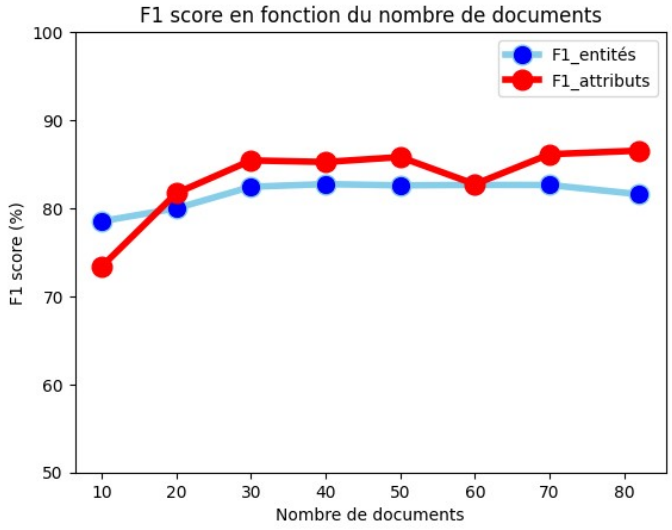



**Figure 21** : Evolution du score F1 en fonction du nombre de documents dans le jeu d'entraînement, pour le type sémantique “médicaments”.

Pour l’expérimentation des données de biologies et de médicaments, nous avons réalisé un transfert de modèle sur un autre projet de l’EDS (TRADIAB) avec les mêmes règles d'annotation discutées entre cliniciens. Les analyses sont en cours pour s’assurer de la robustesse du modèle.

# 5. Discussion et perspectives

## 5.1. Synthèse

Le projet de recherche présenté dans ce manuscrit développe les premières étapes, avec leurs validations intermédiaires, de la construction d'un prototype ayant pour objectif l'aide à la décision thérapeutique en contexte de maladie rare.

### 5.1.1. Contributions principales du projet

Au vu des données de la littérature [63], et en particulier de l'expérience de l'équipe de Stanford [75], nous sommes conscients que de très nombreuses solutions ont été développées et que peu d'entre elles sont finalement utilisées en pratique clinique. Les contraintes majeures identifiées dans les propositions précédentes sont généralement :

- le manque de collaboration entre cliniciens et ingénieurs [63]
- le manque d'interprétabilité des modèles, comme rappelé par l'avis récent du comité national d'éthique [79]
- le manque de modularité des modèles
- le manque de généralisabilité des modèles [39, 75]

Il nous semble que les contributions principales du projet décrit ici tiennent spécifiquement compte de ces contraintes pour répondre aux recommandations du conseil national d'éthique [79] et du centre "Intelligence artificielle centrée sur l'humain" de l'université de Standford [75].

L'interprétabilité en particulier de l'algorithme, en se basant sur le raisonnement clinique, a en effet guidé de nombreux choix :

- l'extraction améliorée des sections (article 1) du compte-rendu hospitalier pour permettre aux cliniciens de retrouver la segmentation des informations telle qu'ils la connaissent ;
- la reconnaissance d'entités nommées pour extraire spécifiquement les symptômes et signes pathologiques des signes physiologiques ;
- la classification de ces symptômes et pathologies par spécialité d'organe ;

- la proposition d'une interface visuelle avec échelle colorimétrique en fonction du degré de similarité, associée à un rendu par nuage de mots pour mettre en évidence les concepts médicaux à l'origine de la cohorte de patients similaires.

Une dimension modulaire sera également apportée dans le prototype final avec un choix possible d'une modulation de la distance entre patients selon certaines spécialités médicales, ou globalement ; le clinicien pourra également choisir d'ajouter ou non l'information des tests biologiques dans le calcul de la distance.

Concernant le caractère généralisable, de nombreuses étapes du modèle ont été évaluées sur des golds standards de la littérature [91] ou comparées à des challenge nationaux [92] et internationaux (participation au challenge n2c2 2022 [51]). Par ailleurs, un grand jeu d'hyperparamètres a été testé : différentes architectures de modèles pour la reconnaissance d'entité nommée et pour l'étape de normalisation, différents modèles de langage, différentes initialisations par plusieurs graines aléatoires, différentes tailles de jeu de données et une multiplicité d'annotations.

Notre modèle sera par ailleurs testé sur un autre projet de recherche de l'EDS concernant notamment l'extraction des données de biologie et de traitement.

Enfin concernant la collaboration avec des cliniciens, plusieurs d'entre eux ont participé au projet et nous réalisons actuellement une campagne d'annotation (projet CSE 21-36) à l'EDS, avec le même objectif d'annotation des symptômes et des signes physiologiques. Cette base d'apprentissage constituera un gold standard local, accessible aux autres projets de l'EDS sur demande au comité scientifique et éthique et permettra également de tester la généralisabilité de la démarche.

Il existe néanmoins un certain nombre de limites à considérer et à exposer avec transparence aux cliniciens.

### 5.1.2. Limites

Malgré les efforts réalisés pour permettre la meilleure généralisation possible, plusieurs risques de biais persistent dans notre travail.

Ainsi, les diverses annotations proposées dans ce projet pour les étapes d'extractions d'entités nommées et de contrôle de la normalisation ont été réalisées par une seule personne- l'auteure de ce manuscrit (en dehors des phénotypes de la cohorte de patients similaires, des jeux de données publics ou issus des challenges). Il peut donc y avoir des biais d'annotation (regard orienté par la qualification en médecine interne, précision importante sur l'examen clinique mais manque de précision, par exemple, sur la sémiologie radiologique ou anatomopathologique, etc.).

Par ailleurs, comme évoqué dans la partie "Enjeux Ethiques" de l'état de l'art, il peut exister des biais secondaires à la base d'apprentissage. Tout d'abord, comme le rappelle Gao et al. [63], la cohorte sur laquelle les modèles sont entraînés est issue de comptes-rendus hospitaliers de centres hospitalo-universitaires parisiens et la majorité des comptes-rendus sont rédigés par des internes, i.e. par des médecins non seniors. Les résultats présentés ici ne sont donc sans doute pas transposables à d'autres hôpitaux, en particulier non universitaires où il y a peu, ou pas, d'interne.

Ensuite, les analyses présentées sont toutes réalisées sur des comptes-rendus hospitaliers, c'est-à-dire le compte-rendu de sortie du patient, qui est généralement plus détaillé mais qui ne reflète peut-être pas au mieux l'évolution clinique en temps réel [63].

Le manque de gold standard comme mentionné partie IV.K. , notamment pour l'étape de pré-sélection des documents, reste également une limite importante pour s'assurer de la performance du modèle notamment à ne pas "manquer" des patients similaires.

Enfin et principalement, le prototype définitif n'a pas encore été testé auprès des cliniciens pour déterminer son apport dans la décision thérapeutique.

## 5.2. Perspectives : prochaines étape de développement

En dehors de l'ajout de l'information de biologie et des traitements déjà précisés, les étapes actuelles de développement sont les suivantes:

1. Application de l'algorithme bout-en-bout sur le cas d'usage de la *crise rénale sclérodermique,* à partir d'un jeu de 3000 comptes-rendus potentiels, extraits par CIM10 de façon élargie, avec relecture manuelle pour s'assurer du phénotype. Puis contrôle de toutes les étapes avec diminution simulée des performances (pré sélection de document, reconnaissance d'entité nommée, classification) pour déterminer les effets sur le *clustering* final.

Après validation de ces étapes, nous déterminerons la présence ou l'absence de prise de corticoïdes systémiques (dans les données structurées et dans les textes) antérieurement à la crise rénale sclérodermique pour déterminer si nous retrouvons effectivement une corrélation comme proposé dans la littérature [97].

2. Le modèle sera par ailleurs par la suite amélioré par une modélisation longitudinale des comptes-rendus patients. En effet, jusqu'alors, le modèle est considéré à l'échelle du compte-rendu uniquement. Ce travail sera conduit dans le cadre d'un travail de thèse de science ayant obtenu une bourse régionale AI4IDF [98].

## 5.3. Perspectives : étude pilote pour l'évaluation du prototype en pratique clinique

Enfin, au vu des résultats prometteurs présentés ici et de l'accueil positif de l'algorithme présenté dans différentes équipes hospitalières de l'AP-HP (service de médecine interne à Saint-Antoine, Service de médecine interne à l'hôpital Bichat, service de néphrologie pédiatrique à l'hôpital Robert Debré, service de néphrologie à l'hôpital Tenon, service de Dermatologie à l'hôpital Saint Louis, service de Santé Publique à la Pitié-Salpêtrière) ce projet va être poursuivi par une étude pilote visant à tester le prototype en réunion thérapeutique dans deux services de médecine interne de l'hôpital Bichat et de l'hôpital Saint Antoine.

Cette étude pilote aura lieu en amont d'un essai randomisé (étude clinique où l'on compare deux groupes de patients : un groupe sur lequel on réalise une intervention-ici l'analyse de la cohorte de patients- et un groupe témoin). Il s'agira à la fois de valider l'outil et de mesurer les changements potentiels dans la prise en charge thérapeutique associés à son utilisation. Cette étude pilote sera une étude prospective de deux ans évaluant l'impact de la synthèse d'informations issues de l'analyse d'une cohorte de patients similaires sur les décisions thérapeutiques prises lors des réunions collégiales de médecine interne.

Seront inclus les patients atteints de maladie de Takayasu, de lupus érythémateux disséminé, de sclérodermie systémique et de syndrome des antiphospholipides suivis dans les services de médecine

interne de l'hôpital Bichat et de l'hôpital Saint Antoine et dont le cas est programmé pour être discuté en réunion thérapeutique collégiale. Pour chaque patient inclus, il y aura deux discussions thérapeutiques au cours d'une même réunion collégiale de médecine interne en 2 phases : la première se déroulera dans les conditions habituelles d'organisation SANS l'outil de décision ; la seconde avec les mêmes informations AVEC les informations complémentaires issues de l'analyse et de la présentation des données obtenues dans la cohorte de patients similaires. La cohorte de patients similaires sera calculée avant la réunion collégiale à partir d'une liste hebdomadaire préétablie de patients. La "décision de traitement de référence" serait la première, afin d'éviter tout effet potentiel sur le patient. La "décision de traitement assistée par ordinateur" serait prise à des fins de recherche, mais ne serait pas appliquée au patient, qui, par conséquent, ne tirerait aucun bénéfice ou préjudice des suggestions de l'outil.

Le principal critère d'évaluation sera de déterminer si les informations d'une cohorte de patients similaires modifient la décision thérapeutique. La modification de la décision thérapeutique est définie comme l'introduction, la suppression ou le changement d'un médicament par rapport à la prescription initiale à la suite de la réunion collégiale de médecine interne sur le traitement.

Les critères d'évaluation secondaires sont les suivants:

En ce qui concerne la santé du patient :

1/ évaluation de la signification clinique du changement de traitement (c'est-à-dire que la décision de traitement a été réellement modifiée et non simplement ajustée à la marge) par un comité d'experts, en aveugle de l'étude ;

2/ pourcentage de changements de posologie, pourcentage de traitements supprimés ou ajoutés et nombre de lignes de traitement modifiées.

3/ pourcentage de réhospitalisation à 30 jours pour une seconde étude interventionnelle prospective.

En ce qui concerne l'opinion des utilisateurs, deux grilles d'évaluation seront utilisées pour évaluer à la fois l'acceptabilité et la facilité d'utilisation de l'outil : le questionnaire SUS validé de 10 questions [99] et le questionnaire USE [100] de 30 questions proposé par [101] permettant une évaluation plus

poussée de la cohérence des résultats. Le temps supplémentaire nécessaire à l'interprétation et à la discussion des résultats de cohortes de patients similaires sera également enregistré.

Ce projet d'étude pilote a été soumis à l'appel à projet MESSIDORE de l'Inserm en octobre 2022, il sera à nouveau soumis avec les nouveaux résultats en septembre 2023.

D'autres applications sont également actuellement envisagées pour ce prototype d'extraction de cohortes de patients similaires, notamment par exemple pour la pré-sélection de patients dans les études cliniques, à partir d'un certain nombre de patients déjà identifiés.

## 5.4. Projets d'interface avec l'entrepôt de données de santé

Pour finir, cette thèse m'a permis de collaborer étroitement avec les différentes équipes de l'EDS sur des projets fonctionnels et de recherche dont quelques-uns sont présentés ici.

### 5.4.1. Classification des types de documents

La classification des types de documents dans l'EDS nous paraît être un bon exemple de l'utilisation des données issues du soin et de la nécessité de collaboration entre médecins et ingénieurs pour l'obtention d'une base de données interprétable et utilisable secondairement pour le pilotage, par exemple pour la remontée d'indicateur de qualité des soins - complétude des compte-rendus hospitaliers, complétude des ordonnances de sorties etc.- ou pour la recherche.
En effet, actuellement, le type de compte-rendu médical (compte-rendu de consultation, compte-rendu d'examen, compte-rendu d'accouchement, etc..), n'est pas remonté de façon structurée. Les informations disponibles pour le type de compte-rendu sont issues d'un champ éditable par le clinicien, par exemple : "CS Mme XXX le 12/10/15" ou encore "CR anapath" etc. Il n'y a donc pas de normalisation en amont de l'EDS. Ces champs sont dénommés "libellé-document".

Les types actuels proposés dans l'EDS sont issus d'un ensemble de règles basées sur des expressions régulières créées au fur et à mesure de la mise en oeuvre de projets de recherche, avec un grand nombre de types qui se recoupent, par exemple "LT-CS", "CR-CS", "LT-type-CS" pour les comptes-rendus de consultation et des types sans pertinence clinique "CRH-EEG" par exemple,

littéralement “compte-rendus hospitaliers - Electro-encéphalogramme”. Le résultat actuel est un nombre important de documents inconnus (environ 7 millions sur les 120 millions de documents).

Partant de ce constat, nous avons fondé une nomenclature complète avec plus de 50 types, à partir du référentiel d’interopérabilité proposé par l’agence du numérique en santé [], comptant de nombreux types supplémentaires pour les examens (compte-rendu de coloscopie, de coronarographie, d’échographie doppler), tous normalisés avec un code issu de l’ontologie médicale internationale LOINC Document Ontology [102].

A partir de cette classification, nous avons construit un jeu d'entraînement pour pouvoir représenter au mieux tous ces documents. Pour mémoire, la répartition initiale était très déséquilibrée, avec un très grand nombre d'ordonnances et peu de comptes-rendus d’examens spécifiques. Pour réaliser ce jeu d'entraînement, nous avons sélectionné les documents à partir du libellé-document qui, bien que bruité, contient en général l’information principale. Ce libellé-document a été modélisé par TF-IDF puis un algorithme de regroupement des termes a été appliqué, par la méthode du K-means, avec initialisation par un terme spécifique pour chaque type de document. Cette méthode a permis une meilleure représentativité des types. 1300 documents ont déjà été annotés avec l’annotation de la cohérence de la date de document et de la cohérence du titre du document.

Les algorithmes de classification de ces documents, pour le moment à partir d’expression régulières sur les titres extraits par l’algorithme présenté section 4.1., sont en cours de développement.

### 5.4.2. Projet Européen Horizon Health

Enfin, j’ai également pu travailler avec l’équipe qualité données à l’EDS notamment pour proposer un ensemble de variables de suivi après audit des variables intérêts de plus de 30 projets de recherche : unicité des IPP patients, conformité des dates de naissance etc.

Ce travail s’inscrit notamment dans le cadre la participation de l’EDS au consortium européen Quantum (Quality, Utility and Maturity Measured), composé de 13 pays et plus de 30 institutions pour développer un label de qualité et utilisabilité de les données, en réponse à l’appel à projet européen Horizon Health 2023 [103], dont je suis la responsable scientifique à l’AP-HP.

# 6. Conclusion

Nous avons présenté ici un travail de recherche visant au développement d'une aide à la décision thérapeutique sur maladies auto-immunes, par identification automatique, à travers les textes cliniques conservés dans l'entrepôt de données de santé de l'AP-HP, de patients similaires.

L'algorithme a été développé en prenant en compte constamment le raisonnement clinique, et ses principaux modules ont été validés par de nombreuses expériences. Au vu de ces premiers résultats prometteurs, une étude pilote a été proposée, pour évaluer le prototype en pratique clinique.

Seules les études et les développements à venir permettront de savoir si l'algorithme atteindra pleinement son objectif d'aide à la décision thérapeutique. Dans l'attente, la conduite de ces travaux a renforcé chez nous la conviction que l'intégration des connaissances médicales au cœur des algorithmes, telle que présentée ici, est l'une des clés pour mettre les possibilités de l'intelligence artificielle au service de la pratique thérapeutique.

# 7. Références

[1] Frankovich, Jennifer, Christopher A. Longhurst, et Scott M. Sutherland. Evidence-Based Medicine in the EMR Era. *New England Journal of Medicine* 365, n° 19 (2 novembre 2011): 1758‑59.

[2] Lamy, Jean-Baptiste, et al. "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach." *Artificial intelligence in medicine* 94 (2019): 42-53.

[3] Suo, Qiuling, et al. "Deep patient similarity learning for personalized healthcare." *IEEE transactions on nanobioscience* 17.3 (2018): 219-227.

[4] Delvin J., Changm. W., Leek. & Toutanovak (2019). 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, 4171–4186.

[5] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[6] https://chat.openai.com

[7] Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, Markus AF, Pera V, Rekkas A, Williams RD, Yang C. Use of unstructured text in prognostic clinical prediction models: a systematic review. Journal of the American Medical Informatics Association. 2022 Jul;29(7):1292-302.

[8] Soni S, Roberts K. (2020). Patient Cohort Retrieval using Transformer Language Models. InAMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1150). American Medical Informatics Association.

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

[10] Shang J, Ma T, Xiao C, Sun J. (2019). Pre-training of graph augmented transformers for medication recommendation. arXiv preprint arXiv:1906.00346. 2019 Jun 2.

[11] Hao T, Huang Z, Liang L, Weng H, Tang B. Health Natural Language Processing: Methodology Development and Applications. JMIR Med Inform. 2021 Oct 21;9(10):e23898. doi: 10.2196/23898. PMID: 34673533; PMCID: PMC8569540.

[13] https://www.snomed.org/?lang=fr

[14] https://www.ncbi.nlm.nih.gov/mesh/

[15] https://www.who.int/tools/atc-ddd-toolkit/atc-classification

[16] https://www.nlm.nih.gov/research/umls/index.html Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan

1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

[17] McCray, A. T.; Burgun, A. & Bodenreider, O.Aggregating UMLS semantic types for reducing conceptual complexity.Studies in health technology and informatics, 2001, 84, 216-220

[18] Tannier X, Wajsbürt P, Calliger A, Dura B, Mouchet A, Hilka M, Bey R. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. arXiv preprint arXiv:2303.13451. 2023 Mar 23.

[19] Bethard S. We need to talk about random seeds. arXiv preprint arXiv:2210.13393. 2022 Oct 24

[20] Tomas Mikolov, Wen-tau Yih, and GeoffreyZweig. 2013a. Linguistic regularities in continuous space word representations. InHu-man Language Technologies: Conference of theNorth American Chapter of the Association of Computational Linguistics, Proceedings, June9-14, 2013, Westin Peachtree Plaza Hotel, At-lanta, Georgia, USA.

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

[22] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017 Dec 1;5:135-46.

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

[24] Martin L., Muller B., Ortizsuárezp. J., Duponty., Romary L.,De La Clergerie É.,Seddahd. & Sagotb.(2020). CamemBERT : a tasty French language model. InProceedings ofthe 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online : Association for Computational Linguistics.

[25] Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D. Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372. 2019 Dec 11.

[26] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.

[27] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.

[28] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.

[29] Rian Touchent, Laurent Romary, Eric Villemonte de La Clergerie. CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. 2023. ⟨hal-04085419v2⟩

[30] Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, Gourraud PA. Drbert: A robust pre-trained model in french for biomedical and clinical domains. bioRxiv. 2023:2023-04.

[31] Dura B, Jean C, Tannier X, Calliger A, Bey R, Neuraz A, Flicoteaux R. Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. arXiv preprint arXiv:2207.12940. 2022 Jul 26.

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

[33] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, Yuan Luo. Journal of the American Medical Informatics Association, Volume 30, Issue 2, February 2023, Pages 340–347, https://doi.org/10.1093/jamia/ocac225

[34] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep 1;17(5):507-13.

[35] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", . June 2001.

[36] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] Giorgi,J.M. and Bader,G.D. (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 34, 4087.

[38] Habibi,M. et al. (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33, i37–i48.

[39] Kühnel, L., Fluck, J. We are not ready yet: limitations of state-of-the-art disease named entity recognizers. *J Biomed Semant* **13**, 26 (2022). https://doi.org/10.1186/s13326-022-00280-6

[40] Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, Li J, Wang G. Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428. 2023 Apr 20.

[41] Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, Su Y. Thinking about gpt-3 in-context learning for biomedical ie? think again. arXiv preprint arXiv:2203.08410. 2022 Mar 16.

[42] Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050. 2003 Jun 12.

[43] https://n2c2.dbmi.hms.harvard.edu/

[44] Yu J., Bohnetb. & Poesiom.(2020). Named Entity Recognition as Dependency Parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p.470–6476, Stroudsburg, PA, USA : Association for Computational Linguistics..

[45] Perceval Wajsbürt. Extraction and normalization of simple and structured entities in medical documents. Santé publique et épidémiologie. Sorbonne Université, 2021. English. ⟨NNT : 2021SORUS541⟩. ⟨tel-03624928v2⟩

[46] Bannour N, Wajsbürt P, Rance B, Tannier X, Névéol A. Privacy-preserving mimic models for clinical named entity recognition in French. Journal of Biomedical Informatics. 2022 Jun 1;130:104073.

[47] Kraljevic Z, Bean D, Mascio A, Roguski L, Folarin A, Roberts A, Bendayan R, Dobson R. MedCAT--medical concept annotation tool. arXiv preprint arXiv:1912.10166. 2019 Dec 18.

[48] Silvestri S, Gargiulo F, Ciampi M. Iterative Annotation of Biomedical NER Corpora with Deep Neural Networks and Knowledge Bases. Applied Sciences. 2022; 12(12):5775. https://doi.org/10.3390/app12125775

[49] Košprdić M, Prodanović N, Ljajić A, Bašaragin B, Milosevic N. From Zero to Hero: Harnessing Transformers for Biomedical Named Entity Recognition in Zero-and Few-Shot Contexts. Available at SSRN 4463335.

[50] Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, Li Y, Li J, Jiang X, Xu H. Zero-shot clinical entity recognition using chatgpt. arXiv preprint arXiv:2303.16416. 2023 Mar 29.

[51] Mahajan, D., Liang, J.J. and Tsou, C.H., 2020. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *arXiv preprint arXiv:2011.08835*.

[52] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34 (5):301–310

[53] Nath N, Lee SH, Lee I. NEAR: Named entity and attribute recognition of clinical concepts. Journal of Biomedical Informatics. 2022 Jun 1;130:104092.

[54] Weiyi Sun and others, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, Journal of the American Medical Informatics Association, Volume 20, Issue 5, September 2013, Pages 806–813

[55] Ramachandran GK, Lybarger K, Liu Y, Mahajan D, Liang JJ, Tsou CH, Yetisgen M, Uzuner Ö. Extracting medication changes in clinical narratives using pre-trained language models. Journal of Biomedical Informatics. 2023 Mar 1;139:104302.

[56] Wajsbürt, Perceval, Sarfati, Arnaud and Tannier, Xavier.(2021) “Medical concept normalization in French using multilingual terminologies and contextual embeddings.” Journal of Biomedical Informatics. Vol. 12, Issue 114.

[57] Z. Afzal, S. A. Akhondi, H. van Haagen, E. M. van Mulligen and J. A. Kors, Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms, in: CLEF 2015 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2015.

[58] D'Souza, Jennifer, and Vincent Ng. "Sieve-based entity linking for the biomedical domain." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015

[59] Robert Leaman and others, DNorm: disease name normalization with pairwise learning to rank, Bioinformatics, Volume 29, Issue 22, November 2013, Pages 2909–2917, https://doi.org/10.1093/bioinformatics/btt474

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

[61] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pages 1188–1196, Beijing, China.

[62] Adhikari A, Ram A, Tang R, Lin J. Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398. 2019 Apr 17.

[63] S. Gao et al., "Limitations of Transformers on Clinical Text Classification," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3596-3607, Sept. 2021, doi: 10.1109/JBHI.2021.3062322.

[64] Paaß, G., Giesselbach, S. (2023). Improving Pre-trained Language Models. In: Foundation Models for Natural Language Processing. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham. https://doi.org/10.1007/978-3-031-23190-2_3

[65] M. Zaheer et al. "Big Bird: Transformers for Longer Sequences". In: Adv. Neural Inf. Process.Syst. 33 (Jan. 8, 2021).

[66] Garcelon, N., Neuraz, A., Salomon, R. et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis 13, 85 (2018).

[67] Salton, G.; Buckley, C. (1988). "Term-weighting approaches in automatic text retrieval" (PDF). Information Processing & Management. 24 (5): 513–523. doi:10.1016/0306-4573(88)90021-0. hdl:1813/6721. S2CID 7725217.

[68] Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform A

[69] Van Aken, Betty, et al. "This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text." arXiv preprint arXiv:2210.08500 (2022).

[70] De Freitas JK, Johnson KW, Golden E, Nadkarni GN, Dudley JT, Bottinger EP, Glicksberg BS, Miotto R. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns. 2021 Sep 10;2(9):100337

[71] Miled ZB, Dexter PR, Grout RW, Boustani M. Feature engineering from medical notes: A case study of dementia detection. Heliyon. 2023 Mar 1;9(3).

[72] Gombar, S., Callahan, A., Califf, R. et al. It is time to learn from patients like mine. npj Digit. Med. 2, 16 (2019). https://doi.org/10.1038/s41746-019-0091-3

[73] Callahan A, Polony V, Posada JD, Banda JM, Gombar S, Shah NH. ACE: the Advanced Cohort Engine for searching longitudinal patient records. Journal of the American Medical Informatics Association. 2021 Jul;28(7):1468-79.

[74] Callahan A, Gombar S, Cahan EM, Jung K, Steinberg E, Polony V, Morse K, Tibshirani R, Hastie T, Harrington R, Shah NH. Using aggregate patient data at the bedside via an on-demand consultation service. NEJM Catalyst Innovations in Care Delivery. 2021 Sep 15;2(10).

[75] https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

[76] Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. Journal of biomedical informatics. 2021 Jan 1;113:103637.

[77] Singh, A., Schooley, B., Lindros, S.H., Brooks, J.M., Kissenberth, M., Pill, S., Faucher, G., Daly, C., Jeray, K. and Floyd, S.B., 2022. The Development of a Proof-of-Concept Physician-Driven Informatics Consult System for the Individualized Treatment of Patients with Orthopaedic Conditions.

[78] https://www.conseil-national.medecin.fr/sites/default/files/external-package/edition/od6gnt/cnomdata_algorithmes_ia_0.pdf

[79] Diagnostic Médical et Intelligence Artificielle : Enjeux Éthiques. Avis commun du CCNE et du CNPEN, Avis 141 du CCNE, Avis 4 du CNPEN. Novembre 2022

[80] Hovy, D. and Prabhumoye, S. (2021).Five sources of bias in natural language processing. Language and Linguistics Compass, 15(8) :e12432.

[81] Thomassin-Naggara I., Ceugnart, L., Tardivon L., Verzaux L., Balleyguier C. et al, "Artificial Intelligence: Place in Breast Cancer Screening in France," Cancer Bulletin 109, no 7 (July 1, 2022): 78085 https://doi.org/10.1016/j.bulcan.2022.04.008.

[82] Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, Mooney C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences. 2021 May 31;11(11):5088.

[83] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124-30.

[84] https://www.ohdsi.org/data-standardization/

[85] https://www.has-sante.fr/upload/docs/application/pdf/2017-03/dir1/pnds_-_lupus_systemique.pdf

[86] Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, Smolen JS, Wofsy D, Boumpas DT, Kamen DL, Jayne D, Cervera R, Costedoat-Chalumeau N, Diamond B, Gladman DD, Hahn B, Hiepe F, Jacobsen S, Khanna D, Lerstrøm K, Massarotti E, McCune J, Ruiz-Irastorza G, Sanchez-Guerrero J, Schneider M, Urowitz M, Bertsias G, Hoyer BF, Leuchten N, Tani C, Tedeschi SK, Touma Z, Schmajuk G, Anic B, Assan F, Chan TM, Clarke AE, Crow MK, Czirják L, Doria A, Graninger W, Halda-Kiss B, Hasni S, Izmirly PM, Jung M, Kumánovics G, Mariette X, Padjen I, Pego-Reigosa JM, Romero-Diaz J, Rúa-Figueroa Fernández Í, Seror R, Stummvoll GH, Tanaka Y, Tektonidou MG, Vasconcelos C, Vital EM, Wallace DJ, Yavuz S, Meroni PL, Fritzler MJ, Naden R, Dörner T, Johnson SR. 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheumatol. 2019 Sep;71(9):1400-1412. doi: 10.1002/art.40930. Epub 2019 Aug 6. PMID: 31385462; PMCID: PMC6827566.

[87] https://www.snfmi.org/content/antiphospholipides-syndrome-des-sapl Rédigé par Nathalie Morel, Véronique Le Guern et Nathalie Costedoat-Chalumeau, Service de Médecine Interne, Centre de Référence Maladies Auto-immunes et Maladies Systémiques Rares, Hôpital Cochin Port-Royal, Paris (mai 2014)

[88] https://www.snfmi.org/sites/default/files/uploads/pnds__sclerodermie_web.pdf

[89] https://www.snfmi.org/sites/default/files/uploads/pnds_takayasu_fair_-_favamulti.pdf

[90] https://huggingface.co/

[91] Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2014. 2014:24-30

[92] Grouin C., Grabar N. & Illouz G.(2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne deft 2021. In Actes de DEFT, Lille.

[93] Hiot, Nicolas, Minard, Anne-Lyse and Badin, Flora. "DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques". Actes de DEFT 2021, Lille, 2021.

[94] Kusner M. Sun Y. Kolkin N. & Weinberger K. (2015, June). From word embeddings to document distances. In International conference on machine learning (pp. 957-966). PMLR.

[95] Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". Kongelige Danske Videnskabernes Selskab. 5 (4): 1–34.

[96] Liu L, Bustamante R, Earles A, Demb J, Messer K, Gupta S. A strategy for validation of variables derived from large-scale electronic health record data. Journal of biomedical informatics. 2021 Sep 1;121:103879.

[97] Teixeira L, Mouthon L, Mahr A, Bérezné A, Agard C, Mehrenberger M, Noël LH, Trolliet P, Frances C, Cabane J, Guillevin L. Mortality and risk factors of scleroderma renal crisis: a French retrospective study of 50 patients. Annals of the rheumatic diseases. 2008 Jan 1;67(1):110-6.

[98] https://www.dataia.eu/index.php/appels-projets/appel-projets-recherche-doctorale-dim-ai4idf

[99] Gronier, G. & Baudet, A. (2021). Psychometric evaluation of the F-SUS: Creation and validation of the French version of the System Usability Scale. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2021.1898828.

[100] Lund, Arnold M. "Measuring usability with the use questionnaire12." *Usability interface* 8.2 (2001): 3-6.

[101] Marcilly, Romaric, et al. "Improving the usability and usefulness of computerized decision support systems for medication review by clinical pharmacists: A convergent, parallel evaluation." *Research in Social and Administrative Pharmacy* (2022).

[102] https://loinc.org/document-ontology/

[103] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2023-2024/wp-4-health_horizon-2023-2024_en.pdf

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

# 8. Annexes

## 8.1. Liste des abréviations

ACE : Advanced Cohort Engine

ACR : American College of Rheumatology

AIIM : Artificial Intelligence in Medicine

AP-HP : Assistance Pulbique- Hôpitaux de Paris

aPL : anticorps antiphospholipides

AREM Archive de l'historique des remontées mensuelles

AUC ROC : Area under the curve (Receiver Operating Characteristic)

ATC : Anatomic Therapeutic Chemical classification system

BERT : Bidirectional Encoder Representations from Transformers

CAPS : Syndrome catastrophique des antiphospholipides

CCAM : Classification commune des actes médicaux

CHU : Centre hospitalo-universitaire

CIM10 : Classification internationale des maladies (version 10)

CREST Syndrome : Calcinose, reflux, E, sclérose

CRF : Conditional random field

CSE : Comité scientifique et éthique

CUI : Concept Unique Identifier

DEFT : Défi Fouille de texte

DME : dossier médical électronique

EDS : Entrepôt de données de santé

EULAR : European League Against Rheumatism

FN : Faux négatifs

FP : Faux positifs

GB : Gigabytes

GEM : Glomérulonéphrite extra-membraneuse

GHM : Groupe homogène de maladies

GPT-3 : Generative Pre-trained Transformer 3

i2b2 : Integrating Biology & the Bedside

LES : lupus érythémateux systémique

LGM : Lésions glomérulaires minimes

LS : Lupus systémique

LSTM : Long-Short-Term-Memory

MeSH : Medical Subject Headings

NLP : Natural Language Processing

OMOP : Observational Medical Outcomes Partnership

OHDSI : Observational Health Data Sciences and Informatics

PMSI Programme de Médicalisation des Systèmes d'Information

PNDS Protocole national de diagnostic et de soins

SAPL : Syndrome des antiphospholipides

SNFMI : Société française médecine interne

SNOMED_CT : Sytematized Nomenclature of Medicine –Clinical Terms

SOSY : Signes et symptômes

TAL Traitement automatique des langues

TF-IDF : Term Frequency-Inverse Document Frequency

TQL : Temporal Query Language

UMLS : Unified Medical Language System

VN : Vrai négatif

VP: Vrai positif

VPP : Valeur prédictive positive

## 8.2. Glossaire

**ATC (Anatomic Therapeutic Chemical classification system) : "**Système de classification anatomique, thérapeutique et chimique international contrôlé par l'Organisation mondiale de la santé. Les médicaments sont divisés en groupes selon l'organe ou le système sur lequel ils agissent. [...].
La classification ATC repose sur cinq niveaux de classement qui correspondent aux organes (ou systèmes d'organes) cibles, et aux propriétés thérapeutiques, pharmacologiques et chimiques des différents produits. La forme générale du code d'une molécule est LCCLLCC, où L représente une

lettre et C un chiffre (exemple : A01AA01). Chaque lettre et chaque doublet de chiffres représente un niveau successif.
Le premier niveau (première lettre) définit le groupe anatomique parmi 14 différents. Le deuxième niveau (deux premiers chiffres) donne le sous-groupe pharmacologique ou thérapeutique principal. Les troisième et quatrième niveaux (deuxième et troisième lettres) correspondent à des sous-groupes chimiques, pharmacologiques ou thérapeutiques. Le cinquième et dernier niveau (deux derniers chiffres) indique la substance chimique." Source : Wikipédia.
Cette classification est traduite par l'Agence nationale de la sûreté du médicament (ANSM) pour sa version française officielle (https://smt.esante.gouv.fr/terminologie-atc/).

**AUC ROC** ***:*** Aire sous la courbe ROC (Receiver Operating Characteristic). La courbe ROC est une mesure de performance d'un classifieur binaire et correspond à la courbe sensibilité/spécificité. L'aire sous la courbe ROC permet d'interpréter cette mesure de performance et varie entre 0 et 1, elle est d'autant plus élevée que le classifieur est performant.

**Clustering hiérarchique ascendant*:*** Classification d'individus réalisée de manière non supervisée (sans annotation). Initialement, chaque individu forme une classe. Puis, de manière itérative, on cherche à réduire le nombre de classes. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes qui deviennent de plus en plus grandes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus « proches », en d'autres termes, celles dont la dissimilarité entre elles est minimale, cette valeur de dissimilarité est appelée indice d'agrégation.

**CNN (Convolutional neural network):** Un réseau de neurones convolutif est un type de réseau de neurones artificiels principalement utilisé pour traiter des images. L'idée principale d'un CNN est d'utiliser des filtres successifs (appelés filtres de convolution), "glissant" sur l'image, pour extraire des caractéristiques importantes d'une image. Ces caractéristiques permettent ensuite de réaliser des tâches spécifiques telles que la classification d'images. Un exemple de réseau de neurones convolutif est fourni figure 1 ci dessous.

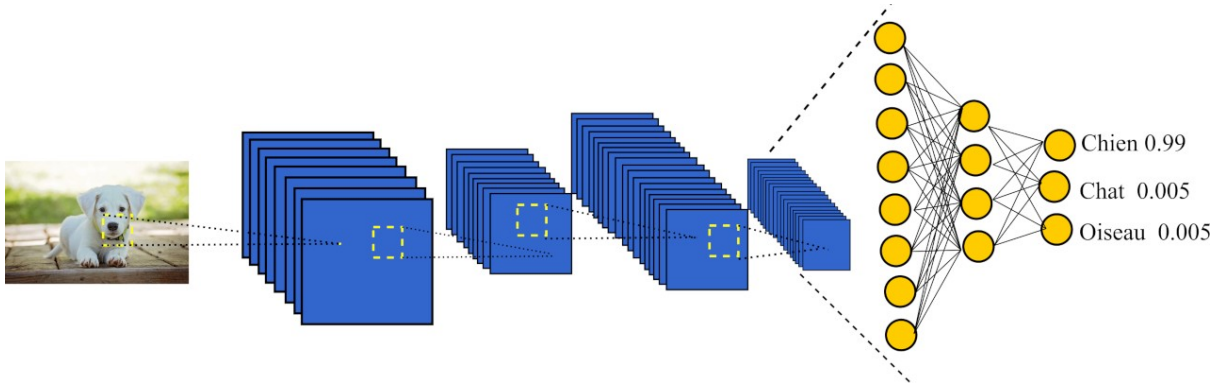



**Figure 1** : exemple de réseau de neurone convolutif. source : https://kongakura.fr/article/R%C3%A9seaux-de-neurones-convolutionnels

**CRF : Le modèle CRF (Conditional Random Field)** est un modèle probabiliste pour la modélisation de séquences. Son objectif est d'attribuer les meilleures étiquettes possibles à chaque élément de la séquence en tenant compte des relations entre les éléments adjacents (et donc, par exemple, un type sémantique à une mention textuelle à partir d'un jeu de données annoté). Il est appelé "random

field" car les étiquettes sont modélisées comme des variables aléatoires interdépendantes, où les valeurs d'une étiquette dépendent des valeurs des étiquettes voisines.
Le modèle CRF définit une fonction qui prend en compte des facteurs locaux et des facteurs globaux. Les facteurs locaux représentent les caractéristiques locales de chaque élément de la séquence, tandis que les facteurs globaux capturent les relations entre les éléments adjacents.
Le modèle CRF peut être entraîné à partir de données d'apprentissage en utilisant des algorithmes d'optimisation, tels que l'algorithme de descente de gradient. Pendant l'entraînement, les poids du modèle sont ajustés pour minimiser l'erreur entre les étiquettes prédites et les étiquettes réelles de l'ensemble d'apprentissage.

**Données structurées :** Dans le domaine des données de santé, on parle de données structurées lorsque ces données sont disponibles dans des tables que l'on peut requêter (tables des données de biologie, tables des données de médicaments etc..) avec des informations standardisées avec des codes (code d'un test de biologie, code ATC pour le médicaments, code CIM10 pour les données de remboursements etc..) et des valeurs numériques.

**Données non structurées :** Contrairement aux données structurées, les données non structurées ne sont pas présentes sous formes de codes et de valeurs numériques mais sous forme de texte libre. Leur traitement présente un défi important d'extraction et de normalisation de l'information.

**Intelligence artificielle** *:* "Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine." (Larousse)

**Machine learning** : méthodes d'apprentissage machine, c'est-à-dire approches basées sur des algorithmes qui, se fondant sur des données, construisent des modèles (arbres de décisions, réseaux de neurones, ou autres) qui seront ensuite appliqués à de nouvelles données pour calculer une décision à partir de celles-ci.

**K-means :** La méthode K-means est un algorithme de regroupement non supervisé utilisé pour partitionner un ensemble de données en K groupes distincts (clusters). L'objectif de l'algorithme est de minimiser la variance intra-cluster, c'est-à-dire la similarité des points à l'intérieur de chaque cluster. A titre d'exemple, il peut être utilisé pour la segmentation d'image, la classification de documents etc…
Cette méthode est très utilisée mais nécessite néanmoins de connaître le nombre de clusters K a priori, il peut par ailleurs converger vers différents résultats selon l'initialisation.

**LOINC :** Classification internationale d'identification des mesures (tests de biologies), observations et documents relatifs à la santé.

**LSTM (Long Short Term Memory [36] ) :** Le LSTM est une architecture de réseau de neurones qui, contrairement aux réseaux neuronaux feedforward standards (passage de l'information dans un seul sens uniquement) possède des connexions de rétroaction et fait partie de ce que l'on appelle un réseau neuronal récurrent (RNN). Il peut traiter non seulement des points de données isolés (tels que des images), mais aussi des séquences entières de données (telles que la parole ou la vidéo).

Les LSTM ont été développés pour pallier au défaut des réseau neuronaux récurrents simples (aussi appelé “vanilla RNN” et définis plus bas) qui ne présentent qu’une boucle simple de rétroaction et ne peuvent pas tenir compte de l’information “à long terme”. Par exemple, dans le traitement de texte, l’information contenu en début de phrase est perdue en fin de phrase.
L’architecture LSTM, plus complexe, composée d’une cellule mémoire et d’une porte d’oubli (en plus des portes d’entrées et de sorties classiques), permet de maintenir des dépendances utiles à long termes pour faire des prédictions à la fois actuelles et futures. (Définition adaptée de Wikipédia).

**OMOP (Observational Medical Outcomes Partnership) common data model :** modèle de données communautaire open-source, conçue pour normaliser la structure et le contenu des données d'observation médicale et pour permettre des analyses efficaces susceptibles de produire des preuves fiables (source :https://ohdsi.github.io/CommonDataModel/). Il a été développé et est entretenu par l’OHDSI (Observational Health Data Sciences and Informatics).

**PMSI (Programme de médicalisation des systèmes d’information) :** permet de décrire de façon synthétique et standardisée l’activité médicale des établissements de santé. Il repose sur l’enregistrement de données médico-administratives normalisées dans un recueil standard d’information. (source : ministère de la santé et de la prévention).

**RNN (Recurrent Neural Network) :** Un réseau neuronal récurrent (RNN) est une classe de réseaux neuronaux où les connexions entre les nœuds peuvent créer un cycle, permettant à la sortie de certains nœuds d'affecter l'entrée ultérieure des mêmes nœuds. Cela lui permet d'afficher un comportement dynamique temporel. Les RNN peuvent utiliser leur état interne (mémoire) pour traiter des séquences d'entrées de longueur variable (source Wikipédia). Un exemple de réseau de neurones récurrents simple est présenté figure 2.

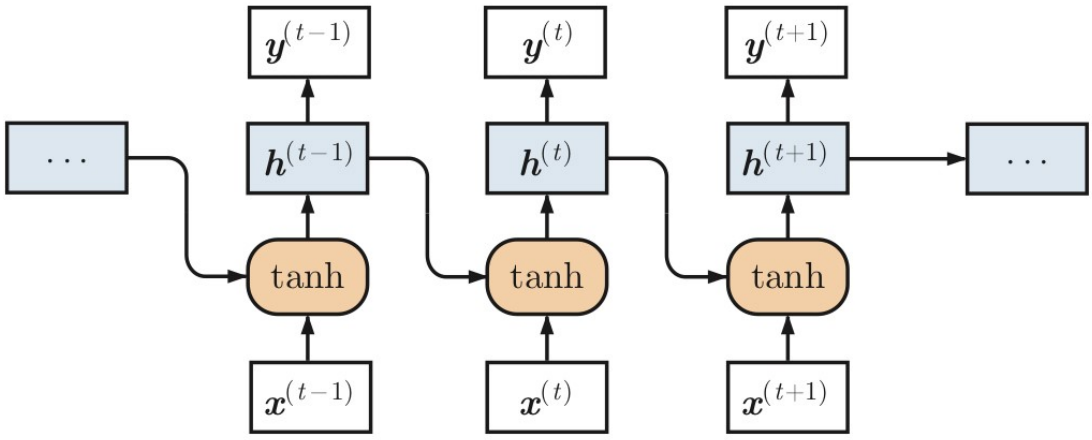



**Figure 2 : Exemple de réseau de neurone récurrent simple (ou “vanilla” en anglais).**
Source : https://blog.acolyer.org/2019/02/25/understanding-hidden-memories-of-recurrent-neural-networks

**Transformers :** Le transformer est une architecture de réseau de neurone complexe proposée par Vaswani et al [23]. Elle est constituée de l'empilement de couches dont certaines sont des couches d'auto-attention multiples (aussi appelées "tête d'attention"), comme présenté figure 3.

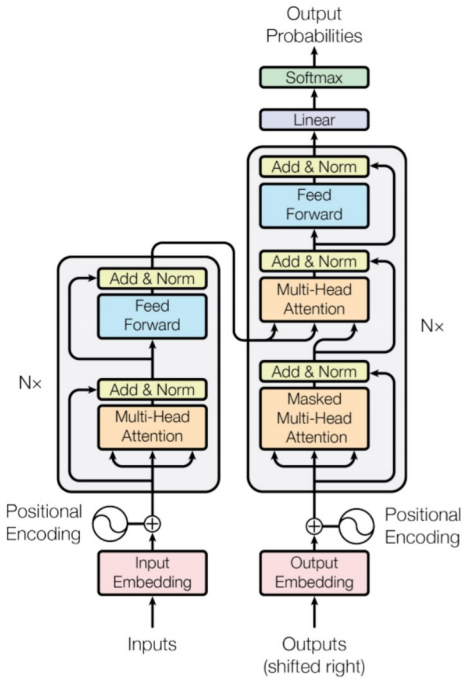



**Figure 3 : Architecture du réseau de neurone transformer [23]**

L'apport principal de cette architecture réside dans la couche d'auto-attention, qui correspond à des produits scalaires multiples entre la représentation vectorielle des mots et la représentation vectorielle des autres mots dans une même phrase, permettant ainsi d'obtenir une représentation du mot dépendante du contexte. L'utilisation de plusieurs "tête d'attention" permet de capturer des dépendances contextuelles à plus longue distance dans le texte.
Par ailleurs, ce mécanisme d'auto-attention nécessite un temps d'entraînement beaucoup plus court que les anciens modèles de mémoire à long terme (LSTM)[36] permettant ainsi d'utiliser de grands corpus de données (linguistiques), tels que le corpus Wikipédia.
Ce modèle a conduit au développement de systèmes pré-entraînés, tels que le GPT (generative pre-trained transformer)[5] et le BERT[4] (Bidirectional Encoder Representations from Transformers).

**Softmax : "**la fonction softmax, aussi appelée fonction softargmax ou fonction exponentielle normalisée, est une généralisation de la fonction logistique. Elle convertit un vecteur de *K* nombres réels en une distribution de probabilités sur *K* choix. Plus précisément, un vecteur z = ( z 1 , … , z K ) est transformé un vecteur σ ( z ) de *K* nombres réels strictement positifs et de somme 1. La fonction est définie par :

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$ pour tout j ∈ { 1 , … , K }". (source : Wikipédia).

**UMLS (Unified Medical Language System) [16] :** "ensemble de fichiers et de logiciels qui rassemble de nombreux vocabulaires et normes de santé et de biomédecine afin de permettre l'interopérabilité entre les systèmes informatiques." (source https://www.nlm.nih.gov/research/umls/index.html). Il

contient plus d'une centaine de vocabulaires (dont par exemple le MeSH [14] ou la SNOMED-CT[12] etc..), dans une trentaine de langues différentes.

**z-score :** le z-score ou la cote Z correspond au nombre d'écarts types séparant un résultat de la moyenne. Il se calcule de la même façon qu'une variable centrée réduite : écart à la moyenne divisé ensuite par l'écart type.

## 8.3. Disponibilité des codes

Les modèles de langage BERT [4], entraînés ou affinés sur l'espace-projet mis à disposition par l'EDS et présentés ici contiennent potentiellement des informations patients identifiantes et ne sont donc pas partageables publiquement. Néanmoins, les modèles peuvent être accessibles sur demande auprès du comité scientifique et éthique pour les espaces projets.

Par ailleurs, l'ensemble des architectures des modèles utilisés sont disponibles et l'ensemble des expérimentations également sur github :

Pour l'article d'extraction de document (article 2): https://github.com/ChristelDG/EHR_2_vec

Pour l'article de traduction (article 3) : https://github.com/ChristelDG/biomed_translation

Pour l'article de construction de cohortes de patients similaires (article 5) - comprenant le classifieur (article 4):

https://github.com/ChristelDG/cohort-similarity

Pour l'article 1 en collaboration avec l'équipe "Science des données" : https://github.com/aphp/edspdf