Identification des patients dans un Entrepôt de Données de Santé Hospitalier : impact de la détection de contexte sur l'extraction des concepts médicaux

Matisse Decilap

U.F.R. DES SCIENCES MEDICALES

Année 2025 Thèse n°3022

THESE POUR L'OBTENTION DU

DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

Par Matisse DECILAP

Né le 07/10/1996 à Orléans

Présentée et soutenue publiquement le 18/04/2025

IDENTIFICATION DES PATIENTS DANS UN ENTREPOT DE DONNEES DE SANTE HOSPITALIER : IMPACT DE LA DETECTION DE CONTEXTE SUR L'EXTRACTION DES CONCEPTS MEDICAUX

Sous la direction du Dr. Vianney JOUHET

Membres du jury :

Dr. GRIFFIER, Romain	Examinateur
Pr. MOUGIN, Fleur	Examinateur
Pr. RICHERT, Laura	Examinateur
Pr. THIEBAUT, Rodolphe	Président

Remerciements

Au Dr. Vianney JOUHET

Merci d'avoir été mon directeur de thèse. Cela a été une vraie chance de pouvoir bénéficier de la richesse de tes connaissances et compétences techniques, toujours transmises avec la bienveillance qui te caractérise. Ce que j'ai appris à tes côtés m'est déjà précieux aujourd'hui, et le sera encore davantage dans ma vie professionnelle à venir, comme sur le plan personnel.

Au Pr. Rodolphe THIEBAUT

C'est un honneur de t'avoir comme président de jury. Je te suis reconnaissant pour tes enseignements et la confiance que tu m'as témoignée, en m'ouvrant notamment la porte vers de futures opportunités.

Au Pr. Laura RICHERT

Merci pour ta présence dans mon jury. Je te remercie également pour l'accueil au sein de l'unité de méthodologie, qui m'a beaucoup fait progresser dans le champ de la recherche clinique.

Au Pr. Fleur MOUGIN

Merci d'avoir accepté de rapporter ce travail. Je tiens à t'exprimer ma gratitude pour l'enseignement que tu m'as transmis durant le master SITIS, sans lequel ce travail n'aurait tout simplement pas pu exister.

Au Dr. Romain GRIFFIER

C'est un honneur de t'avoir dans mon jury. Merci pour tes enseignements, sans lesquels ce travail n'aurait jamais pu voir le jour. Ton ardeur et ton sens de la transmission, qui n'ont d'égal que ton humour, font de toi un exemple qui m'inspire.

Aux membres des équipes d'IAM, de l'USMR et du COREVIH

Merci de m'avoir accueilli et soutenu tout au long de ce travail. J'exprime toute ma gratitude aux annotateurs, Alexandre, les deux Antoine, Camille, Guillaume, Lucas et Sullivan qui ont consacré de nombreuses heures pour rendre ce projet possible.

Au Pr. Thierry SCHAEVERBEKE et au Dr. Kevin SALLES

Merci de m'avoir permis de m'appuyer sur l'étude ArthroVIH. La réutilisation de vos travaux a été précieuse pour nourrir cette réflexion et ancrer ce projet dans une réalité clinique.

Aux encadrants universitaires, hospitaliers, et à tous les professeurs qui m'ont accompagné tout au long de ma scolarité

Merci pour vos enseignements et votre accompagnement, de la petite école au lycée, puis lors de mon externat en Touraine et jusqu'à mon internat à Bordeaux. Vous avez toutes et tous, à votre manière, contribué à mon parcours.

A mes amis

Merci à tous mes amis de lycée, pour leur amitié sans faille depuis l'adolescence. Et non, mes études ne sont toujours pas terminées. Une dédicace toute particulière à mon ami Polo, pour nos appels et le partage d'histoires de vie dans nos hôpitaux respectifs mais également pour ton aide et tes conseils avisés.

Merci à mes amis de la fac de Tours, et tout particulièrement à ceux de la Fanfare, pour avoir fait de mes années d'études de médecine des années de joie, de pouets et de rires. Je pense notamment à vous, Célia et Gatien, qui avez affûté à la fois mon sens critique… et mon humour.

Merci à tous mes co-internes, pour les moments partagés en stage, à l'internat, au Tondu, et lors des soirées organisées ou improvisées. Et un merci particulier à Nathan, dont les travaux m'ont inspiré pour la concrétisation de mon travail, tant écrit qu'oral.

A Polochon

Cousin, colocataire, compagnon de route… Tu as suivi cette thèse au plus près, du début à la fin. Aujourd'hui, on prend des chemins différents, mais je sais qu'on restera connectés. À très vite. J'ai hâte de lire nos correspondances.

A Line

Merci de m'avoir soutenu tout au long de ce travail. J'ai écrit cette thèse un peu partout, mais toujours avec toi à mes côtés. Tes conseils, ton écoute et ta présence ont compté plus que tu ne l'imagines.

A ma famille

Merci à Clément, Hugo et Tom, pour vos rôles de frères. Vos parcours, si différents, sont pour moi des sources d'inspiration. Vous êtes mes modèles, chacun à votre façon, et j'ai beaucoup appris à vous observer, à vous écouter, à partager avec vous.

Merci à mes parents, pour votre amour, vos encouragements constants et votre soutien indéfectible. Vous avez toujours été là, dans les moments simples comme dans les étapes décisives. Si cette thèse existe aujourd'hui, c'est d'abord grâce à vous.

Les honneurs aujourd'hui vous reviennent.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

AVC : Accident Vasculaire Cérébral

CIM-10 : Classification Internationale des Maladies, 10^{e} Révision

CNIL : Commission Nationale de l'Informatique et des Libertés

CNN : Convolutional Neural Network (Réseau de Neurones Convolutifs)

EDS : Entrepôt de Données de Santé

EDSH : Entrepôt de Données de Santé Hospitalier

ETL : Extract, Transform, Load

HTAP : Hypertension Artérielle Pulmonaire

i2b2 : Informatics for Integrating Biology & the Bedside

LLM : Large Language Model (Modèle de Langage de Grande Taille)

LSTM : Long Short-Term Memory (Mémoire à Long Terme)

NER : Named Entity Recognition (Reconnaissance d'Entités Nommées)

OHDSI : Observational Health Data Sciences and Informatics

OMOP : Observational Medical Outcomes Partnership

PMSI : Programme de Médicalisation des Systèmes d'Information

RGPD : Règlement Général sur la Protection des Données

SIH : Système d'Information Hospitalier

TAL : Traitement Automatique du Langage

TDM : Tomodensitométrie

UMLS : Unified Medical Language System

1 Introduction

1.1 Utilisation secondaire des données et entrepôts de données de santé hospitaliers

L'exploitation des données de santé a connu un essor considérable avec l'informatisation des Systèmes d'Information Hospitaliers (SIH). Ces derniers centralisent l'ensemble des données de soins produites lors de la prise en charge des patients. Ces données incluent des informations administratives (identité, date de naissance, sexe…) ainsi que des données médicales (constantes vitales, examens cliniques, imagerie, biologie, prescriptions, comptes rendus, etc.). Elles sont collectées à partir de multiples logiciels qui forment l'architecture du SIH (1). Ces données sont essentiellement produites pour le soin, et leur usage principal est de garantir une prise en charge optimale des patients : on parle d'utilisation primaire des données de santé. Elles retracent l'histoire complète du patient, de ses symptômes initiaux jusqu'aux examens complémentaires et aux traitements administrés. Toutefois, ces données ont un potentiel d'exploitation bien au-delà de leur usage clinique immédiat. L'utilisation secondaire des données de santé désigne leur exploitation dans un but autre que la prise en charge du patient (2). Cette réutilisation couvre un large spectre d'activités, incluant :

- Le pilotage des établissements de santé,
- La recherche clinique et épidémiologique,
- L'évaluation de la qualité et de la sécurité des soins,
- L'aide à la constitution de cohortes et bases clinico-biologiques,
- Le développement d'algorithmes d'intelligence artificielle en santé (3),
- La tarification et l'optimisation des ressources hospitalières.

Cependant, plusieurs verrous freinent cette exploitation secondaire (4,5). Ils se classent en deux grandes catégories : Les verrous réglementaires et éthiques, car les données de santé sont des données sensibles régies par le Règlement Général sur la Protection des Données (RGPD) (6) et la Loi Informatique et Libertés (7) sous le contrôle de la Commission Nationale de l'Informatique et des Libertés (CNIL). L'accès aux données doit respecter des cadres stricts pour garantir la confidentialité et le secret médical. Les verrous techniques, liés à l'architecture même des SIH. Ces derniers ne sont pas conçus pour des interrogations massives et croisées sur un grand nombre de patients (2,8). Par exemple, si l'on cherche à identifier tous les patients jeunes vivant avec le VIH présentant des douleurs articulaires avec atteintes radiologiques et une CRP élevée en biologie, cela nécessite de croiser plusieurs sources de données :

- Les données administratives pour l'âge,
- Les comptes rendus médicaux pour repérer les mentions de VIH et de douleurs articulaires dans les consultations,
- Les prescriptions médicamenteuses pour éventuellement confirmer le diagnostic de VIH,
- Les comptes rendus d'imagerie pour détecter les atteintes articulaires (érosions, pincement articulaire, ostéonécrose…),

- Les résultats biologiques pour repérer une CRP élevée.

Or, chaque logiciel métier (imagerie, biologie, dossier patient informatisé) possède sa propre structure de données, et le SIH n'est pas conçu pour faciliter des interrogations transversales. De plus, l'hétérogénéité syntaxique et sémantique des données complexifie ces analyses : L'hétérogénéité syntaxique correspond à la multiplicité des formats et langages d'interrogation (bases de données relationnelles, formats propriétaires, etc.). L'hétérogénéité sémantique est le fait qu'une même information peut être stockée de plusieurs manières : en texte libre, dans un formulaire structuré ou encore via un codage du Programme de Médicalisation des Systèmes d'Information (PMSI). Par exemple, la notion de VIH peut être décrite dans un compte rendu médical, enregistrée sous forme de code CIM-10 ou notée comme une observation clinique structurée.

C'est ici que l'Entrepôt de Données de Santé Hospitalier (EDSH) apparait comme une solution pour lever ces verrous techniques grâce à la mise en place d'une infrastructure dédiée. Celui-ci permet de désiloter les données de santé en les centralisant dans une structure commune et interrogeable massivement. Un EDSH repose sur un processus ETL (Extract, Transform, Load) (2) :

- Extract (Extraction des données) : Sélection et récupération des données pertinentes depuis les logiciels métiers (dossiers patients, comptes rendus médicaux, résultats biologiques, prescriptions…).
- Transform (Transformation des données) : Harmonisation des formats pour réduire l'hétérogénéité syntaxique, pseudonymisation des données pour renforcer la protection de la vie privée, remodélisation de la base pour optimiser les requêtes complexes.
- Load (Chargement des données) : Intégration dans l'EDSH des données transformées pour leur exploitation.

L'EDS présente plusieurs caractéristiques essentielles :

- Données répliquées : L'EDS contient une copie des données de soins, sans impact sur les bases de production, avec une actualisation régulière.
- Données désilotées : Les informations sont réunies en une base unique permettant une interrogation transversale et multi-sources sur des données massives.
- Intégration sélective des données : Seules les données pertinentes sont intégrées pour optimiser les performances, ce qui peut toutefois limiter certains usages secondaires.
- Réduction de l'hétérogénéité syntaxique grâce à l'utilisation d'un format commun, bien que l'hétérogénéité sémantique persiste.
- Meilleure valorisation des données : La centralisation facilite leur accès et leur exploitation, bien qu'elles restent des données de vie réelle et donc imparfaites.

Un exemple d'EDSH est celui du CHU de Bordeaux développé depuis 2018, alimenté quotidiennement grâce à un ETL basé sur le logiciel Talend (9). Cet entrepôt contient les données de plus de 2,5 millions de patients et 20 millions de séjours, représentant près de 3,5 milliards d'observations dont 72 millions de documents textuels entre 2005 et aout 2024 (10). Celui-ci est basé sur un modèle open source en étoile de type I2B2 pour Informatics for

Integrating Biology & the Bedside développé par l'université d'Harvard (11). L'EDSH du CHU de Bordeaux est également composé d'un datamart, c'est-à-dire un sous-ensemble de l'EDSH, depuis 2021 sous le format OMOP, permettant de participer à des réseaux fédérés via la communauté OHDSI (12).

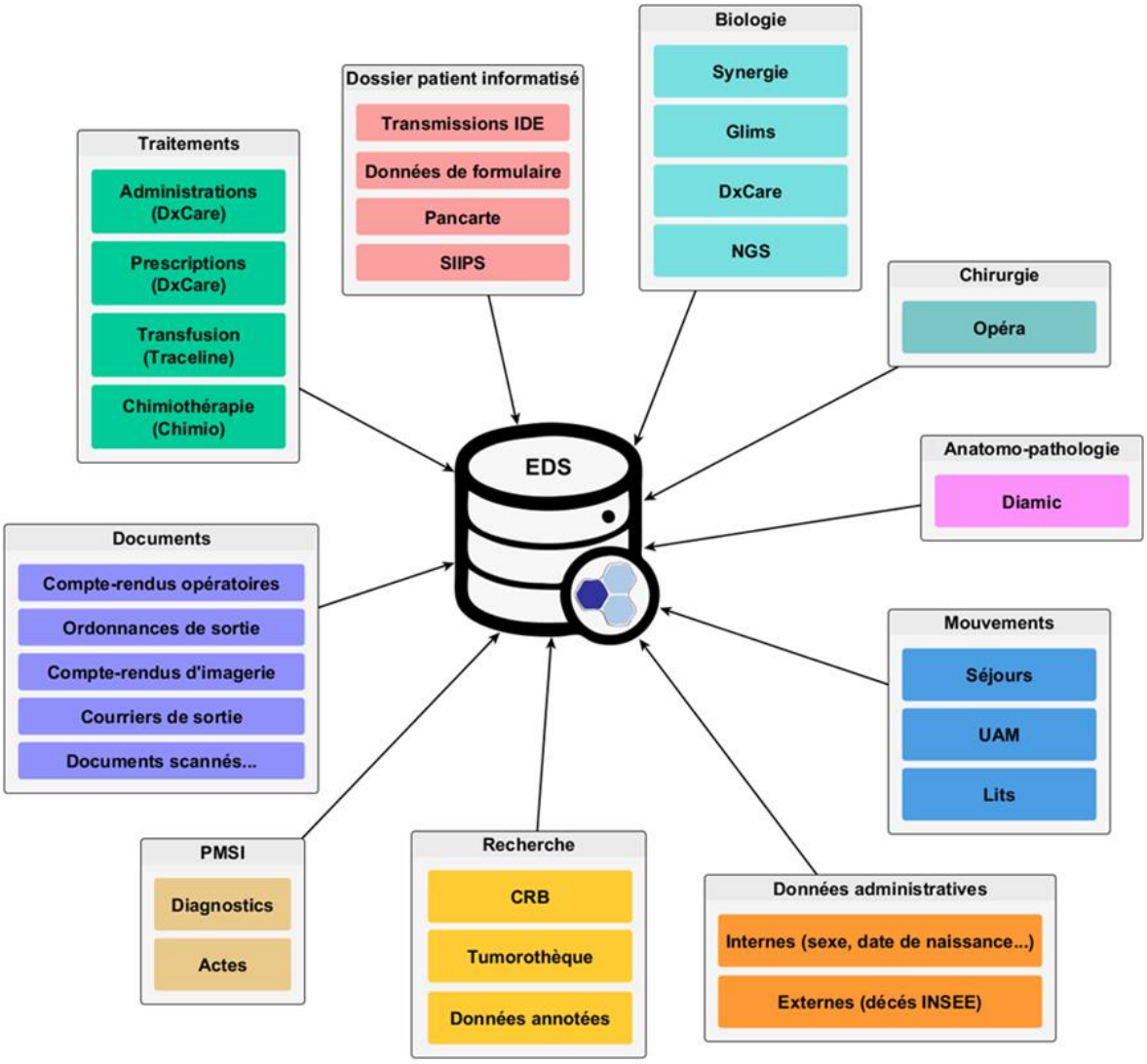


Figure 1 : Sources de données de l'EDS du CHU de Bordeaux (Source : Présentation de l'EDS par Vianney Jouhet, mars 2023)

1.2 Cas d'usages et importance de la détection de contexte dans les données non structurées

L'exploitation des données non structurées en santé, issues principalement des comptes rendus médicaux, des courriers, des observations cliniques et des notes de suivi, représente un défi majeur pour leur réutilisation dans le cadre de la recherche et de l'optimisation des soins. Contrairement aux données structurées, directement exploitables dans des bases relationnelles, les données textuelles sont riches en informations cliniques mais restent difficiles à analyser en raison de leur grande hétérogénéité syntaxique et sémantique.

Dans ce contexte, plusieurs cas d'usage des données non structurées dans les EDSH peuvent être identifiés, le principal décrit dans la littérature étant le phénotypage. Le phénotypage est le processus qui permet d'identifier des patients correspondant à des critères précis pour constituer des cohortes ou des indicateurs. Par exemple, Grosjean et al. ont exploité ces données pour repérer des patients vulnérables dans le cadre de la stratégie vaccinale contre la COVID-19 (13). De manière plus large, les travaux de Garcelon ont permis la création de plus d'une centaine de cohortes en moins d'un an (14). Une autre étude récente a montré qu'il était possible de reconstituer des cohortes issues de registres à partir des EDSH (15). Plusieurs études montrent l'intérêt des données non structurées pour le recrutement de patient (16). Une étude réalisée en 2014 essayait en effet de démontrer le caractère indispensable de ce type de données pour le recrutement dans les essais cliniques (17). Leurs résultats montrent que 59% à 77% des critères de recrutement étaient présents dans les données non structurées. Ils soulignent par ailleurs l'importance d'une utilisation conjointe des données structurées et non structurées pour pouvoir améliorer la capacité de recueil de l'information (16). Une autre étude a révélé que 79,3 % des concepts médicaux représentant les comorbidités des patients n'étaient détectables que dans le texte libre (18).

Pour mener ce type d'étude, l'utilisation d'algorithmes prenant en compte le contexte est indispensable afin d'améliorer la pertinence des résultats. Plusieurs recherches ont utilisé ces approches, comme celle de Salmasian et al. sur la détection de la surconsommation de médicaments (19), celle de Mendonça et al. sur les pneumonies pédiatriques (20) ou encore celle de Shiner et al. sur les chutes des sujets âgés (21). On note également l'étude de Garcelon qui utilise la détection de contexte pour faire des algorithmes de similarités de patients (22). Ces travaux illustrent la diversité des usages des données non structurées, en particulier lorsqu'elles sont enrichies par la prise en compte du contexte.

La détection du contexte est essentielle pour éviter des erreurs d'interprétation dans l'analyse des données médicales. Une mauvaise prise en compte du contexte peut conduire à des erreurs d'analyse et de classification des patients. Par exemple, un algorithme ne tenant pas compte de la négation pourrait interpréter à tort la phrase "Il n'y a pas de douleur articulaire" comme une affirmation de la présence d'une pathologie articulaire. De même, la mention "suspicion de VIH" ne signifie pas un diagnostic certain. Il est également crucial de distinguer les relations entre concepts médicaux. Par exemple, un patient mentionnant "Antécédent familial de VIH chez le père" ne doit pas être classé à tort comme porteur du VIH. L'analyse contextuelle permet d'éviter ces erreurs et d'assurer une classification plus précise des informations médicales.

La détection du contexte reste néanmoins une tâche complexe. En effet, les données recueillies de cette manière ne sont pas parfaites et leur qualité est variable en fonction de la nature des données (23,24). Aussi, la complexité de la langue et l'absence d'une approche standardisée compliquent encore davantage cette problématique. Prenons l'exemple des marqueurs de négation : ceux-ci peuvent se manifester sous forme de préfixes, tels que an-, in-, im-, ir- ou dis-, qui inversent le sens d'un mot. Cependant, la simple identification de ces marqueurs ne suffit pas (25). Il est également crucial de calculer leur portée, c'est-à-dire de déterminer l'étendue de leur influence sur la phrase. Cette portée peut être orientée vers la droite, la gauche, s'étendre des deux côtés, être discontinue ou encore chevauchante. Par ailleurs, une négation peut en annuler une autre, rendant l'analyse encore plus complexe (25). À ce jour, nous n'avons pas identifié d'étude internationale intégrant pleinement la prise en compte du contexte dans l'analyse de données non structurées multilingues, notamment dans le cadre de réseaux fédérés comme OHDSI. L'optimisation des méthodes de détection du contexte constitue donc un enjeu majeur pour l'exploitation secondaire des données de santé non structurées.

1.3 Présentation de l'étude ArthroVIH

L'étude ArthroVIH, menée par Kévin Salle et al. en 2023 (26), a été réalisée au CHU de Bordeaux pour explorer les douleurs articulaires des mains chez les patients vivant avec le VIH, avec un focus sur l'atteinte structurale des articulations métacarpophalangiennes. Elle repose sur une description épidémiologique des causes de ces douleurs et à en comprendre les mécanismes sous-jacents, en particulier leur origine dégénérative ou métabolique en lien avec l'infection VIH et les traitements antirétroviraux. Les résultats ont montré que l'arthrose était la principale cause d'arthralgies, tandis que les maladies auto-immunes restaient rares, à des niveaux similaires à ceux de la population générale. De plus, l'étude a mis en évidence une association entre l'atteinte des articulations métacarpophalangiennes et une immunodépression plus marquée, une exposition prolongée aux antiprotéases et une présence accrue d'autres atteintes articulaires, suggérant un rôle potentiel du VIH et des traitements antirétroviraux dans ces manifestations articulaires.

L'identification des patients et donc la faisabilité de l'étude a été permise grâce à l'EDSH du CHU de Bordeaux, en combinant une recherche par mots-clés dans les données non structurées et l'utilisation des codes CIM-10 pour le VIH. Cette extraction a initialement identifié 2 324 patients répondant aux critères de recherche. Toutefois, après vérification manuelle, seuls 706 patients remplissaient les critères d'inclusion de l'étude, soulignant une proportion importante de faux positifs due aux limites de la recherche automatique via mots-clés. Ainsi, bien que l'EDSH ait permis de constituer la cohorte de l'étude, son utilisation a nécessité un tri manuel conséquent, tâche fastidieuse et à faible valeur ajoutée. L'hypothèse que nous portons est que l'ajout d'une détection du contexte les données extraites permettrait d'affiner la sélection des patients, en minimisant le bruit et ainsi la charge de validation manuelle, grâce à des méthodes de traitement automatique du langage naturel.

1.4 Traitement du langage naturel dans le cadre de la détection de contexte

Le traitement automatique du langage naturel (TAL), ou Natural Language Processing (NLP), vise à développer des programmes capables de traiter automatiquement les langues naturelles. Le TAL englobe diverses applications telles que la classification de texte, la traduction, la génération automatique de texte, l'analyse de sentiments et la reconnaissance d'entités nommées (27). Dans le cadre de l'analyse des données médicales non structurées, le TAL joue un rôle essentiel pour identifier les concepts médicaux, contextualiser leur signification et limiter les erreurs d'interprétation (25).

La reconnaissance des entités nommées ou Named Entity Recognition (NER) consiste à identifier et classifier des éléments spécifiques dans un texte, comme des maladies, des médicaments, des symptômes, des dates ou des valeurs biologiques. Toutefois, la simple détection d'une entité ne suffit pas : il est impératif de la contextualiser pour éviter les erreurs d'interprétation.

Pour la détection du contexte, notamment des négations et autres modalités linguistiques influençant la signification des entités médicales, plusieurs approches ont été développées. On retrouve principalement trois types de méthodes : Les approches basées sur des règles : Le premier exemple populaire dans le domaine médical est l'algorithme *NegEx*, développé en 2001 (28) suivi de son extension *ConText* en 2009 (29) pour la langue anglaise. « NegEx est l'un des premiers systèmes permettant de détecter le contexte des conditions cliniques pour la langue anglaise. Il utilise des expressions régulières et des indicateurs lexicaux (modificateurs, pseudo-modificateurs et termes de terminaison). L'idée de base est de considérer une condition comme affirmée par défaut et de la marquer comme niée si elle apparaît sous la portée d'un modificateur ("aucun signe", "absence de", "est écarté", etc.). Outre sa simplicité, cet algorithme est rapide et efficace. » (30). L'adaptation française de *NegEx* est apparue en 2012 (31), suivie de celle de *ConText* en 2017(30,32). D'autres algorithmes combinent indicateurs lexicaux et règles grammaticales élaborées manuellement (33,34). En parallèle, Garcelon et al. ont proposé un autre modèle en 2016 (35) ; Les approches basées sur l'apprentissage supervisé(25). Celles-ci reposent sur l'annotation de données, l'extraction de caractéristiques linguistiques et l'entraînement de modèles comme les machines à vecteurs de support ou les forêts aléatoires (36,37). Des bases de données telles que *i2b2, MedNLI* et *BioScope* (38–40) sont utilisées pour entrainer ces modèles. Toutefois, ces approches souffrent de plusieurs limites : elles nécessitent un grand volume de données annotées de qualité, leur évaluation est souvent restreinte à l'échelle de la phrase et non du document entier, et certaines études montrent que des algorithmes à base de règles comme *NegEx* reste plus performant que certaines méthodes supervisées (41). Pour pallier ces limites, l'apprentissage profond a été exploré, notamment avec les réseaux de neurones convolutifs (CNN) et de mémoires à long terme (LSTM). Bien que prometteurs, ces modèles exigent d'importants volumes de données étiquetées et peuvent sous-performer sur des ensembles de données de petite taille ou déséquilibrés (42) ; Les approches basées sur les modèles de langage de grande taille ou Large Language Model (LLM). Entraînés sur de vastes corpus, ces modèles captent des relations complexes entre les mots et adaptent leur compréhension à

divers contextes. Cela permet une meilleure généralisation face aux variations linguistiques et une prise en compte du contexte global d'un document de manière plus performante (42). De plus, ils permettent un apprentissage par transfert (fine-tuning) sur des corpus spécialisés (43). On note par exemple des modèles basé sur BERT : Biomedical BERT (44), ClinicalBERT (45), CamembertBERT (46) ou basé sur GPT-3 et LLaMA2 (42). Cependant, ces approches posent plusieurs défis, notamment en termes d'interprétabilité, reproductibilité, de biais potentiels et de nécessité d'une puissance de calcul élevée que nous ne développerons pas ici.

1.5 Défis de la détection de contexte au sein d'un EDSH

L'application des méthodes d'extraction de concepts et de détection de contexte dans un EDSH présente plusieurs défis techniques et méthodologiques. Tout d'abord, le volume massif de documents à traiter, pouvant atteindre plusieurs dizaines de millions, exige des solutions capables de traiter efficacement ces données en un temps raisonnable, notamment pour une utilisation quotidienne. L'optimisation des algorithmes et l'utilisation de méthodes parallélisées sont essentiels pour garantir une extraction rapide et fiable du contexte des entités médicales. En lien également avec le volume de données, la question du stockage des résultats est importante. L'extraction et l'annotation des entités nommées et de leur contexte génèrent une quantité importante de métadonnées qu'il convient de stocker pour une interrogation rapide. Un autre enjeu important concerne l'impact carbone lié au traitement et au stockage des données. Face aux préoccupations environnementales croissantes, il devient essentiel d'optimiser la consommation énergétique des infrastructures informatiques utilisées (47). On note d'ailleurs de plus en plus d'études mesurer l'impact carbone de leurs algorithmes (15,48). Enfin, la définition même du « contexte » constitue un défi fondamental. Il est crucial de clarifier ce que recouvre cette notion en milieu médical afin que les utilisateurs puissent interpréter correctement les résultats et exploiter efficacement les données dans un moteur de recherche ou un système d'aide à la décision.

1.6 Objectifs

Cette étude vise à évaluer l'impact des algorithmes de détection de contexte sur le nombre de concepts identifiés chez les patients présents dans l'Entrepôt de Données de Santé Hospitalier (EDSH) du CHU de Bordeaux, en appliquant différentes stratégies de filtrage contextuel.

Les objectifs secondaires étaient d'évaluer les performances de deux bibliothèques de détection de contexte afin d'évaluer leur efficacité et leur précision et analyser leur capacité de filtration en vie réelle en réutilisant une étude projet permise par l'EDSH du CHU de Bordeaux, afin d'en mesurer les implications sur la qualité des données extraites et leur potentiel d'application en recherche clinique.

2 Matériel et méthodes

Pour répondre à notre objectif nous avons construit une chaine de traitement afin d'évaluer l'impact de la prise en compte du contexte dans les documents médicaux. La figure 2 présente les différentes étapes de conception de cette chaine que nous détailleront par la suite.

Figure 2 : Chaine de traitement de TAL dans le cadre de la détection de conteste (Source : GitHub EDS-NLP)

2.1 Extraction des données non structurées de l'EDSH du CHU de Bordeaux

Pour évaluer l'impact des algorithmes de détection de contexte au sein de l'EDSH du CHU de Bordeaux, nous avons réalisé deux extractions distinctes. Une extraction à grande échelle, visant à mesurer l'impact et les performances des algorithmes sur un volume massif de données issues de l'EDSH. Une extraction ciblée, destinée à évaluer leur impact et performances en conditions réelles dans une cohorte clinique : la cohorte ArthroVIH.

La première extraction repose sur un grand échantillon de l'EDSH, comprenant les données textuelles d'environ 15 000 patients tirés aléatoirement ayant eu au moins une consultation ou hospitalisation en 2023 pour un objectif d'environ 1 millions de documents. Cette sélection a été réalisée à partir de la base relationnelle i2b2, représentée par le modèle relationnel présenté en Figure 3. Les documents textuels associés à ces patients ont été extraits depuis la table *observation_fact*, où le contenu est stocké dans la variable *observation_blob*. Ces documents incluent des rapports d'hospitalisation, des comptes rendus opératoires, des lettres de consultation, des courriers médicaux. Les documents ont été préalablement pseudonymisés via des algorithmes internes exécutés en routine lors du processus ETL (Extract, Transform, Load). Cette pseudonymisation consiste à supprimer toutes les données directement identifiantes (nom, prénom, etc.) tout en conservant le reste du corps des documents. À partir de ce grand échantillon, nous avons extrait un jeu de données d'exploration constitué de 10 000 documents sélectionnés aléatoirement (maximum un par patient). Ce jeu a été utilisé pour développer les chaines de traitements, mesurer les temps d'exécution et servir de base à la construction d'un gold standard. Le gold standard, composé de 100 documents issus du jeu d'exploration, a été annoté manuellement selon la méthodologie détaillée en section 2.3. Il constitue la référence permettant l'évaluation des performances des algorithmes.

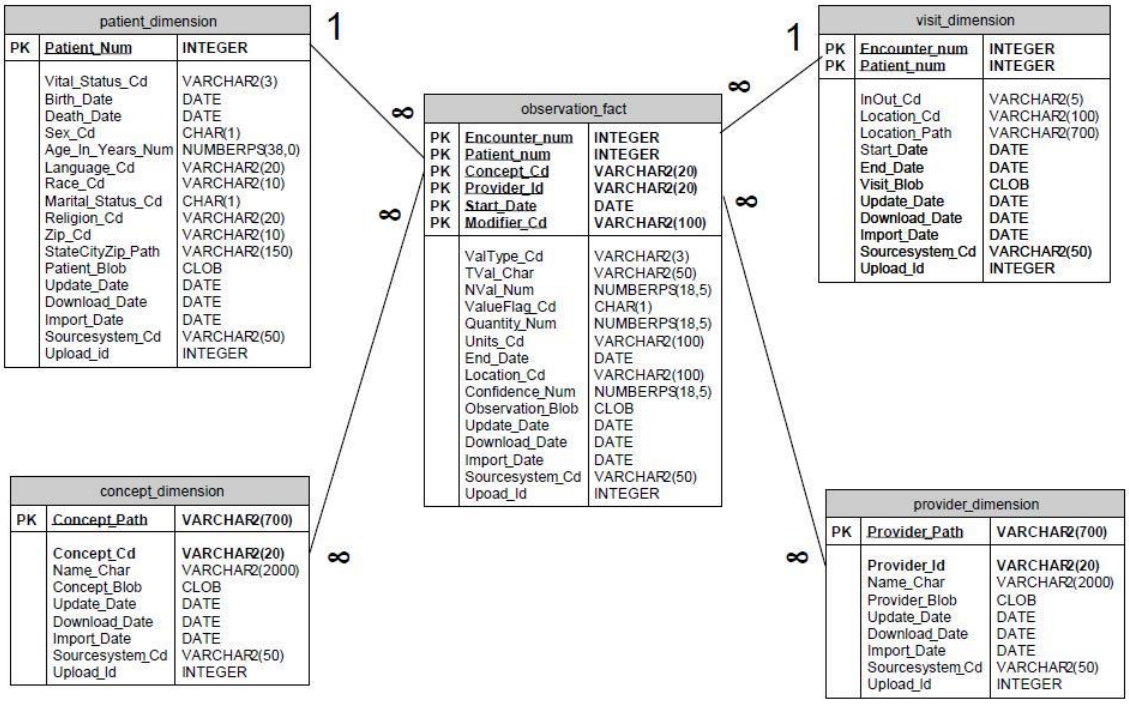


Figure 3 : Modèle relationnel i2b2

Pour des mesures en conditions réelles, nous avons procédé à une extraction ciblée portant sur les patients de la cohorte ArthroVIH. Cette cohorte, constituée lors d'une étude antérieure menée au CHU de Bordeaux, avait pour objectif d'analyser les causes des douleurs de la main chez les patients séropositifs. L'identification des patients inclus dans cette cohorte s'est appuyée sur une recherche par mots-clés, détaillée en section 2.5, et effectuée à l'aide de l'outil Query Builder de l'EDSH. Cet outil permet d'extraire des cohortes de patients répondant à des critères spécifiques en interrogeant, entre autres, les données textuelles stockées dans l'EDSH. Dans notre étude, nous avons reproduit cette recherche par mots-clés et extrait l'ensemble des documents associés aux patients identifiés par cette requête.

2.2 Développement d'une chaine de traitement de TAL

2.2.1 Sélection des outils et technologies adaptés

Dans le cadre de cette étude, nous avons choisi de nous concentrer sur une approche basée sur des règles pour la détection du contexte dans les documents médicaux. De nombreux algorithmes et frameworks de TAL dédiés au domaine médical existent dans différents langages de programmation, chacun offrant des capacités spécifiques. Nous avons initialement testé des bibliothèques open source en Java et en Python afin de comparer leurs performances et leur pertinence pour notre application. En Java, nous avons évalué IAMsystem 2.2.0 (49) pour la tâche de NER et FastContext 1.3.1.9 (50) (implémentation Java de l'algorithme ConText) pour la détection du contexte. En parallèle, nous avons testé en Python les bibliothèques EDS-NLP v0.14.0 (51) et MedSpaCy 1.3.0 (52), toutes deux basées sur l'architecture de spaCy (53). Toutefois, pour des raisons de rapidité d'exécution, de maintenabilité et de popularité, nous avons décidé d'abandonner Java au profit de Python. MedSpaCy a été développé par la Veterans Health Administration et l'Université de l'Utah

pour adapter SpaCy aux textes médicaux et propose un traitement multilangue, bien qu'il soit principalement optimisé pour l'anglais. En revanche, EDS-NLP, conçu par l'Assistance Publique - Hôpitaux de Paris (AP-HP), est spécifiquement dédié aux documents médicaux en français. La suite de cette étude ne traitera que des analyses avec les bibliothèques python. Enfin, afin d'optimiser le temps d'exécution, nous avons utilisé une approche par parallélisation. La parallélisation consiste à répartir le calcul sur plusieurs unités de traitement (CPU ou GPU) afin d'exécuter différentes tâches simultanément, plutôt que de manière séquentielle. Cette approche permet de réduire considérablement le temps d'exécution, en particulier lorsqu'un grand volume de documents est analysé.

2.2.2 Prétraitement des documents

Afin d'être analysables, les documents ont subi plusieurs étapes de prétraitement, comprenant la tokenisation, la normalisation et le découpage en phrases(27). La tokenisation consiste à segmenter un texte en unités élémentaires appelées tokens (mots, nombres, ponctuation). Nous appelons span, un ensemble de tokens, cela peut être un concept en plusieurs mots, une phrase, une partie d'un document etc. La normalisation vise à uniformiser le texte, notamment par la mise en minuscule, la suppression des accents et des guillemets dans notre cas. Le découpage en phrases segmente le texte en unités syntaxiques, nécessaire à certaines bibliothèques pour l'analyse contextuelle. EDS-NLP propose nativement ces trois fonctions, avec des optimisations spécifiques au domaine médical (voir annexe 1), tandis que MedSpaCy repose uniquement sur les fonctions natives de SpaCy, qui ne prennent pas en compte les spécificités des textes cliniques. Pour assurer une meilleure précision avec une méthode simple d'implémentation, nous avons choisi la tokenisation, la normalisation et le découpage en phrases d'EDS-NLP. De plus, nous avons comparé les temps d'exécution à chaque étape des différentes bibliothèques sur 10 000 documents.

2.2.3 Reconnaissance d'entités nommées

Pour la reconnaissance d'entités nommées (NER), nous avons opté pour une approche basée sur dictionnaire, en nous appuyant sur l'Unified Medical Language System (UMLS) (54) pour la partie étude sur grand échantillon de l'EDSH. Développé par la National Library of Medicine (NLM), l'UMLS est un métathésaurus regroupant et unifiant différentes terminologies médicales afin de faciliter l'interopérabilité et l'extraction d'informations à partir des bases de données de santé. Afin d'adapter l'UMLS à notre étude, nous avons appliqué un filtrage, en ne conservant que les terminologies disponibles en français, à savoir MedDRA, MeSH et LOINC. Nous avons ensuite affiné cette sélection en filtrant selon les types sémantiques d'intérêt, présentés dans le Tableau 1. Enfin, nous avons exclu les termes d'une ou deux lettres, ainsi que les mots *"par", "les", "maladie", "pris"* qui représentaient 25 % des concepts de notre corpus, jugés non pertinents dans le cadre d'une utilisation au sein d'un EDSH. Cette démarche avait pour objectif de réduire les temps d'exécution et d'optimiser l'espace de stockage en éliminant les données inutiles. Au final, le dictionnaire retenu comprennait 103 589 termes, appelés entités, correspondant à 62 293 concepts répartis en 16 catégories sémantiques.

Dans le cadre de notre étude pour l'exploitation de la cohorte ArthroVIH, la stratégie de NER a également reposé sur une approche par dictionnaire, non basée sur l'UMLS, mais sur trois groupes de mots-clés spécifiques, définis en amont par les investigateurs de l'étude. Nous avons strictement conservé la liste de mots-clés élaborée. Par ailleurs, nous avons maintenu la classification des entités selon leur groupe d'appartenance, telle que définie initialement dans l'étude.

Tableau 1 : Types sémantiques UMLS conservés

Version utilisé	*UMLS 2023AB*
Types sémantiques conservés	Acquired Abnormality, Anatomical Abnormality, Antibiotic, Cell or Molecular Dysfunction, Congenital Abnormality, Diagnostic Procedure, Disease or Syndrome, Finding, Gene or Genome, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Pharmacologic Substance, Sign or Symptom, Therapeutic or Preventive Procedure.
Termes supprimés	*Termes de 1 et 2 caractères et appartenant à la liste suivante : ["par", "les", "maladie", "pris"]*
Nombre total de termes	*103 589*
Nombre total de concepts	*62 293*

2.2.4 Détection du contexte

Pour la détection de contexte, nous avons utilisé les modules d'EDS-NLP et de MedSpaCy. EDS-NLP propose une série de chaînes de traitement permettant de qualifier les entités préalablement extraites lors de l'étape de reconnaissance d'entités nommées (NER), pour leur ajouté des modalités linguistiques. Ces chaînes reposent sur des algorithmes basés sur des règles et annotent les entités selon cinq modalités linguistiques : la négation (*eds.negation*), le contexte familial (*eds.family*), la spéculation (*eds.hypothesis*), le discours rapporté (*eds.reported_speech*) et les antécédents médicaux (*eds.history*). La qualification des entités repose sur l'extraction préalable de celles-ci dans le texte. Une fois identifiées, la chaîne de traitement recherche des indices contextuels, appelés *cues*, qui ajoutent une ou des modalités linguistiques à ces entités (exemple : « Le patient n'a ***pas*** de douleur » - négation). Ces indices peuvent apparaître avant ou après l'entité, ou être portés par des verbes modulant le sens de l'entité (exemple : « le patient ***pourrait*** avoir leucémie aigüe » - hypothèse). La segmentation du texte se fait en phrases et en propositions syntaxiques, réalisée par le module *sentencizer* et une liste de motifs de terminaison délimitant ce qui est appelé des syntagmes. Ces syntagmes assurent la propagation correcte des indices contextuels sur les entités (spans). L'attribution d'un indice à une entité dépend de leur position relative : un indice précédent est pris en compte s'il se situe entre le début du syntagme et l'entité ; un indice suivant est retenu s'il se trouve entre la fin de l'entité et la fin de la proposition. La chaîne de traitement intègre également un mécanisme de gestion des pseudo-indices, permettant d'exclure les expressions contenant un indice apparent sans en avoir la valeur sémantique attendue. Par exemple, l'expression "sans doute" contient le mot "sans", mais

n'exprime pas une négation. Ces pseudo-indices sont identifiés afin d'éviter qu'ils n'interfèrent avec l'analyse contextuelle. Les résultats de ces chaînes de traitement sont stockés au format SpaCy. Après identification des entités et des syntagmes, l'algorithme détecte les indices contextuels, classés en trois catégories : les indices précédents, situés avant l'entité à qualifier ; les indices suivants, situés après l'entité ; et les verbes, pouvant moduler le sens de l'entité. Le fonctionnement des différents modalités linguistiques est produit comme suit :

- *Négation* : utilisation des indices précédents, suivants, des pseudo-indices, des verbes de négation et d'une liste de termes de terminaisons.
- *Contexte familial* : utilisation d'indices se rapportant aux membres de la famille dans la proposition.
- *Antécédents médicaux* : identification d'indices dans la proposition, avec possibilité d'utilisation des sections du document. Nous n'avons pas utilisé cette fonctionnalité dans notre cas.
- *Hypothèse* : utilisation des indices précédents, suivants, des pseudo-indices, et des verbes indiquant une hypothèse, tels que "douter", ainsi que des verbes conjugués au conditionnel, indiquant une éventualité, comme "pourrait".
- *Discours rapporté* : utilisation des indices précédents, suivants, des verbes, et prise en compte des guillemets pour détecter les citations ou discours indirects.

Medspacy est une implémentation python de l'algorithme *ConText* en français. L'algorithme *ConText* repose sur une approche basée sur des expressions régulières et de termes déclencheurs (*trigger terms*), équivalent aux *cues*, pour attribuer des modalités linguistiques de contexte aux entités médicales identifiées dans un texte. Chaque entité est initialement considérée comme affirmée, et c'est seulement en présence d'une modalité linguistique dans une fenêtre d'analyse donnée que son statut est modifié. *ConText* attribue initialement trois types de modalités aux entités médicales et la version française en rajoute une quatrième : La négation, la temporalité qui peut être historique ou hypothétique, l'expérienceur et enfin la certitude. La portée des termes déclencheurs est déterminée par chaque règle et correspond à un nombre de token autour du terme déclencheur, contrairement à EDS-NLP qui analyse à l'échelle de la phrase au maximum. En général, elle s'étend depuis le terme déclencheur jusqu'à un terme de terminaison (*termination term*). Toutefois, certains déclencheurs, notamment ceux liés à la négation, peuvent également agir à rebours et modifier les entités situées avant eux. *ConText* prend également en compte des pseudo-déclencheurs, qui contiennent des termes indicateurs sans en avoir la signification attendue comme EDS-NLP. L'algorithme applique ses règles dans un ordre strict : il identifie d'abord les termes déclencheurs, puis il applique la portée définie par ces déclencheurs en respectant les termes de terminaison, et enfin il ajoute la ou les modalités linguistiques aux entités concernées (29,55).

Pour fonctionner, MedSpaCy nécessite un fichier externe de règles au format .json. Cependant, les règles originales avaient été produites sous format .tsv (équivalent à un fichier texte .txt). L'un des contributeurs du projet MedSpaCy a effectué une conversion des règles de ConText et les a directement intégrées à la bibliothèque. Toutefois, nous avons identifié plusieurs erreurs dans cette implémentation : certaines règles ne permettaient pas de

détecter certaines modalités linguistiques, tandis que d'autres entraînaient une sur-détection, générant des résultats aberrants. Afin de corriger ces problèmes, nous avons repris la version originale des règles disponible sur le GitHub des créateurs de *ConText* en français (56) et avons procédé à une conversion automatisé, corrigé en format .json. Le fichier de règles final, contenant 10 161 règles, a été utilisé dans notre étude et soumis au projet MedSpaCy. Le fichier étant trop volumineux, un échantillon des règles utilisées sont disponibles en annexe 2.

Tableau 2 : Définition des différentes modalités linguistiques d'après EDS-NLP et MedSpaCy

Critères	EDS-NLP	MedSpaCy (ConText)
Approche	Basée sur des règles	Basée sur des règles
Modificateurs		
Négation	***negation*** : Qui détecte les span négatifs	***negation*** : Affecté par un terme déclencheur comme « non », « écarté » etc.
Hypothèse	**hypothesis** : Qui détecte les spans qui sont des spéculations plutôt que des déclarations certaines	***hypothetical*** (*Temporality*) : Affecté par un terme déclencheur comme « non », « écarté » etc. **uncertain** : Pas de définition
Historique	***history*** : Qui détecte les spans explicitement décrits comme faisant partie des antécédents médicaux, c'est-à-dire précédées d'un synonyme de « antécédents médicaux ».	***historical*** (*Temporality*) : Affecté par un terme déclencheur comme « antécédent de », « statut post » etc. ou contenant une valeur temporelle explicite supérieure à 14 jours comme « il y a trois mois ».
Expérimentateur	***family*** : Qui détecte les spans qui décrivent un membre de la famille (ou un antécédent familial) du patient plutôt que le patient lui-même.	***other*** : Affecté par un terme déclencheur comme « antécédent familial », « sa mère » etc.
Discours rapporté	**reported_speech :** Qui détecte les spans qui relate un discours rapporté (par exemple, lorsque le médecin cite le patient)	Non applicable
Unité d'analyse	Phrase et proposition syntaxique	Dépend de la portée des termes déclencheurs (règles).
Méthode de détection	Identification d'indices contextuels avant et après les entités	Identification de termes déclencheurs avant et après les entités
Portée des modificateurs	Limite la portée à la phrase et à la proposition	Dépend de chaque règle : La portée est étendue au-delà de la phrase jusqu'à un terme de terminaison.
Gestion des pseudo-modificateurs	Oui	Oui
Version utilisé	v13.1	Medspacy 1.2.0 ConText : V2 modifié (voir Annexe 2)

Nous présentons en figure 4 la chaine de traitement que nous avons utilisé pour faire notre analyse de texte et de contexte pour obtenir une base de données contenant l'ensemble des concepts retrouvés par le NER dans les documents médicaux avec l'ensemble des modalités linguistiques associées. Nous avons mesuré pour chaque étape le temps nécessaire à l'exécution de celle-ci sur notre jeu d'exploration en parallélisant pour traiter l'ensemble de nos jeux de données.

```
def create_pipeline():
    # Création chaine de traitement "vide" basée sur SpaCy
    nlp = edsnlp.blank("eds")

    # Normalisation du texte
    nlp.add_pipe(
        eds.normalizer(
            lowercase=True,
            accents=True,
            quotes=True,
            spaces=False,
            pollution=False,
        )
    )

    # Segmentation en phrases
    nlp.add_pipe(eds.sentences())

    # Ajout du module de reconnaissance d'entitées nommées avec dictionnaire UMLS filtré
    nlp.add_pipe(
         eds.matcher(
             terms=umls_dict,
             attr="NORM"
         )
     )

    # Ajout d'un module MedSpaCy avec ajout des règles ConText pour détection des assertions
    nlp.add_pipe(
        "medspacy_context",
        config={"language_code": "fr", "rules": "../util/rules.json"},
    )

    # Détection des assertions spécifiques à EDS-NLP
    nlp.add_pipe(eds.negation())
    nlp.add_pipe(eds.hypothesis())
    nlp.add_pipe(eds.family(use_sections=False))
    nlp.add_pipe(eds.history())
    nlp.add_pipe(eds.reported_speech())

    return nlp
```

Figure 4 : Chaine de traitement pour analyse du contexte avec EDS-NLP et MedSpaCy

2.3 Analyse descriptive du jeu d'exploration, du gold standard et évaluation des algorithmes de contexte

Après application de notre chaîne de traitement sur l'ensemble des documents des différents jeux de données, nous avons extrait les entités analysées pour en déterminer leur contexte. Nous avons ensuite calculé le pourcentage de modalités linguistiques retrouvées, cette analyse servant de base à la construction du gold standard. Pour évaluer les performances des algorithmes, nous avons créé un gold standard constitué de 100 documents annotés manuellement après extraction automatique des concepts via le NER. Sept évaluateurs ont été sollicités pour cette annotation. La sélection de ces 100 documents a suivi une méthodologie spécifique, fondée sur un système de score. En effet, nous avons constaté dans la littérature que la mesure des performances peut être limitée lorsque les documents contiennent peu de modalités linguistiques. Un tirage purement aléatoire aurait pu conduire à un corpus sans suffisamment de modalités rares, empêchant ainsi une évaluation pertinente des algorithmes. Pour éviter ce risque, nous avons d'abord appliqué notre chaîne de traitement sur le jeu de données d'exploration et attribué un score à chaque document. Ce score était calculé en faisait la somme du nombre d'entités détectées, pondéré par la rareté des modificateurs dans l'ensemble du corpus c'est-à-dire l'inverse de la fréquence de la modalité dans l'ensemble du corpus, par document. Une entité comportant plusieurs modalités était comptabilisée autant de fois que de modalités qu'elle comportait. L'objectif était de sélectionner des documents riches en entités, tout en maximisant la présence de modalités linguistiques rares. Un guide d'annotation, disponible en annexe 3, a été fourni aux évaluateurs pour standardiser l'identification des modalités dans les documents médicaux. L'annotation a été réalisée à l'aide de DocAnno, un outil d'annotation de texte open source, en utilisant six types de modalités : Négation, Incertitude, Historique, Hypothétique, Non patient et discours rapporté. Les définitions des modalités étaient alignées entre EDS-NLP et MedSpaCy (*ConText*), sauf pour "historique", dont l'interprétation différait entre les deux bibliothèques. Pour mesurer l'impact de cette différence, les évaluateurs ont été répartis en deux groupes, chacun utilisant une définition distincte :

- Définition 1 : Historique = Entité explicitement identifiée comme faisant partie des antécédents médicaux du patient.
- Définition 2 : Historique = Entité ayant débuté il y a plus de 2 semaines par rapport à la date de la visite/document ou explicitement décrite comme un antécédent médical.

Enfin, pour garantir une correspondance entre les modificateurs des deux bibliothèques, nous avons effectué les équivalences suivantes :

- Négation (EDS-NLP) ↔ Négation (MedSpaCy)
- Hypothesis (EDS-NLP) ↔ Uncertain + Hypothetical (MedSpaCy)
- Family (EDS-NLP) ↔ Other (MedSpaCy)
- Discours rapporté (EDS-NLP) : non pris en charge par MedSpaCy

Pour évaluer les performances globales des algorithmes, nous avons utilisé les métriques classiques suivantes à l'échelle de l'entité :

- Rappel : Mesure la capacité du modèle à identifier correctement toutes les observations positives. Mathématiquement, il peut être représenté par le rapport suivant :

$$Rappel = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Négatifs}$$

- Précision : Mesure la capacité des modèles à ne pas générer de faux positifs. Mathématiquement, elle peut être représentée par le rapport suivant :

$$Précision = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Positifs}$$

- F-mesure : La F-mesure ou score F1 est une métrique combinant à la fois le rappel et la précision, calculée comme la moyenne harmonique entre ces deux mesures. Permet d'obtenir un équilibre pour mesurer les performances des modèles. Mathématiquement, elle peut être représentée par le rapport suivant :

$$F_{mesure} = 2 * \frac{Précision * Rappel}{Précision + Rappel}$$

Nous avons mesuré ces métriques à l'échelle de l'entité et du document. Nous avons stratifié ces résultats sur le type sémantique, sur la fréquence du concept dans le corpus et le nombre d'occurrence du concept au sein d'un même document. L'objectif était d'apprécier différents déterminants autour de la détection de contexte.

2.4 Mesures d'impact au sein de l'EDSH

Après application de notre chaîne de traitement sur l'ensemble du jeu de données, nous avons analysé l'impact des algorithmes de détection de contexte en procédant à plusieurs niveaux d'évaluation. Pour l'analyse descriptive de notre jeu de données, nous avons calculé des statistiques descriptives classiques (moyenne, écart-type, valeurs minimales et maximales, médiane, intervalle interquartile) sur les entités et concepts extraits, en tenant compte de plusieurs niveaux d'agrégation : Document, Séjour, Patient. Ensuite, nous avons catégorisé l'ensemble des entités extraits selon leur concept rattaché puis leur type sémantique, en évaluant leur répartition à l'aide de la médiane et de l'intervalle interquartile. L'évaluation de l'impact des algorithmes de contexte a été réalisée en quantifiant le nombre d'entités et de concepts éliminés à l'échelle du patient, dans une perspective de phénotypage. L'objectif était de mesurer la perte de concepts induite par l'application des algorithmes. Nous avons identifié les modalités linguistiques pertinents pour une stratégie de filtrage des patients, à savoir la négation, l'hypothèse et la notion de non patient. Les modalités "discours rapporté" et "historique" ont été exclus de cette analyse. En effet, il n'est pas souhaitable de

filtrer sur la temporalité car un utilisateur veut filtrer sur une affection qu'a actuellement le patient ou a eu le patient, que ce soit pour un critère d'inclusion ou d'exclusion. La notion d'antécédent n'apporte pas d'information supplémentaire pour notre cas d'usage. À partir de cette sélection, nous avons défini six stratégies de filtrage : Négation seule ; hypothèse seule ; non patient seul ; Négation + Hypothèse ; Négation + Non patient ; Négation + Hypothèse + Non patient. Pour chaque stratégie, nous avons calculé le nombre le nombre médian de concepts perdus par patient (avec l'intervalle interquartile associé) et nous avons stratifiés les résultats par type sémantique, afin d'identifier les catégories de concepts les plus impactées par le filtrage.

2.5 Mesures d'impact en vie réelle : Cohorte ArthroVIH

Afin d'évaluer l'impact des algorithmes de détection de contexte dans un cadre clinique réel, nous avons appliqué notre méthodologie de traitement à la cohorte ArthroVIH. Cette cohorte, initialement constituée au CHU de Bordeaux, vise à étudier les causes des douleurs des mains chez les patients suivis pour une infection par le VIH. L'objectif de cette analyse était de mesurer dans quelle mesure la prise en compte du contexte dans l'extraction des entités cliniques permet de faciliter l'identification des patients en réduisant le bruit, et ainsi diminuer la charge du tri manuel.

La population cible de cette étude était constituée des patients adultes suivis au CHU de Bordeaux, ayant reçu un diagnostic d'infection par le VIH et ayant exprimé, au moins une fois lors du suivi, une plainte concernant une douleur des mains, avec au moins une venue au CHU depuis 2010. Les critères de sélection utilisés reposaient sur une recherche par mots-clés, ciblant trois groupes d'entités :

- Le diagnostic du VIH, identifié à partir des codes CIM-10 spécifiques au VIH et de mots-clés associés aux traitements antirétroviraux (ex. : « Abacavir »).
- Plainte douloureuse (exemple : « Rhumatisme »)
- Localisation douloureuse (exemple : « main »)

L'ajout des traitements antirétroviraux comme critère de sélection visait à confirmer indirectement l'infection par le VIH, en supposant que leur prescription était exclusive aux patients séropositifs. Cette stratégie permettait d'éviter l'inclusion de mentions de recherche diagnostique ou de suspicions de VIH, qui auraient introduit du bruit dans la sélection des patients. Elle constituait ainsi un moyen empirique de limiter les biais liés à l'absence de prise en compte du contexte dans l'extraction des données médicales non structurées. L'ensemble des mots-clés utilisés sont disponible en annexe 6.

Nous avons donc reconstitué la population source de l'étude originelle en réutilisant la requête effectuée au sein de l'EDSH. Pour l'ensemble de ces patients, tous les documents qui leur étaient associés dans l'EDSH ont été extraits. Ces documents ont ensuite été analysés à l'aide de notre chaîne de traitement, avec pour NER la liste de mots-clés initialement utilisée dans la requête d'extraction et non l'UMLS. Par la suite, nous avons appliqué les algorithmes de détection de contexte EDS-NLP et MedSpaCy. À partir des résultats obtenus après application des algorithmes, nous avons appliqué différentes stratégies de filtrage afin de

nous rapprocher le plus possible de la population d'étude. Ces populations obtenues après filtrage sont désignées comme "populations filtrées". Nous avons ensuite procédé à une analyse descriptive du jeu de données, conformément à la méthodologie décrite en section 2.3. Nous n'avons pas décrit la partie "concept", car il était attendu que les patients présentent au maximum trois concepts : le diagnostic du VIH, la plainte douloureuse et la localisation de la douleur. Nous avons ensuite analysé :

- Les variations du nombre de patients sélectionnés après application de différentes stratégies de filtrage par le contexte.
- Les performances des algorithmes en termes de précision, en évaluant le nombre de faux positifs et d'exclusions erronées (silence dans les données).
- La distribution des erreurs commises par les algorithmes en fonction des différentes stratégies de filtrage appliquées.

2.6 Environnement de l'EDSH

Les données de l'EDS sont stockées dans une base de donnée dédiée, sécurisée, conforme au référentiel de la CNIL sur les EDS. Cette base de données est hébergée localement au CHU de Bordeaux. L'accès aux données individuelles patients est médiée par une couche d'authentification au niveau de la plateforme de l'EDS et une couche d'habilitation dans le contexte spécifique d'un projet spécifique.

3 Résultats

3.1 Mesures des temps d'exécutions

Afin d'orienter la conception de notre chaîne de traitement finale et d'évaluer le temps de traitement des différents jeux de données, nous avons mesuré les temps d'exécution des différents algorithmes utilisés pour le prétraitement des documents ainsi que pour l'ensemble de la chaîne de TAL. Le prétraitement des documents comprenait trois étapes principales : la tokenisation, la normalisation et le découpage en phrases. Nous avons comparé les performances de la bibliothèque EDS-NLP aux fonctions natives de SpaCy sur un corpus de 10 000 documents (jeu d'exploration). De manière générale, pour l'ensemble des fonctions, EDS-NLP était plus rapide. Pour la tokenisation, EDS-NLP effectue cette tâche en 9 secondes, contre 22 secondes pour SpaCy. Cela représente un facteur de 2,5 fois plus rapide, inférieur au rapport 5 à 6 fois plus rapide mentionné dans la documentation d'EDS-NLP. Cependant, lorsque l'exécution est réalisée sans parallélisation, nous retrouvons bien ce facteur de 5 à 6 annoncé (57). L'ajout de la normalisation augmente le temps de de prétraitement avec EDS-NLP à 10s. Nous rappelons qu'il n'existe pas de fonction native de normalisation pour SpaCy. Enfin, en intégrant également le découpage en phrases, le temps de traitement total atteint 14 secondes. À l'inverse, le module de découpage en phrases de SpaCy est plus lent. En combinant la tokenisation et la normalisation d'EDS-NLP avec le découpage en phrases de SpaCy, le temps total atteint 19 secondes. En plus de la logique utilisé pour le traitement des documents, ces résultats renforcent l'intérêt d'utiliser EDS-NLP pour l'ensemble de la phase de prétraitement, solution retenue pour notre chaîne de traitement finale. Ce choix permet un gain de temps significatif, facilitant le traitement de grands volumes de données.

Après avoir optimisé la phase de prétraitement, nous avons mesuré les temps d'exécution des modules de NER et de détection de contexte en utilisant EDS-NLP et MedSpaCy, avec un prétraitement effectué par EDS-NLP. Pour la reconnaissance des entités nommées, nous avons utilisé une approche basée sur un dictionnaire UMLS filtré, comme décrit dans la section méthodes. Avec EDS-NLP, cette tâche est effectuée en moins de 15 secondes, tandis qu'avec MedSpaCy, le traitement prend environ 42 secondes. Les résultats produits par les deux outils sont strictement identiques en termes de nombre d'entités détectées et de leur position dans le document. Face à cette différence de performance, nous avons opté pour EDS-NLP pour le NER.

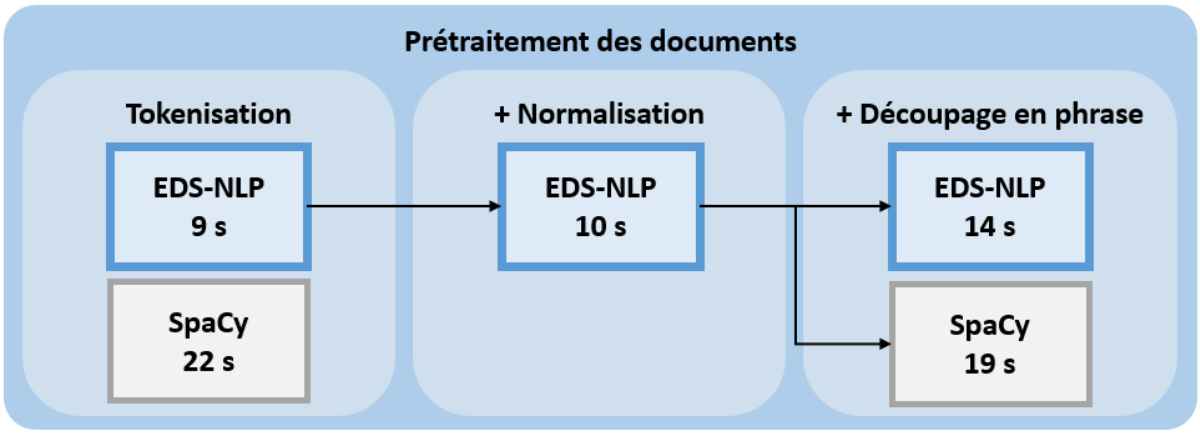


Figure 5 : Evaluation des temps d'exécution des méthodes de prétraitement de texte avec EDS-NLP et SpaCy sur 10 000 documents

En ce qui concerne la détection de contexte, nous observons des différences notables entre les deux bibliothèques. EDS-NLP réalise cette tâche en 37 secondes, alors que MedSpaCy ne prend que 13 secondes. Cette disparité s'explique par les différences de règles de détection de contexte entre les deux outils. Afin de pouvoir comparer les résultats des deux bibliothèques et mesurer leurs performances et impacts, nous avons retenu les deux bibliothèques dans notre chaîne de traitement finale : EDS-NLP pour le prétraitement et le NER, et à la fois EDS-NLP et MedSpaCy pour la détection de contexte. Le temps total d'exécution du TAL sur 10 000 documents, est de 1m22s, auxquels s'ajoutent 20 secondes d'initialisation. Cette initialisation comprend le chargement du dictionnaire UMLS et la préparation des documents avant leur traitement.

L'analyse des temps d'exécution met en évidence un point clé : la détection de contexte représente à elle seule 50 % du temps total de traitement. L'intégration d'un module de détection de contexte peut donc être particulièrement chronophage, un élément crucial à prendre en compte pour une mise en production en conditions réelles. Toutefois, plusieurs stratégies permettent d'optimiser ce temps de traitement. D'une part, l'ajustement des règles de détection peut réduire significativement la charge computationnelle. Si certaines modalités (négation, non patient, famille, discours rapporté, historique) ne sont pas nécessaires pour l'analyse finale, leur suppression permettrait un gain de temps substantiel. D'autre part, une

optimisation du module de NER contribuerait également à diminuer le temps total d'exécution de la chaîne de TAL.

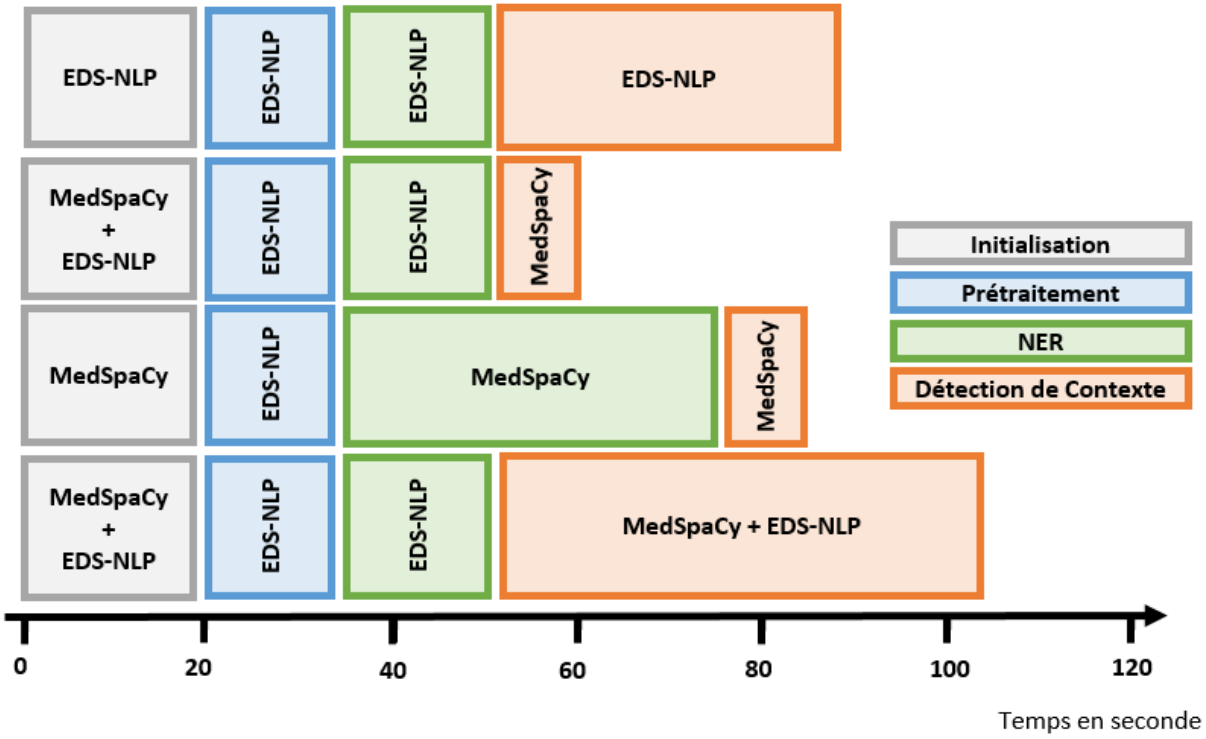


Figure 6 : Evaluation des temps d'exécution des chaines de TAL avec EDS-NLP et MedSpaCy sur 10 000 documents

Le tableau 3 résume le temps d'exécution des chaînes de traitement sur nos différents jeux de données. Un détail plus approfondi des temps d'exécution, en fonction des modalités linguistiques détectées pour la bibliothèque EDS-NLP, est présenté en annexe 4. À partir de ces données, nous estimons que si l'on choisit de ne conserver que la détection de la négation, de l'hypothèse et du non-patient, tout en utilisant uniquement EDS-NLP, il faudrait environ 3 à 6 jours pour traiter l'ensemble de l'entrepôt de données du CHU de Bordeaux soit 70 millions de documents selon le NER choisi.

Tableau 3 : Temps d'exécution des chaînes de TAL sur les différents jeux de données

Jeu de données	N documents	N entités	Temps d'exécution total (N docs/s)*
Jeu d'exploration	10 000	205 877	1min42s (98 docs/s)
Jeu de données EDS à grande échelle	1 007 583	15 498 881	11h38min24s (24 docs/s)
Jeu de données ArthroVIH	810 087	416 139	1h7min11s (201 docs/s)

** Évaluation des performances des chaînes de TAL appliquées à différents jeux de données, en mesurant le temps d'exécution total et le nombre de documents traités par seconde. Les résultats ont été obtenus avec une configuration de 16 Go de RAM et 8 CPU. Les documents ont été traités en un seul paquet pour le jeu d'exploration, tandis que les jeux de données à plus grande échelle ont été traités par paquets de 5 000 documents avec des batch de 10, afin d'optimiser l'utilisation de la mémoire.*

3.2 Description du jeu d'exploration et du gold standard

Le tableau 4 présente la description des entités retrouvées ainsi que les modalités linguistiques de contexte associées dans le jeu d'exploration et le Gold Standard. Dans le jeu d'exploration, le NER a permis d'identifier 205 877 entités, parmi lesquelles EDS-NLP et MedSpaCy ont extrait les différentes modalités. La négation est la modalité la plus fréquemment détectée, retrouvée dans 28 194 entités (13,7 %) avec EDS-NLP et 20 820 entités (10,1 %) avec MedSpaCy. La modalité hypothèse est identifiée dans 9 770 entités (4,7 %) avec EDS-NLP, tandis que MedSpaCy retrouve 7 028 entités (3,3 %). Pour cette modalité, nous avons regroupé l'incertitude et la notion d'hypothétique sous une seule modalité. La catégorie non patient est détectée dans 3 257 cas (1,6 %) avec EDS-NLP et 1 646 cas (0,8 %) avec MedSpaCy. Concernant la modalité historique, EDS-NLP identifie 2 862 entités (1,4 %), tandis que MedSpaCy en retrouve 2 406 (1,2 %). Enfin, la modalité discours rapporté est retrouvée uniquement par EDS-NLP, avec 8 873 entités (4,3 %). Dans le Gold Standard, construit à partir de 100 documents annotés manuellement, on observe une augmentation générale des modalités linguistiques détectées. La négation passe à 1 475 entités (16,6 %) avec EDS-NLP et 1 054 entités (11,8 %) avec MedSpaCy. La modalité hypothèse est identifiée dans 609 entités (6,8 %) avec EDS-NLP, tandis que MedSpaCy en retrouve 512 (5,7 %). Pour la modalité non patient, on retrouve 297 entités (3,3 %) détectées par EDS-NLP et 309 (3,5 %) par MedSpaCy. En ce qui concerne la modalité historique, EDS-NLP détecte 398 entités (4,5 %) contre 240 (2,3 %) avec MedSpaCy. Enfin, la modalité discours rapporté, absente de MedSpaCy, est retrouvée dans 463 entités (5,2 %) avec EDS-NLP. L'augmentation des fréquences des modalités linguistiques dans le Gold Standard par rapport au jeu d'exploration valide la méthodologie adoptée, qui visait à maximiser la présence de ces contextes dans les documents sélectionnés.

Tableau 4 : Description des entités retrouvés avec EDS-NLP et MedSpaCy

Jeu de données*	**Modalité linguistique**	**EDS-NLP**		**MedSpaCy**	
		N	%	N	%
Jeu d'exploration	Négation	28 194	13,7	20 820	10,1
	Hypothèse	9 770	4,7	7 028	3,3
	Non patient	3 257	1,6	1 646	0,8
	Historique	2 862	1,4	2 406	1,2
	Discours rapporté	8 873	4,3	-	-
Gold Standard	Négation	1 475	16,6	1 054	11,8
	Hypothèse	609	6,8	512	5,7
	Non patient	297	3,3	309	3,5
	Historique	398	4,5	2406	2,3
	Discours rapporté	463	5,2	-	-

*Le jeu d'exploration contenait 10 000 document pour 205 877 entités retrouvés. Le gold standard était composé de 100 document pour 8 895 entités retrouvés. Nous avons fait le choix de regroupé les modalités linguistiques d'incertitude et hypothétique de MedSpaCy pour n'en faire qu'une modalité hypothèse. Pour rappel, il n'est pas possible de retrouver la modalité de discours rapporté avec MedSpaCy

3.3 Mesures des performances

Pour les performances globales, l'ensemble des modalités linguistiques seront présentés, pour les performances stratifiées, seul la négation sera présentée. L'ensemble des résultats est disponible en annexe 5.

3.3.1 Mesures des performances à l'échelle de l'entité

3.3.1.1 Performances globales

L'évaluation des performances des algorithmes de détection de contexte à l'échelle des entités met en évidence des disparités selon les modalités linguistiques analysées. L'ensemble des résultats est présenté dans le tableau 5. De manière générale, EDS-NLP détecte un plus grand nombre de contextes quelle que soit la modalité (rappel plus élevé), tandis que MedSpaCy est plus précis, générant ainsi moins de faux positifs.

Concernant la négation, les deux algorithmes affichent des performances relativement proches, avec des scores F1 de 0,79 pour EDS-NLP (rappel : 0,92, précision : 0,69) et 0,80 pour MedSpaCy (rappel : 0,79, précision : 0,82). Pour les autres modalités linguistiques, les performances sont globalement plus faibles. Concernant l'hypothèse, EDS-NLP atteint un score F1 de 0,34 (rappel : 0,70, précision : 0,22) contre 0,25 pour MedSpaCy (rappel : 0,48, précision : 0,18). Pour la modalité Non patient, les deux modèles affichent des résultats similaires avec un score F1 de 0,54 pour EDS-NLP (rappel : 0,65, précision : 0,47) et 0,54 pour MedSpaCy (rappel : 0,51, précision : 0,57). Nous avons également testé deux définitions différentes de la modalité "historique", décrivant soit uniquement les antécédents médicaux, soit les antécédents médicaux plus les événements passés de plus de 15 jours. Globalement, les deux bibliothèques ont du mal à repérer ces notions, bien que la précision atteigne 0,71 pour EDS-NLP et 0,79 pour MedSpaCy. Lorsque nous utilisons la définition d'EDS-NLP (focalisée sur les antécédents médicaux stricts), la précision d'EDS-NLP augmente à 0,81, tandis que MedSpaCy atteint 0,75. C'est le seul cas où EDS-NLP est plus précis que MedSpaCy. En revanche, lorsque l'on adopte la définition de MedSpaCy (intégrant une notion plus large d'événements passés), MedSpaCy redevient plus précis avec une précision de 0,81, contre 0,65 pour EDS-NLP. Ces résultats soulignent l'importance de bien définir les modalités linguistiques même pour des notions d'apparence non complexe. Enfin pour la modalité discours rapporté, EDS-NLP avait un score F1 à 0,18 (rappel : 0,45, précision : 0,11), cette modalité n'étant pas prise en charge par MedSpaCy. Pour la suite des analyses, nous ne détaillerons que la détection de la négation, l'ensemble des analyses pour les différentes modalités linguistiques est disponible en annexe.

3.3.1.2 Performances stratifiés par type sémantique

L'analyse des performances des algorithmes EDS-NLP et MedSpaCy pour la détection de la négation met en évidence des variations selon le type sémantique (Figure 9-10-11). Parmi les types les plus représentés, *Disease or Syndrome, Finding* et *Sign or Symptom* affichent des performances similaires, avec des F1-scores proches de 0,8 pour les deux algorithmes. Toutefois, une diminution des performances est observée pour *Therapeutic or Preventive Procedure, Diagnostic Procedure et Pharmacologic Substance*, où les F1-scores sont inférieurs à 0,7. *Diagnostic Procedure* présente un décrochement plus marqué, notamment avec MedSpaCy, indiquant une difficulté accrue dans la détection de la négation pour cette catégorie. Concernant le rappel et la précision, la tendance observée reste la même : EDS-NLP présente un rappel plus élevé, tandis que MedSpaCy affiche une meilleure précision. L'analyse pas type sémantique pour l'ensemble des modalités linguistiques est disponible en Annexe.

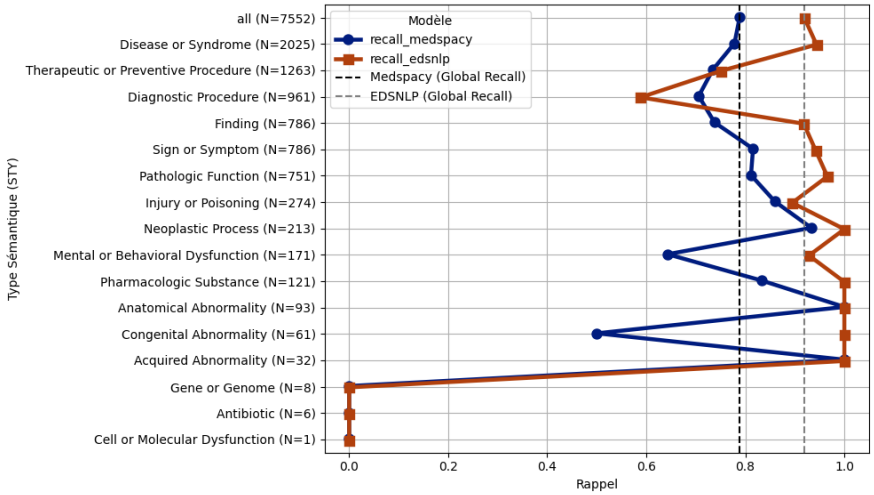


Figure 7 : Diagramme en forêt des rappels de ESD-NLP et MedSpaCy stratifiés par le type sémantique

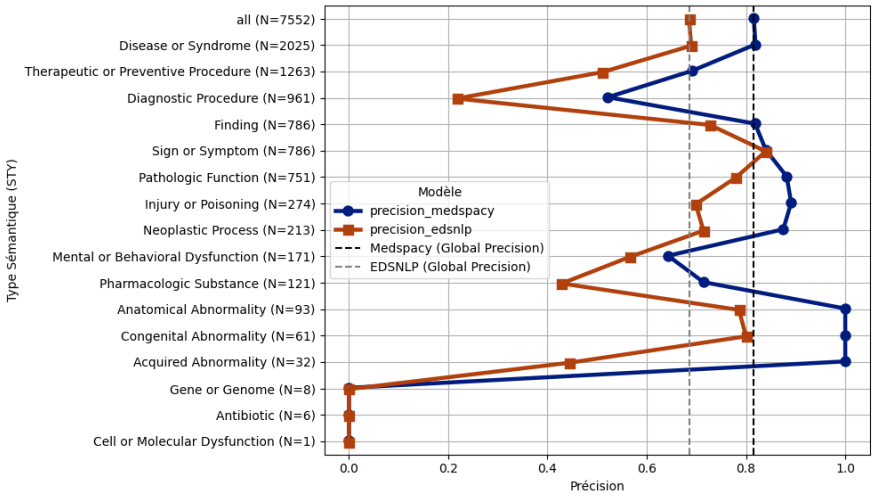


Figure 8 : Diagramme en forêt des précisions de ESD-NLP et MedSpaCy stratifiés par le type sémantique

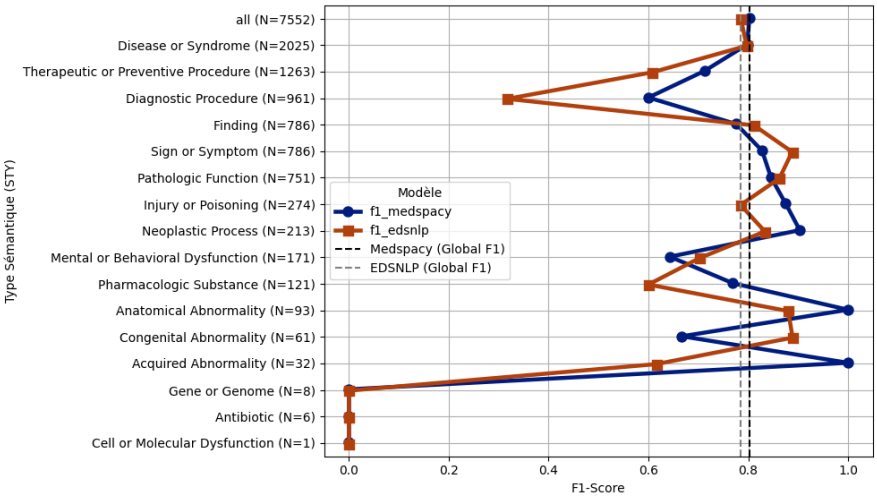


Figure 9 : Diagramme en forêt des F1 scores de ESD-NLP et MedSpaCy stratifiés par le type sémantique

3.3.1.3 Stratifiés par fréquence du concept dans le corpus

L'analyse des performances des algorithmes EDS-NLP et MedSpaCy pour la détection de la négation met en évidence des variations selon la fréquence d'apparition des concepts dans le corpus. De manière générale, les performances pour l'ensemble des métriques sont meilleures pour les concepts modérément fréquents. Parmi les concepts très fréquents (> 20 %), on retrouve des termes comme « TDM », « urgences », « dyspnée », « échographie », « atteinte », « douleur », « fièvre », « diagnostic », « lésions », « surveillance », « antécédents familiaux », « thérapeutique » et « pathologie ». Dans la catégorie 10 à 20 %, on identifie des concepts tels que « chute », « sténose », « asthme », « dépistage », « appendicectomie », « biopsie », « intervention », « sevrage », « hypertrophie » et « AVC ». Les concepts apparaissant entre 5 et 10 % incluent « médicaments », « cancer », « dyspnée d'effort », « hernie », « scanner cérébral », « kyste », « HTAP », « surpoids » et « cataracte ». Dans la catégorie 1 à 5 %, on retrouve des termes comme « hypothyroïdie », « dénutrition », « radiothérapie », « dialyse », « zona », « psoriasis », « inflammation », « traumatisme », « reprise de traitement » et « rupture ». Enfin, les concepts les plus rares (< 1 %) incluent « CIVD », « crise vaso-occlusive », « ataxie », « maladie de Kawasaki », « maladie de Ménière », « cataracte congénitale », « signe rénal » et « greffe de moelle osseuse ».

Les figures 12 et 13 présentent les performances de MedSpaCy et EDS-NLP en fonction de cette stratification. Une légère augmentation des performances est observée dans la catégorie 5 à 10 % pour l'ensemble des métriques et également 10 à 20 % pour MedSpaCy. L'hypothèses pouvant expliquer ce phénomène si celle-ci n'est pas fortuite repose sur l'idée que les performances devraient logiquement augmenter avec la fréquence des concepts, car les règles et la construction des algorithmes ont probablement été optimisées à partir des concepts les plus fréquents, leur permettant ainsi d'être plus efficaces sur ces termes. Toutefois, ce n'est pas exactement ce que nous observons. Il est possible que les performances sur les concepts très fréquents soient impactées par la présence de nombreuses entités et concepts peu informatifs, pour lesquelles la détection du contexte est moins pertinente et donc moins bien détectée, comme surveillance, thérapeutique ou antécédents familiaux.

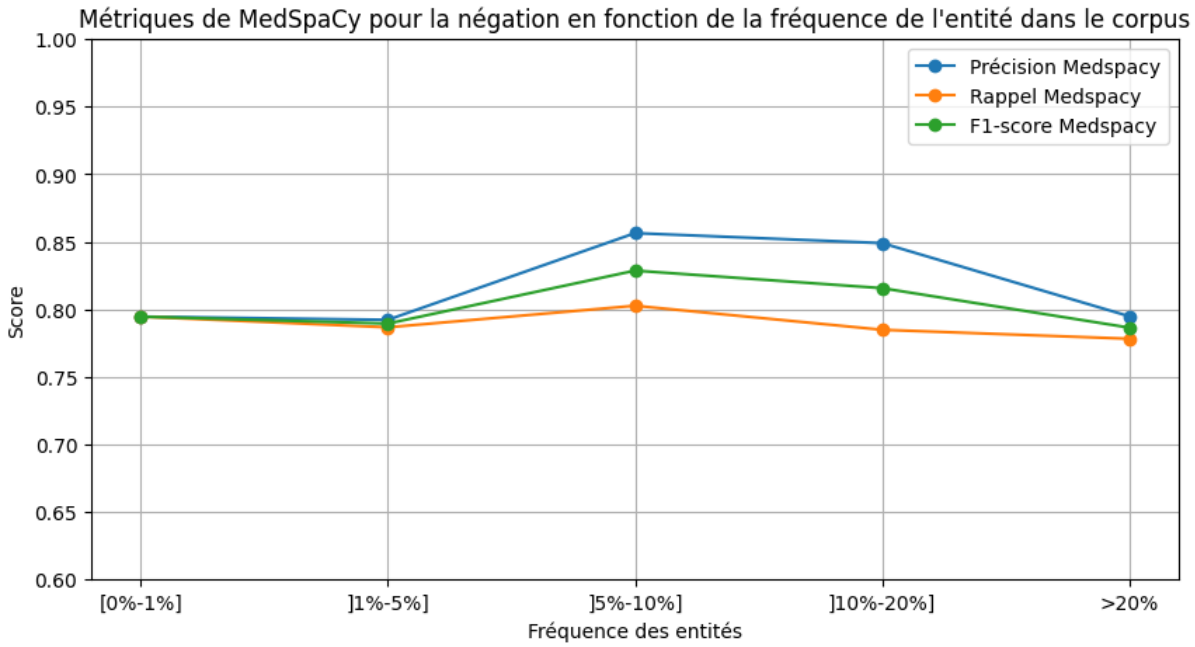


Figure 10 : Performances de MedSpaCy pour la négation en fonction de la fréquence d'apparition du concept dans le corpus

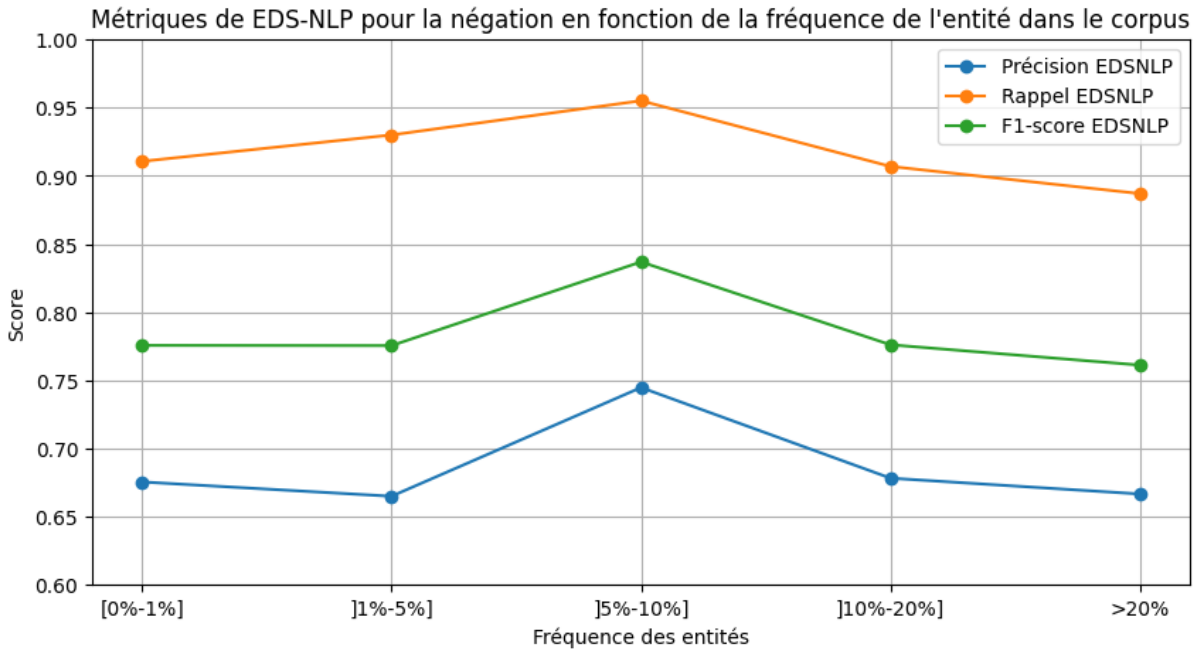


Figure 11 : Performances de EDS-NLP pour la négation en fonction de la fréquence d'apparition du concept dans le corpus

3.3.1.4 Stratifiés par fréquence du concept dans le document

L'analyse des performances des algorithmes EDS-NLP et MedSpaCy pour la détection de la négation met en évidence des variations en fonction de la fréquence d'apparition des concepts au sein d'un même document. Les concepts les plus fréquents dans un document sont principalement associés à des examens médicaux, traitements et pathologies récurrentes, tels que les concepts de « scanner », « chimiothérapie », « lésions » mais également des concepts de « métastases », « récidives », « interventions », « asthme », « COVID-19 » etc. À l'inverse, certains concepts plus rares, apparaissant une seule fois dans un document, incluent des termes plus spécifiques comme « utérus rétroversé », « sténose pulmonaire infundibulaire », « constipation chronique » ou « troubles alimentaires ». Les performances des algorithmes en fonction de la fréquence du concept dans un document sont présentées dans les figures 14 et 15. Une tendance générale de diminution des performances est observée lorsque la fréquence du concept dans un même document augmente. Avec MedSpaCy, le score F1 passe de 0,81 (rappel : 0,79, précision : 0,83) pour des concepts apparaissant une seule fois à 0,50 (rappel : 0,61, précision : 0,46) lorsque ces concepts sont répétés sept fois dans un document. EDS-NLP suit la même tendance avec un score F1 passant de 0,84 (rappel : 0,94, précision : 0,75) à 0,50 (rappel : 0,61, précision : 0,46) pour des concepts atteignant sept occurrences. Après un examen manuel des données, il apparaît que les concepts les plus fréquents dans un même document sont souvent des termes liés aux Diagnostic Procedures, comme le « TDM », ou des concepts peu informatifs pour lesquels la prise en compte du contexte est moins pertinente comme « maladie ». Cela pourrait expliquer la baisse des performances observée lorsque ces concepts sont répétés dans un même document.

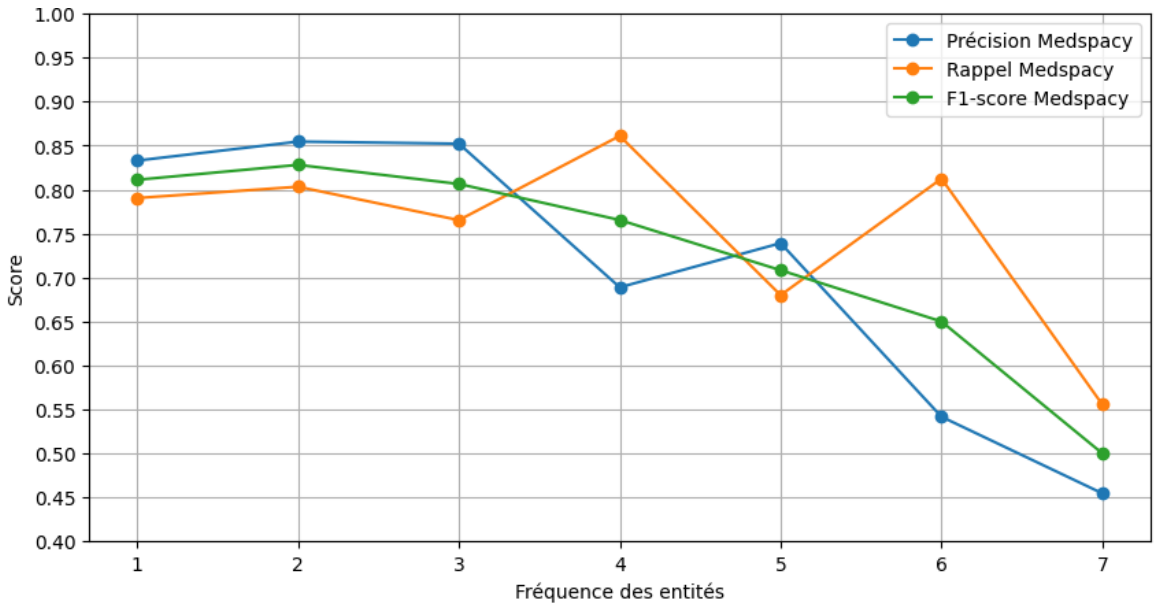


Figure 12 : Performances de MedSpaCy pour la négation en fonction de la fréquence d'apparition d'un même concept au sein d'un document

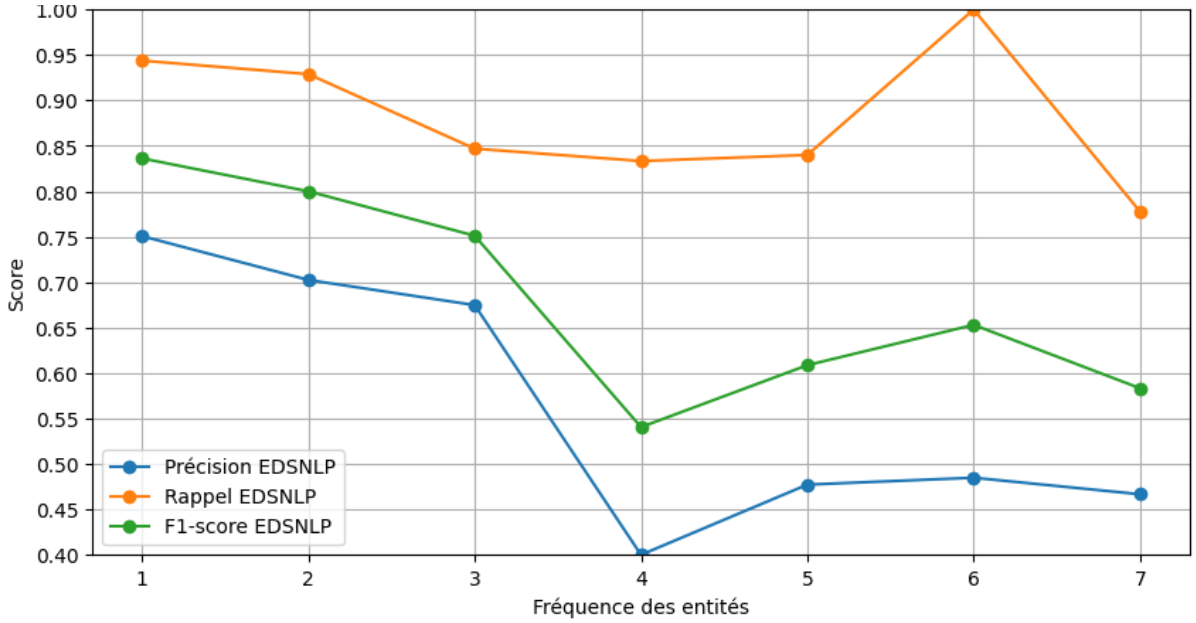


Figure 13 : Performances de EDS-NLP pour la négation en fonction de la fréquence d'apparition d'un même concept au sein d'un document

3.3.2 Mesures des performances à l'échelle du document

3.3.2.1 Globale

Les résultats des performances des algorithmes EDS-NLP et MedSpaCy à l'échelle du document sont présentés dans le Tableau 5. Contrairement à l'échelle de l'entité, où chaque occurrence d'un concept est évaluée individuellement, l'échelle du document prend en compte l'ensemble des entités associées à un même concept dans un même document. L'objectif est de déterminer si, à l'échelle du document, un concept possède une modalité linguistique donnée (négation, hypothèse, historique, etc.). Pour cela, chaque occurrence du concept est annotée avec sa propre modalité linguistique. Ensuite si chaque entité de ce concept possède une modalité particulière, alors le concept obtiendra cette modalité. Ainsi, si un concept apparaît plusieurs fois et que certaines de ses occurrences sont négatives tandis que d'autres non, le concept ne sera pas considéré comme négatif, car il est effectivement mentionné dans le document avec au moins une occurrence positive.

À l'échelle du document, les performances sont globalement similaires à celles observées à l'échelle de l'entité, avec une légère amélioration de l'ensemble des métriques. Concernant la négation, EDS-NLP atteint un score F1 de 0,81 (rappel : 0,94, précision : 0,75), contre 0,81 également pour MedSpaCy (rappel : 0,78, précision : 0,84). Pour la modalité hypothèse, les performances restent faibles avec un F1 de 0,40 pour EDS-NLP (rappel : 0,73, précision : 0,23) et 0,38 pour MedSpaCy (rappel : 0,48, précision : 0,38). Concernant la modalité non patient, les performances sont très proches de celles observées à l'échelle de l'entité avec un F1 de 0,56 pour EDS-NLP (rappel : 0,63, précision : 0,51) et 0,54 pour MedSpaCy (rappel : 0,50, précision : 0,57). Pour la modalité historique, les scores restent globalement faibles, avec un F1 de 0,14 pour EDS-NLP (rappel : 0,11, précision : 0,84) et 0,14 pour MedSpaCy (rappel : 0,09, précision : 0,18). Enfin, pour la modalité discours rapporté, seul

EDS-NLP permet une détection, avec un F1 de 0,18 (rappel : 0,45, précision : 0,11). Cette légère amélioration du rappel et de la précision s'explique par le fait que certaines erreurs de détection du contexte sont compensées par d'autres occurrences du même concept dans le document. Par exemple, considérons les deux phrases suivantes dans un même document : "Le patient a arrêté son suivi pour sa maladie de Crohn." et "La maladie de Crohn est pourtant active.". Dans ce cas, si une erreur est commise dans la première phrase en attribuant à tort une négation au concept maladie de Crohn, la deuxième phrase où ce même concept est affirmé permet de rectifier l'erreur à l'échelle du document. Finalement, le concept « maladie de Crohn » est bien considéré comme ne possédant pas de modalité linguistique, malgré une annotation erronée sur une occurrence de ce concept. Cette approche à l'échelle du document permet d'appréhender l'impact des erreurs de classification, des faux positifs sur notre jeu de données.

Tableau 5: Performances des algorithmes EDS-NLP et MedSpaCy pour la détection de contexte à l'échelle des entités et des documents

Echelle	Modalité linguistique	EDS-NLP			MedSpaCy		
		Rappel	Précision	F1	Rappel	Précision	F1
Entité							
	Négation	0,92	0,69	0,79	0,79	0,82	0,80
	Hypothèse	0,70	0,22	0,34	0,45	0,18	0,25
	Non patient	0,65	0,47	0,54	0,57	0,51	0,54
	Historique*	0,11	0,71	0,19	0,08	0,79	0,14
	Historique 1	*0,18*	*0,81*	*0,29*	*0,13*	*0,75*	*0,22*
	Historique 2	*0,08*	*0,65*	*0,15*	*0,06*	*0,81*	*0,11*
	Discours rapporté	0,41	0,07	0,12	-	-	-
Document							
	Négation	0,94	0,75	0,81	0,78	0,84	0,81
	Hypothèse	0,73	0,24	0,36	0,47	0,19	0,27
	Non patient	0,63	0,51	0,57	0,56	0,57	0,57
	Historique	0,11	0,71	0,19	0,08	0,81	0,14
	*Historique 1**	*0,18*	*0,84*	*0,30*	*0,14*	*0,81*	*0,23*
	*Historique 2**	*0,09*	*0,63*	*0,15*	*0,06*	*0,81*	*0,11*
	Discours rapporté	0,45	0,11	0,18	-	-	-

* La modalité historique a été évaluée selon deux définitions : Historique 1, où une entité est considérée comme historique si elle fait explicitement partie des antécédents médicaux, et Historique 2, où elle est aussi incluse si elle est également mentionné comme un événement passé de plus de 15 jours. Le jeu de données a été scindé, avec quatre évaluateurs pour Historique 1 et trois pour Historique 2. La modalité historique correspond aux performances sur l'ensemble du jeu de données.

3.3.2.2 Stratifiés par type sémantique

Les performances des algorithmes EDS-NLP et MedSpaCy à l'échelle du document, stratifiées par type sémantique, montrent des différences notables selon la catégorie analysée. Les résultats sont présentés en Tableau 6. Pour la majorité des catégories sémantiques, EDS-NLP affiche des scores F1 supérieurs à ceux de MedSpaCy. Par exemple, pour *Disease or Syndrome* (N=1283), EDS-NLP atteint un F1 de 0,84 contre 0,81 pour MedSpaCy. Pour *Finding* (N=568), les performances sont également en faveur d'EDS-NLP, avec

un F1 de 0,85 contre 0,76 pour MedSpaCy. De même, pour *Sign or Symptom* (N=500) et *Pathologic Function* (N=414), EDS-NLP atteint respectivement 0,87 et 0,89, contre 0,83 et 0,83 pour MedSpaCy. Cependant, pour les catégories moins fréquentes, MedSpaCy semble légèrement mieux performer, avec des scores F1 souvent plus proches de 1,0, ce qui suggère une meilleure gestion des concepts rares. Un point important est que EDS-NLP a davantage bénéficié du gommage des erreurs pour les catégories les plus fréquentes, car avait un rappel plus élevé. L'amélioration des performances à l'échelle du document est plus marquée pour cet algorithme, car les erreurs de détection du contexte sur certaines occurrences sont compensées par d'autres affirmations du même concept dans le document. Toutefois, pour *Therapeutic or Preventive Procedure* (N=604) et *Diagnostic Procedure* (N=516), les performances restent faibles pour les deux bibliothèques, avec des scores F1 ne dépassant pas 0,70. Cela suggère que ces modèles ne sont pas conçus pour analyser le contexte de ces types d'informations, qui sont souvent utilisées dans des formulations médicales où la négation ou l'incertitude ne sont pas explicitement exprimées de manière standardisée.

Tableau 6 : Evaluation des performances de EDS-NLP et MedSpaCy pour la négation stratifiées sur le type sémantique à l'échelle du document

Type sémantique*	EDS-NLP			MedSpaCy		
	Rappel	Précision	F1	Rappel	Précision	F1
Disease or Syndrome (N = 1283)	0,94	0,76	0,84	0,77	0,85	0,81
Therapeutic or Preventive Procedure (N = 604)	0,79	0,63	0,70	0,77	0,70	0,73
Finding (N = 568)	0,94	0,78	0,85	0,76	0,87	0,81
Diagnostic Procedure (N = 516)	0,82	0,24	0,38	0,36	0,31	0,33
Sign or Symptom (N = 500)	0,96	0,86	0,91	0,81	0,85	0,83
Pathologic Function (N = 414)	0,97	0,82	0,89	0,77	0,90	0,83
Injury or Poisoning (N = 120)	0,92	0,79	0,85	0,92	0,85	0,88
Mental or Behavioral Dysfunction (N = 117)	1,00	0,71	0,83	0,60	0,67	0,63
Neoplastic Process (N = 100)	1,00	0,63	0,77	1,00	1,00	1,00
Pharmacologic Substance (N = 92)	1,00	0,42	0,59	1,00	0,83	0,91
Anatomical Abnormality (N = 69)	1,00	0,89	0,94	1,00	1,00	1,00
Congenital Abnormality(N = 32)	1,00	0,75	0,86	0,33	1,00	0,50
Acquired Abnormality (N = 12)	1,00	0,50	0,67	1,00	1,00	1,00

* Il n'a été retrouvé aucun concept négatif après filtration par le gold standard pour certains types sémantiques qui ne sont donc pas présenté dans ce tableau

3.3.2.3 Stratifiés par fréquence du concept dans le corpus

Les performances des algorithmes EDS-NLP et MedSpaCy pour la détection de la négation, stratifiées selon la fréquence du concept dans le corpus, sont présentées dans le Tableau 7. De manière générale, EDS-NLP affiche des performances légèrement supérieures à celles de MedSpaCy, avec des scores F1 systématiquement plus élevés, variant entre 0,79 et 0,87 pour EDS-NLP contre 0,79 à 0,83 pour MedSpaCy. Concernant l'impact de la fréquence des concepts, les résultats ne montrent pas de variation majeure des performances en fonction de la fréquence d'apparition des concepts dans le corpus. Contrairement à ce que l'on pourrait attendre, les performances ne s'améliorent pas significativement pour les concepts les plus fréquents (> 20 %). Au contraire, les meilleures performances sont observées pour la catégorie [0%-1%], où EDS-NLP atteint un F1 de 0,86 et MedSpaCy de 0,80.

Tableau 7 : Performances d'EDS-NLP et MedSpaCy pour la négation stratifiés par fréquence du concept au sein du corpus

Fréquence	EDS-NLP			MedSpaCy		
	Rappel	Précision	F1	Rappel	Précision	F1
[0%-1%]	0,97	0,77	0,86	0,75	0,86	0,80
]1%-5%]	0,94	0,72	0,81	0,79	0,82	0,80
]5%-10%]	0,93	0,72	0,81	0,79	0,80	0,80
]10%-20%]	0,94	0,76	0,84	0,79	0,82	0,81
>20%	0,94	0,80	0,87	0,77	0,89	0,83

3.3.2.4 Stratifiés par le nombre d'occurrence du concept au sein d'un même document

Les performances des algorithmes EDS-NLP et MedSpaCy pour la détection de la négation, stratifiées selon le nombre d'occurrences du concept dans un même document, sont présentées dans le Tableau 8. De manière générale, les performances restent relativement stables tant que le nombre d'occurrences est de 3 ou moins. EDS-NLP affiche un score F1 légèrement supérieur lorsque le concept n'apparaît qu'une seule fois dans le document, avec un score de 0,86 contre 0,80 pour MedSpaCy, grâce à son rappel plus élevé (0,97 contre 0,75). En revanche, lorsque le concept est présent deux ou trois fois, les performances des deux algorithmes deviennent très similaires, avec des scores F1 oscillant entre 0,79 et 0,81, suggérant que le nombre d'occurrences n'a pas d'impact majeur sur les performances tant qu'il reste dans cette limite. Cependant, aucune performance n'a pu être calculée pour les concepts apparaissant plus de trois fois dans un même document, car le gold standard ne retrouvait aucun concept négatif dans ces cas. Autrement dit, lorsqu'un concept apparaissait plus de trois fois, les évaluateurs retrouvaient systématiquement au minimum une entité sans une modalité spécifique qui rend le concept affirmé. Cette observation suggère que la détection de la négation devient inutile pour les concepts très répétés. Si un concept est mentionné quatre fois ou plus dans un document, au moins une de ses occurrences sera affirmée, rendant toute correction par un algorithme de détection du contexte superflue. Autrement dit, l'identification de la négation n'a d'intérêt que pour les concepts apparaissant peu dans un document, car au-delà de trois occurrences, la présence du concept est déjà établie sans que l'on puisse lever l'ambiguïté. Cette information pourrait être utile pour optimiser les ressources de calcul en évitant d'appliquer des modèles de détection du contexte lorsque la fréquence d'un concept dépasse ce seuil.

Tableau 8 : Performances d'EDS-NLP et MedSpaCy pour la négation stratifiées par le nombre d'occurrence du concept au sein d'un même document

Nombre d'occurrence*	EDS-NLP			MedSpaCy		
	Rappel	Précision	F1	Rappel	Précision	F1
1	0,97	0,77	0,86	0,75	0,86	0,80
2	0,94	0,72	0,81	0,79	0,82	0,80
3	0,93	0,72	0,81	0,79	0,80	0,80

* Nombre d'occurrence du concept au sein d'un même document. Aucune performance n'a pu être calculé à partir de 3 occurrences car aucun concept avec plus de 3 occurrences négatives étaient retrouvé dans le gold standard

3.4 Mesures d'impact sur l'EDSH

3.4.1 Description grand échantillon EDSH

L'analyse du grand échantillon de l'EDSH a porté sur 14 445 patients, 185 545 séjours et 1 007 583 documents, permettant l'extraction de 15 498 881 entités. Parmi eux, 14 418 patients, 185 305 séjours et 957 372 documents contenaient au moins une entité. La médiane par patient était de 6 séjours [2-15] et 24 documents [7-61]. Concernant l'identification des concepts, une médiane de 2 concepts [1-10] pour 9 entités [4-20] a été retrouvée par document, 11 concepts [3-31] pour 31 entités [12-74] par séjour et 94 concepts [28-199] pour 363 entités [78-1 133] par patient. Une forte variabilité est observée, certains documents pouvant contenir plus de 1 000 entités et 300 concepts, tandis qu'un patient peut accumuler jusqu'à 54 239 entités et 1 480 concepts. On compte donc moins de 1 % des patients ne possédant aucun concept dans leurs documents. Ces résultats sont détaillés dans le Tableau 9.

Tableau 9 : Description du jeu de données grand échantillon de l'EDSH

Métrique		Entités N = 15 498 881	Concepts* N = 15 260	Documents N = 1 007 583	Séjours N = 185 545	Patients N = 14 445
Par document				-	-	-
	Moyenne (ET)	15,4 (22,0)	8,2 (14,9)	-	-	-
	Médiane [25%-75%]	9 [4-20]	2 [1-10]	-	-	-
	Min ; Max	0 ; 1056	0 ; 350	-	-	-
Par séjour					-	-
	Moyenne (ET)	83,5 (227,6)	24,1 (35,1)	5,4 (16,6)	-	-
	Médiane [25%-75%]	31 [12-74]	11 [3-31]	2 [1-5]	-	-
	Min ; Max	0 ; 21 253	0 ; 562	1 ; 2 036	-	-
Par patient						-
	Moyenne (ET)	1 073,0 (2162,2)	139,0 (146,4)	69,8 (145,7)	12,8 (19,1)	-
	Médiane [25%-75%]	363 [78-1133]	94 [28-199]	24 [6-71]	6 [2-15]	-
	Min ; Max	0 ; 54 239	0 ; 1 480	1 ; 3 709	1 ; 247	-

3.4.2 Impact selon les différentes stratégies de filtrage

L'évaluation de l'impact du contexte à travers différentes stratégies de filtration a été réalisée en appliquant successivement les filtres de négation, hypothèse et non-patient, seuls ou en combinaison. Globalement, la prise en compte du contexte a entraîné une diminution médiane du nombre de concepts extraits par patient, avec des variations selon les bibliothèques utilisées et la stratégie de filtrage adoptée. Les résultats sont présentés en Tableau 10,11 et annexe 9.

L'application du filtre de négation seule réduit la médiane des concepts extraits à 82 concepts [25-178] avec EDS-NLP, soit une diminution de 12 concepts (-12,8 %), et à 86 concepts [26-183] avec MedSpaCy, représentant une réduction de 8 concepts (-8,5 %). Les types sémantiques les plus impactés sont *Disease or Syndrome* (-4 pour EDS-NLP et -3 pour MedSpaCy) et *Pathologic Function* (-2 concepts pour EDS-NLP et -1 concept pour MedSpaCy). Le filtrage des concepts hypothétiques entraîne une réduction médiane à 89 concepts [27-189] pour EDS-NLP (-5 concepts, -5,3 %) et à 92 concepts [27-194] pour MedSpaCy (-2 concepts, -2,1 %). Le filtrage des entités non liées au patient a un impact plus limité, avec une réduction médiane de 1 (-1%) pour les deux bibliothèques.

L'application conjointe de plusieurs filtres accentue l'impact de la filtration. La combinaison de la négation et de l'hypothèse réduit la médiane à 78 concepts [23-167] pour EDS-NLP (-16 concepts, -17,0 %) et à 84 concepts [25-178] pour MedSpaCy (-10 concepts, -10,6 %). Lorsque la négation et le filtre non-patient sont combinés, la médiane diminue à 82 concepts [24-174] pour EDS-NLP (-12 concepts, -12,8 %) et à 85 concepts [26-182] pour MedSpaCy (-7 concepts, -7,4 %). Enfin, l'application simultanée des trois filtres aboutit à la plus forte réduction, avec une médiane de 77 concepts [23-166] pour EDS-NLP (-17 concepts, -18,1 %) et 83 concepts [25-178] pour MedSpaCy (-11 concepts, -11,7 %). L'effet additif de la combinaison des filtres est marqué. Lorsqu'un concept est à la fois en hypothèse et en négation, l'application d'un seul filtre ne suffit pas toujours à l'exclure, car l'autre modalité reste active. En revanche, la combinaison de plusieurs filtres permet une suppression plus efficace des concepts, expliquant ainsi l'impact plus marqué des stratégies combinées par rapport aux filtres appliqués individuellement.

Tableau 10 : Impact de la filtration avec MedSpaCy

Type sémantique	Sans filtrage	Négation		Hypothèse		Non Patient		Négation + Hypothèse + Non Patient	
	Médiane [25%-75%]	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif
Total (Distribution globale)*	**94 [28-199]**	**86 [26-183]**	**-8 [-2;-16]**	**92 [27-194]**	**-2 [-1;-5]**	**93 [28-197]**	**-1 [;-2]**	**83 [25-178]**	**-11 [-3;-21]**
Disease or Syndrome	16 [5-40]	14 [4-35]	-2 [-1;-5]	15 [5-39]	-1 [;-1]	16 [5-40]		13 [4-34]	-3 [-1;-6]
Therapeutic or Preventive Procedure	14 [4-32]	13 [3-31]	-1 [-1;-1]	14 [4-32]		14 [4-32]		13 [3-30]	-1 [-1;-2]
Finding	10 [3-24]	9 [2-22]	-1 [-1;-2]	10 [3-24]		10 [3-24]		9 [2-22]	-1 [-1;-2]
Diagnostic Procedure	11 [4-22]	11 [3-21]	[-1;-1]	11 [4-21]	[;-1]	11 [4-22]		11 [3-21]	[-1;-1]
Sign or Symptom	7 [2-15]	6 [1-13]	-1 [-1;-2]	6 [1-15]	-1 [-1;]	7 [2-15]		5 [1-12]	-2 [-1;-3]
Pathologic Function	7 [2-15]	6 [2-12]	-1 [;-3]	7 [2-14]	[;-1]	7 [2-15]		5 [2-12]	-2 [;-3]
Pharmacologic Substance	4 [1-8]	4 [1-7]	[;-2]	3 [1-7]	-1 [;-2]	4 [1-8]	[;-1]	3 [1-7]	-1 [;-2]
Injury or Poisoning	3 [1-6]	2 [0-5]	-1 [-1;-1]	3 [1-6]		3 [1-6]		2 [0-5]	-1 [-1;-1]
Mental or Behavioral Dysfunction	1 [0-4]	1 [0-3]	[;-1]	1 [-4]		1 [0-4]		1 [0-3]	[;-1]
Neoplastic Process	1 [0-3]	0 [0-3]	-1 [;-1]	0 [-3]	-1 [;-1]	1 [0-3]	[;-1]	0 [0-2]	-1 [;-2]
Anatomical Abnormality	1 [0-2]	1 [0-2]	[;-2]	1 [-2]	[;-2]	1 [0-2]	[;-2]	1 [0-2]	[;-2]
Congenital Abnormality	1 [0-1]	1 [0-1]	[;-3]	1 [-1]	[;-3]	1 [0-1]	[;-3]	1 [0-1]	[;-3]

Tableau 11 : Impact de la filtration avec EDSNLP

Type sémantique	Sans filtrage	Négation		Hypothèse		Non Patient		Négation + Hypothèse + Non Patient	
	Médiane [25%-75%]	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif	Médiane [25%-75%]	Dif
Total (Distribution globale)*	**94 [28-199]**	**82 [25-178]**	**-12 [-3;-21]**	**89 [27-189]**	**-5 [-1;-10]**	**93 [28-197]**	**-1 [;-2]**	**77 [23-166]**	**-17 [-5;-33]**
Disease or Syndrome	16 [5-40]	13 [4-33]	-3 [-1;-7]	15 [4-37]	-1 [-1;-3]	16 [5-39]	[;-1]	12 [3-29]	-4 [-2;-11]
Therapeutic or Preventive Procedure	14 [4-32]	13 [3-30]	-1 [-1;-2]	14 [4-32]		14 [4-32]		13 [3-29]	-1 [-1;-3]
Finding	10 [3-24]	9 [2-21]	-1 [-1;-3]	10 [2-23]	[-1;-1]	10 [3-24]		8 [2-20]	-2 [-1;-4]
Diagnostic Procedure	11 [4-22]	10 [3-21]	-1 [-1;-1]	11 [4-21]	[;-1]	11 [4-22]		10 [3-20]	-1 [-1;-2]
Sign or Symptom	7 [2-15]	5 [1-12]	-2 [-1;-3]	6 [1-15]	-1 [-1;]	7 [2-15]		5 [1-12]	-2 [-1;-3]
Pathologic Function	7 [2-15]	5 [1-11]	-2 [-1;-4]	6 [2-14]	-1 [;-1]	7 [2-15]		4 [1-10]	-3 [-1;-5]
Pharmacologic Substance	4 [1-8]	3 [1-7]	-1 [;-2]	3 [1-7]	-1 [;-2]	4 [1-8]	[;-1]	3 [1-7]	-1 [;-2]
Injury or Poisoning	3 [1-6]	2 [0-5]	-1 [-1;-1]	2 [1-5]	-1 [;-1]	3 [1-6]		2 [0-4]	-1 [-1;-2]
Mental or Behavioral Dysfunction	1 [0-4]	1 [0-3]	[;-1]	1 [0-4]	[;]	1 [0-4]		1 [0-3]	[;-1]
Neoplastic Process	1 [0-3]	0 [0-2]	-1 [;-2]	0 [0-3]	-1 [;-1]	1 [0-3]	[;-1]	[0-2]	-1 [;-2]
Anatomical Abnormality	1 [0-2]	1 [0-2]	[;-2]	1 [0-2]	[;-2]	1 [0-2]	[;-2]	1 [0-2]	[;-2]
Congenital Abnormality	1 [0-1]	1 [0-1]	[;-3]	1 [0-1]	[;-3]	1 [0-1]	[;-3]	1 [0-1]	[;-3]

3.5 Mesures d'impact en vie réelle : Cohorte ArthroVIH

3.5.1 Description du jeu de données ArthroVIH

Un total de 2 316 patients a été identifié pour 68 132 séjours, 112 036 documents à partir de la requête initialement faite par les auteurs de l'étude ArthroVIH. Pour ces patients on retrouve en moyenne 29,4 séjours (ET = 22,7) et 48,4 documents (ET=39,3). Il a été au total 416 139 entités à partir de la liste de mots-clés décrite en annexe 6. On retrouvait en moyenne 3,7 entités (ET = 5,3) par document, 6,1 (ET = 11,5) par séjour et 179,6 (ET=203,0) par patient. 416 139 entités ont été identifié dans l'ensemble du corpus de document. En moyenne il a été retrouvé 29,4 (ET = 22,7) séjours par patients. Ces résultats montrent une forte variabilité interindividuelle, avec certains patients ayant un grand nombre d'entités identifiées. L'ensemble des résultats sont présenté dans le Tableau 12.

Tableau 12 : Description du jeu de données ArthroVIH

Métrique		Entités N = 416 139	Documents N = 112 036	Séjours N = 68 132	Patients N = 2 316
Par document			-	-	-
	Moyenne (ET)	3,7 (5,3)	-	-	-
	Médiane [25%-75%]	2 [1-4]	-	-	-
	Min ; Max	1 ; 213	-	-	-
Par séjour				-	-
	Moyenne (ET)	6,1 (11,5)	1,6 (1,6)	-	-
	Médiane [25%-75%]	3 [2-6]	1 [1-2]	-	-
	Min ; Max	1 ; 838	1 ; 62	-	-
Par patient					-
	Moyenne (ET)	179,6 (203)	48,4 (39,3)	29,4 (22,7)	-
	Médiane [25%-75%]	126 [51-238]	41 [20-67]	26 [12-41]	-
	Min ; Max	1 ; 4 455	1 ; 556	1 ; 151	-

3.5.2 Capacité de filtration et performances

L'analyse de l'impact de la filtration par le contexte a été réalisée à partir d'une population source de 2 316 patients pour lesquels on retrouve 131 869 séjours et 810 087 documents.

L'extraction des entités via le NER a permis d'identifier un total de 416 139 entités, réparties en 309 512 entités de diagnostic VIH (74,4 %), 74 578 plaintes douloureuses (17,9 %) et 38 970 entités de localisation douloureuse (9,4 %). Ces entités provenaient de 112 036 documents, dont 79 851 contenaient au moins un concept de diagnostic VIH, 36 110 une plainte douloureuse et 19 747 une localisation douloureuse. Concernant les séjours, 68 132 ont été analysés, parmi lesquels 53 183 comportaient un concept de diagnostic VIH, 23 183 une plainte douloureuse et 13 738 une localisation douloureuse. Au niveau des patients, sur les 2 316 inclus, 2 225 présentaient au moins un concept de diagnostic VIH, 2 098 une plainte douloureuse et 2 094 un concept de localisation douloureuse. Parmi ces patients, 367 ne remplissaient pas les trois critères d'inclusion, réduisant ainsi la population source à une population pré-filtrée de 1 949 patients, 3 154 séjours et 2 776 documents répondant aux

critères d'inclusion. Tous les patients n'avaient pas nécessairement les trois critères, car nous n'avions pas accès aux codes CIM-10 des patients. Par ailleurs, des écarts dans la détection des concepts peuvent être attribués à l'utilisation de NER différents entre l'EDS et la chaîne de traitement de cette étude. Le NER de l'EDS utilise notamment un algorithme de reconnaissance type fuzzy matching, permettant de corriger certaines fautes d'orthographe et de gérer des variantes lexicales, telles que l'ajout ou la suppression d'un "s" en fin de mot, ce qui peut expliquer certaines divergences.

Avec une stratégie de filtration par la négation, l'hypothèse et le non patient, EDS-NLP a conduit à l'exclusion de 131 patients (-7 %), réduisant ainsi la population filtrée à 1 818 patients. De plus, le nombre de séjours et de documents a été réduit respectivement à 2 622 (-17 %) et 2 282 (-18 %). On observe que les concepts liés aux plaintes douloureuses sont ceux ayant subi le plus fort impact, avec 20 % des entités associées à une modalité, contre 8 % pour les localisations douloureuses et 6 % pour le diagnostic VIH. Au total, 3 patients (0 %) n'avaient plus aucun concept, 2 232 séjours (-3 %) ne contenaient plus d'entité identifiable et 4 687 documents (-4 %) ne comportaient plus aucun concept après filtration. L'application de la filtration contextuelle avec MedSpaCy a conduit à l'exclusion de 85 patients (-4 %), réduisant ainsi la population filtrée à 1 864 patients. De plus, le nombre de séjours et de documents a été réduit respectivement à 2 810 (-11 %) et 2 446 (-12 %). Comme pour EDS-NLP, les concepts liés aux plaintes douloureuses ont été les plus impactés, avec 14 % des entités associées à une modalité, contre 3 % pour les localisations douloureuses et 2 % pour le diagnostic VIH. Au total, 2 patients (0 %) n'avaient plus aucun concept, 1 173 séjours (-2 %) ne comportaient plus d'entité identifiable et 2 479 documents (-2 %) ne contenaient plus aucun concept après application des filtres. Ainsi, EDS-NLP a montré une capacité de filtration plus importante que MedSpaCy, ce qui était attendu compte tenu du rappel plus élevé précédemment mesuré. Cependant, malgré cette capacité de filtration, aucune des deux bibliothèques ne parvient à atteindre la population cible de 706 patients, soit une perte de 64 % de la population pré-filtrée. Ces résultats suggèrent que, malgré l'apport de la filtration contextuelle, un tri manuel des données reste indispensable. L'automatisation complète du processus d'identification des patients apparaît donc impossible dans ces conditions. Les résultats de cette stratégie sont présentés en figure 14 et l'ensemble des résultats pour chaque stratégie est présenté en annexe 7.

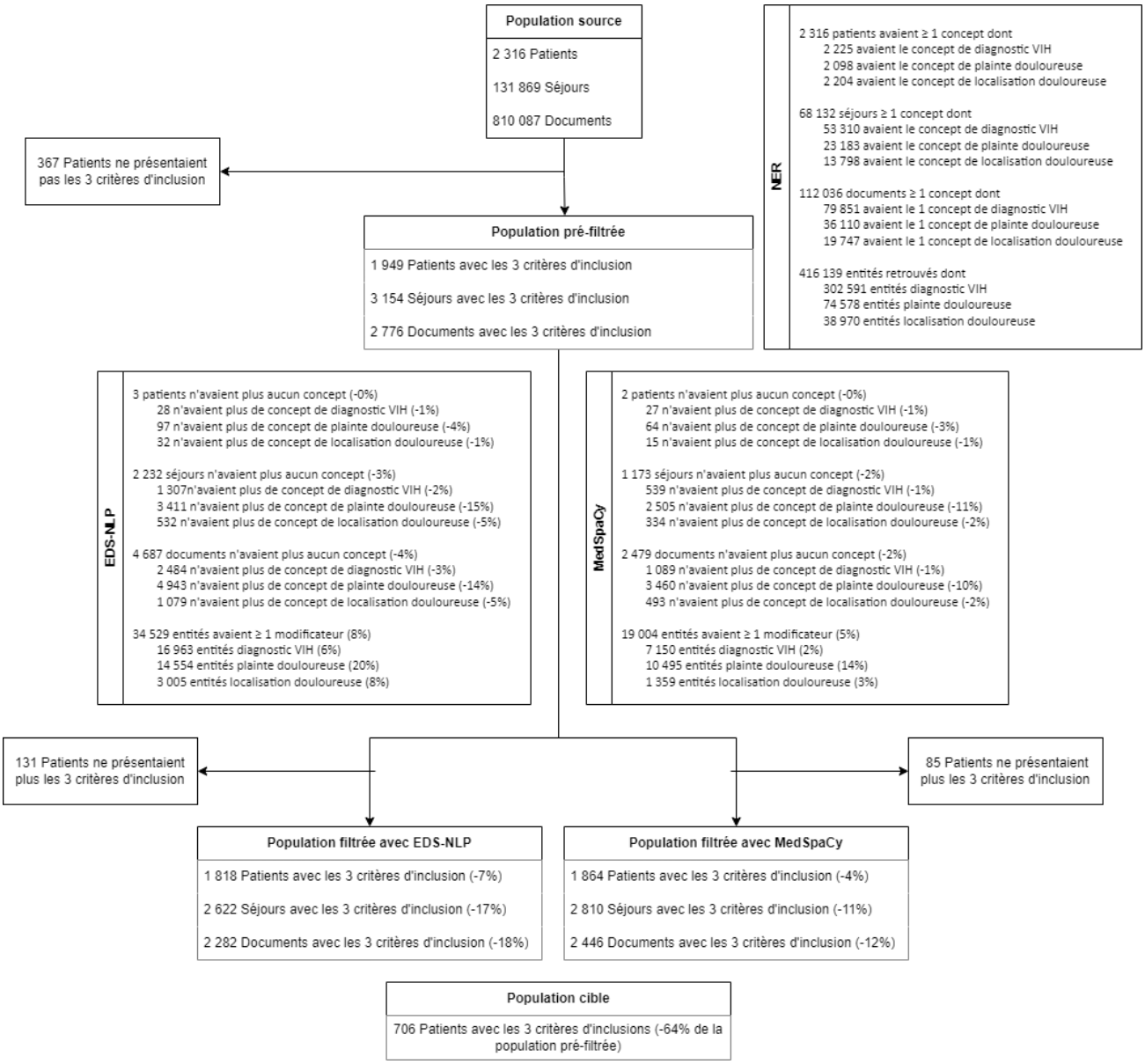


Figure 14 : Diagramme de flux de la cohorte ArthroVIH par EDS-NLP et MedSpaCy avec la stratégie de filtration par la négation, l'hypothèse et le non patient

L'évaluation de la capacité de filtration des différentes stratégies de filtrage a été réalisée pour EDS-NLP et MedSpaCy, en analysant l'impact sur le nombre de patients, séjours et documents filtrés. Les résultats sont détaillés dans le tableau 13. L'application du filtre de négation seul a conduit à l'exclusion de 95 patients (5 %), 420 séjours (13 %) et 402 documents (14 %) avec EDS-NLP, tandis que pour MedSpaCy, ce filtre a entraîné l'exclusion de 56 patients (3 %), 255 séjours (8 %) et 251 documents (9 %). Le filtre d'hypothèse seul a eu un impact moindre, avec une exclusion de 26 patients (1 %), 112 séjours (4 %) et 87 documents (3 %) avec EDS-NLP, et une filtration de 28 patients (1 %), 84 séjours (3 %) et 70 documents (3 %) avec MedSpaCy. L'application du filtre non-patient seul a eu un effet très limité, éliminant 2 patients (<1 %), 8 séjours (<1 %) et 8 documents (<1 %) avec EDS-NLP, contre 3 patients (<1 %), 18 séjours (1 %) et 13 documents (<1 %) avec MedSpaCy. L'utilisation conjointe de la négation et de l'hypothèse a entraîné une augmentation marquée de la filtration, avec une exclusion de 129 patients (7 %), 524 séjours (17 %) et 487 documents (18 %) avec EDS-NLP, tandis que pour MedSpaCy, cette combinaison a filtré 84 patients (4 %), 336 séjours (11 %) et 321 documents (12 %). L'association de la négation et du filtre non-patient a également renforcé la filtration, avec une suppression de 97 patients (5 %), 428 séjours (14 %) et 409 documents (15 %) avec EDS-NLP, tandis que MedSpaCy a exclu 57 patients (3 %), 263 séjours (8 %) et 260 documents (9 %). Enfin, l'application simultanée des trois filtres (négation, hypothèse et non-patient) a abouti à la plus forte réduction, excluant 131 patients (7 %), 532 séjours (17 %) et 494 documents (18 %) avec EDS-NLP. Pour MedSpaCy, cette combinaison a filtré 85 patients (4 %), 344 séjours (11 %) et 330 documents (12 %). Ces résultats montrent que la filtration combinée des négation, hypothèse et non-patient a un effet additif, augmentant la suppression des concepts extraits. EDS-NLP a un impact plus important sur la filtration des patients, séjours et documents par rapport à MedSpaCy, bien que la capacité de filtration éloignée de celle effectuée manuelle par les investigateurs de l'étude ArthroVIH (-64%). L'ensemble des erreurs en fonction de l'atteinte retrouvée au sein de l'étude est présentée en annexe 8.

Tableau 13 : Capacité de filtration des différentes Stratégies de Filtrage avec EDS-NLP et MedSpaCy (N Total Patients = 1949 ; N Total Séjours = 3154 ; N Total Documents = 2776)

Stratégie de Filtrage	EDS-NLP						MedSpaCy					
	Patients		Séjours		Documents		Patients		Séjours		Documents	
	N	%	N	%	N	%	N	%	N	%	N	%
Négation	95	5	420	13	402	14	56	3	255	8	251	9
Hypothèse	26	1	112	4	87	3	28	1	84	3	70	3
Non patient	2	0	8	0	8	0	2	0	8	0	10	0
Négation + Hypothèse	129	7	524	17	487	18	84	4	336	11	321	12
Négation + Non patient	97	5	428	14	409	15	57	3	263	8	260	9
Négation + Hypothèse + Non patient	131	7	532	17	494	18	85	4	344	11	330	12

L'évaluation des performances des différentes stratégies de filtration a été réalisée en analysant le nombre de patients filtrés, le taux de faux positifs (FP) et la précision pour EDS-NLP et MedSpaCy. Les résultats sont détaillés dans le tableau 12. L'application du filtre de

négation seul a conduit à l'exclusion de 95 patients (5 %) avec EDS-NLP, avec 18 faux positifs et une précision de 0,81. Pour MedSpaCy, ce filtre a exclu 56 patients (3 %) avec 12 faux positifs et une précision légèrement inférieure de 0,79. Le filtre d'hypothèse seul a eu un impact plus limité, avec une filtration de 26 patients (1 %) pour EDS-NLP, entraînant 4 faux positifs et une précision de 0,85. Pour MedSpaCy, cette stratégie a exclu 28 patients (1 %) avec 2 faux positifs et une précision plus élevée de 0,93, indiquant une meilleure sélectivité de cet outil pour ce filtre. L'application du filtre non-patient seul a eu un effet très faible, excluant 2 patients (<1 %) avec EDS-NLP, sans faux positifs, mais avec une précision faible de 0,50. MedSpaCy a exclu 3 patients (<1 %), générant 1 faux positif, avec également une précision de 0,50. Ces résultats montrent que ce filtre, utilisé isolément, manque de fiabilité et entraîne des exclusions limitées. L'utilisation conjointe de la négation et de l'hypothèse a entraîné une augmentation du nombre de patients filtrés, avec 129 patients exclus (7 %) pour EDS-NLP, 25 faux positifs et une précision de 0,81. Pour MedSpaCy, cette combinaison a filtré 84 patients (4 %), générant 15 faux positifs avec une précision légèrement supérieure de 0,82. L'association de la négation et du filtre non-patient a entraîné l'exclusion de 97 patients (5 %) avec EDS-NLP, avec 19 faux positifs et une précision de 0,80. Pour MedSpaCy, cette combinaison a exclu 57 patients (3 %), avec 12 faux positifs et une précision similaire de 0,79. Enfin, l'application simultanée des trois filtres (négation, hypothèse et non-patient) a abouti à la plus forte filtration, avec 131 patients exclus (7 %) pour EDS-NLP, 26 faux positifs et une précision de 0,80. Pour MedSpaCy, cette combinaison a exclu 85 patients (4 %), générant 15 faux positifs, avec une précision de 0,82. Ces résultats montrent que MedSpaCy tend à produire moins de faux positifs, avec des précisions légèrement supérieures à celles de EDS-NLP, notamment pour le filtre d'hypothèse. En revanche, EDS-NLP a un impact plus important sur la filtration des patients, entraînant une suppression plus marquée des entités extraites.

Tableau 14 : Performances des différentes Stratégies de Filtrage avec EDS-NLP et MedSpaCy (N Total = 1949 patients)

Stratégie de Filtrage	EDS-NLP				MedSpaCy			
	N Patients	% Total	FP*	Précision	N Patients	% Total	FP*	Précision
Négation	95	5	18	0,81	56	3	12	0,79
Hypothèse	26	1	4	0,85	28	1	2	0,93
Non patient	2	0	1	0,50	2	0	1	0,50
Négation + Hypothèse	129	7	25	0,81	84	4	15	0,82
Négation + Non patient	97	5	19	0,80	57	3	12	0,79
Négation + Hypothèse + Non patient	131	7	26	0,80	85	4	15	0,82

* Faux positifs : Correspond au nombre de patients filtré à tort.

4 Discussion

L'évaluation des bibliothèques EDS-NLP et MedSpaCy a révélé une variabilité significative des performances, avec des scores F1 allant de 0,11 à 0,81 selon la modalité linguistique analysée. Globalement, leurs performances sont relativement similaires : EDS-NLP présente un meilleur rappel, tandis que MedSpaCy affiche une meilleure précision. Toutefois, à l'échelle du document, certaines erreurs d'EDS-NLP sont corrigées par la redondance des informations, ce qui améliore les performances globales. Concernant la négation, nos résultats sont cohérents avec ceux rapportés dans la littérature, où EDS-NLP avait obtenu un F1-score de 71 % sur le jeu CAS/ESSAI et de 88 % sur NegParHyp. Il en est de même pour la modalité hypothèse, nos scores sont similaires aux références connues (47% contre 49 % sur CAS/ESSAI et 52 % sur NegParHyp) (58). Nous n'avons pas trouvé de mesures de performance pour les autres modalités avec EDS-NLP. A l'inverse, les performances annoncées par les développeurs des règles de FastContext sont nettement supérieures à celles obtenues dans notre étude. La première version des règles de négation affichait un F1-score de 78 % sur des dossiers patients, et la seconde version (celle utilisée ici) atteignait 95 % (55) . Or, ces valeurs n'ont pas été retrouvées dans notre évaluation. Plusieurs éléments peuvent expliquer ces divergences. D'une part, les définitions des modalités linguistiques ne sont pas toujours bien établies, ce qui peut entraîner une variabilité dans l'interprétation des annotations et une hétérogénéité dans l'évaluation des algorithmes. D'autre part, la qualité du gold standard peut constituer un facteur limitant. Bien que certaines études aient rapporté des coefficients de Kappa de Cohen satisfaisants et des accords interannotateurs élevés (55), nos annotateurs ont rencontré des difficultés lors de l'annotation des textes, et ce, pour l'ensemble des modalités linguistiques. De plus, le sens clinique de la notion d'« historique » demeure sujet à débat, notamment la règle de 14 jours pour définir un antécédent qui semble discutable et mérite d'être réévaluée. Au final, ces algorithmes ne permettent par définition de ne repérer que les mentions explicites des modalités ce qui est vision restreinte de la notion de contexte. Un autre facteur essentiel mis en évidence dans notre étude est la dépendance de la qualité de la détection du contexte à la reconnaissance des entités nommées (NER). En effet, un NER de mauvaise qualité entraîne une classification contextuelle erronée. Cette corrélation directe souligne la nécessité d'améliorer la reconnaissance des entités avant d'aborder la détection du contexte. Par ailleurs, la nature des documents analysés joue un rôle déterminant dans les performances obtenues. Nos mesures ont été effectuées exclusivement sur des comptes rendus d'hospitalisation relativement longs, en raison du système de score utilisé, alors que notre base de données contient plus de 600 types de documents différents. Cette hétérogénéité complexifie l'analyse et pourrait influencer les performances des algorithmes. Enfin, un point critique de notre étude est la difficulté de détecter correctement la négation au-delà de quatre entités par document. Cette limitation restreint considérablement leur application en conditions réelles dans un EDS. En vie réelle, sur la cohorte ArthroVIH, nous avons observé une réduction de 5 à 10 % du nombre de patients détectés comme conformes aux critères de recherche après application des algorithmes de détection de contexte. Toutefois, cela ne permet pas encore d'automatiser intégralement l'identification des patients. Une approche combinée pourrait néanmoins s'avérer bénéfique : en ne sélectionnant que les documents où les trois concepts d'intérêt apparaissent, nous pourrions réduire le nombre de documents à examiner de 10 à 20 %, offrant ainsi un gain de temps

significatif pour les investigateurs. Si l'application de la détection de la négation en conditions réelles reste incertaine, les performances des autres modalités sont clairement insuffisantes pour une exploitation en EDS. Des approches supervisées, basées sur l'apprentissage profond ou l'utilisation de modèles de langage de grande taille (LLM), pourraient être explorées. Ces méthodes permettraient d'aller au-delà de l'analyse contextuelle selon un nombre restreint de modalités fixes et de certifier directement si un concept représente une affection réelle du patient. D'après notre étude, qui constitue la première évaluation des algorithmes de détection de contexte à plusieurs échelles (du concept au patient), nous recommandons d'adopter une approche prudente quant à leur utilisation dans un EDSH. Ces outils présentent des performances encore insuffisantes pour une utilisation dans des cas où l'exhaustivité des résultats est indispensable, comme la constitution de cohortes complètes ou la production d'indicateurs de santé publique robustes. En revanche, ils pourraient être envisagés dans des contextes où l'exhaustivité n'est pas requise, par exemple pour identifier une population cible restreinte ou présélectionner des documents pertinents dans le cadre de recherches exploratoires ou d'études de faisabilité. Dans ces cas d'usage, les algorithmes peuvent offrir un gain de temps appréciable, à condition d'être associés à un système NER performant. Par ailleurs, l'impact en termes de stockage ne doit pas être négligé. L'annotation contextuelle des documents peut générer un volume important de données, ce qui soulève des questions d'infrastructure et de coûts qu'il convient d'évaluer en amont de tout déploiement. Enfin, l'implémentation de tels outils nécessite un accompagnement adapté des utilisateurs finaux. Il est indispensable de les sensibiliser aux limites des algorithmes, de les former à leur usage et de développer des recommandations opérationnelles afin de garantir une exploitation maîtrisée et pertinente des résultats au sein des EDS.

5 Bibliographie

1. Hansell A, Bottle A, Shurlock L, Aylin P. Accessing and using hospital activity data. J Public Health Med. 2001 Mar;23(1):51–6.

2. Griffier R, Jouhet V, Thiessard F, Cossin S. Identification des verrous et des leviers à la réutilisation secondaire des données dans un établissement de santé. Revue d'Épidémiologie et de Santé Publique. 2020 Mar 1;68:S49–50.

3. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020 Feb;145(2):463–9.

4. Griffier R, Jouhet V, Thiessard F, Cossin S. Identification des verrous et des leviers à la réutilisation secondaire des données dans un établissement de santé. Revue d'Épidémiologie et de Santé Publique. 2020 Mar 1;68:S49–50.

5. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014 Nov 5;14(1):1144.

6. Le règlement général sur la protection des données - RGPD [Internet]. [cited 2025 Jan 29]. Available from: https://www.cnil.fr/fr/reglement-europeen-protection-donnees

7. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

8. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014 Nov 5;14:1144.

9. Sreemathy J, Joseph V. I, Nisha S, Prabha I. C, Priya R.M. G. Data Integration in ETL Using TALEND. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020 Mar;1444–8.

10. Griffier R. Intégration et utilisation secondaire des données de santé hospitalières hétérogènes : des usages locaux à l'analyse fédérée [Internet] [phdthesis]. Université de Bordeaux; 2024 [cited 2025 Feb 3]. Available from: https://theses.hal.science/tel-04880743

11. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.

12. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600–6.

13. Grosjean J, Pressat-Laffouilhère T, Ndangang M, Leroy JP, Darmoni SJ. Using Clinical Data Warehouse to Optimize the Vaccination Strategy Against COVID-19: A Use Case in France. Stud Health Technol Inform. 2022 Jun 6;290:150–3.

14. 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.

15. Loffler A. Utiliser un entrepôt de données de santé pour reproduire les résultats d'une cohorte vaccinale COVID-19 chez des patients atteints d'hypogammaglobulinémie. 2024 Nov 12;74.

16. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform. 2019 Sep;129:13–9.

17. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci Proc. 2014;2014:218–23.

18. Escudié JB, Rance B, Malamut G, Khater S, Burgun A, Cellier C, et al. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. BMC Med Inform Decis Mak. 2017 Sep 29;17(1):140.

19. Salmasian H, Freedberg DE, Abrams JA, Friedman C. An automated tool for detecting medication overuse based on the electronic health records. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):183–9.

20. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. Journal of Biomedical Informatics. 2005 Aug 1;38(4):314–21.

21. Shiner B, Neily J, Mills PD, Watts BV. Identification of Inpatient Falls Using Automated Review of Text-Based Medical Records. J Patient Saf. 2020 Sep;16(3):e174–8.

22. 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 Assoc. 2017 May 1;24(3):607–13.

23. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak. 2019 Oct 28;19(1):202.

24. Setia MS. Methodology Series Module 1: Cohort Studies. Indian J Dermatol. 2016;61(1):21–5.

25. Dalloux C, Claveau V, Grabar N. Détection de la négation : corpus français et apprentissage supervisé. In: SIIM 2017 - Symposium sur l'Ingénierie de l'Information Médicale [Internet]. Toulouse, France; 2017 [cited 2024 Apr 10]. p. 1–8. Available from: https://hal.science/hal-01659637

26. Salles K. Description épidémiologique des causes de douleurs des mains chez les patients suivis pour une infection par le VIH au CHU de Bordeaux (étude ArthroVIH). 2023 Oct 6;33.

27. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Comput Surv. 2002 Mar;34(1):1–47.

28. 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 Oct;34(5):301–10.

29. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 Oct 1;42(5):839–51.

30. Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français.

31. Deléger L, Grouin C. Detecting negation of medical problems in French clinical notes. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium [Internet]. New York, NY, USA: Association for Computing Machinery; 2012 [cited 2025 Feb 17]. p. 697–702. (IHI '12). Available from: https://doi.org/10.1145/2110363.2110443

32. Jonquet C, Shah NH, Musen MA. The Open Biomedical Annotator. Summit on Translat Bioinforma. 2009 Mar 1;2009:56–60.

33. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc. 2007;14(3):304–11.

34. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak. 2005 May 5;5:13.

35. 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 Assoc. 2017 May 1;24(3):607–13.

36. Rokach L, Romano R, Maimon O. Negation recognition in medical narrative reports. Inf Retrieval. 2008 Dec 1;11(6):499–538.

37. Morante R, Daelemans W. A Metalearning Approach to Processing the Scope of Negation. In: Stevenson S, Carreras X, editors. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009) [Internet]. Boulder, Colorado: Association for Computational Linguistics; 2009 [cited 2025 Feb 17]. p. 21–9. Available from: https://aclanthology.org/W09-1105/

38. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552–6.

39. Shivade C. MedNLI - A Natural Language Inference Dataset For The Clinical Domain [Internet]. PhysioNet; [cited 2025 Feb 17]. Available from: https://physionet.org/content/mednli/

40. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008 Nov 19;9 Suppl 11(Suppl 11):S9.

41. Goryachev S, Sordo M, Zeng Q, Ngo LH. Implementation and Evaluation of Four Different Methods of Negation Detection. In 2007 [cited 2025 Feb 17]. Available from: https://www.semanticscholar.org/paper/Implementation-and-Evaluation-of-Four-Different-of-Goryachev-Sordo/49517539055234e73bfa6140a7a84b74cfc12685

42. Ji Y, Yu Z, Wang Y. Assertion Detection in Clinical Natural Language Processing Using Large Language Models. In: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) [Internet]. 2024 [cited 2025 Feb 17]. p. 242–7. Available from: https://ieeexplore.ieee.org/document/10628639

43. Olsson C, Elhage N, Nanda N, Joseph N, DasSarma N, Henighan T, et al. In-context Learning and Induction Heads [Internet]. arXiv; 2022 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2209.11895

44. 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):1234–40.

45. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission [Internet]. arXiv; 2020 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/1904.05342

46. Touchent R, Romary L, De La Clergerie E. CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In: Servan C, Vilnat A, editors. Actes de CORIA-TALN 2023 Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), \ volume 1 : travaux de recherche originaux – articles longs [Internet]. Paris, France: ATALA; 2023 [cited 2025 Feb 17]. p. 323–34. Available from: https://hal.science/hal-04130187

47. « Intelligence artificielle, données, calculs : quelles infrastructures dans un monde décarboné ? » : The Shift Project publie son rapport intermédiaire – The Shift Project [Internet]. [cited 2025 Mar 13]. Available from: https://theshiftproject.org/article/rapport-intermediaire-ia/

48. Efficient extraction of medication information from clinical notes: an evaluation in two languages [Internet]. [cited 2025 Feb 19]. Available from: https://arxiv.org/html/2502.03257v1#bib.bib22

49. Cossin S. scossin/IAMsystem [Internet]. 2024 [cited 2025 Feb 19]. Available from: https://github.com/scossin/IAMsystem

50. Shi J. jianlins/FastContext [Internet]. 2024 [cited 2025 Feb 19]. Available from: https://github.com/jianlins/FastContext

51. Wajsburt P, Petit-Jean T, Dura B, Cohen A, Jean C, Bey R. EDS-NLP: efficient information extraction from French clinical notes [Internet]. Zenodo; 2024 [cited 2025 Feb 19]. Available from: https://zenodo.org/doi/10.5281/zenodo.6424993

52. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2022 Feb 21;2021:438–47.

53. spaCy · Industrial-strength Natural Language Processing in Python [Internet]. [cited 2025 Feb 19]. Available from: https://spacy.io/

54. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 Jan 1;32(suppl_1):D267–70.

55. Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: A publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform. 2021 May;117:103733.

56. practikpharma/FrenchFastContext [Internet]. PractiKPharma Project; 2020 [cited 2025 Feb 26]. Available from: https://github.com/practikpharma/FrenchFastContext

57. Tokenizers - EDS-NLP [Internet]. [cited 2025 Mar 6]. Available from: https://aphp.github.io/edsnlp/latest/tokenizers/

58. Redjdal A. Oncolog-IA : symbolic and numeric artificial intelligence for learning complexity of breast cancer cases and providing decision support for their therapeutic management [Internet] [phdthesis]. Sorbonne Université; 2023 [cited 2025 Mar 26]. Available from: https://theses.hal.science/tel-04330919

6 Annexes

Annexe 1. Comparaison des tokenizers de EDS-NLP et SpaCy français (https://aphp.github.io/edsnlp/latest/tokenizers/)

Tokenizers

In addition to the standard spaCy `FrenchLanguage` (`fr`), EDS-NLP offers a new language better fit for French clinical documents: `EDSLanguage` (`eds`). Additionally, the `EDSLanguage` document creation should be around 5-6 times faster than the `fr` language. The main differences lie in the tokenization process.

A comparison of the two tokenization methods is demonstrated below:

Example	FrenchLanguage	EDSLanguage
`ACR5`	`[ACR5]`	`[ACR, 5]`
`26.5/`	`[26.5/]`	`[26.5, /]`
`\n \n CONCLUSION`	`[\n \n, CONCLUSION]`	`[\n, \n, CONCLUSION]`
`l'artère`	`[l', artère]`	`[l', artère]` (same)
`Dr. Pichon`	`[Dr, ., Pichon]`	`[Dr., Pichon]`
`B.H.HP.A.7.A`	`[B.H.HP.A.7.A]`	`[B., H., HP., A, 7, A, 0]`

Annexe 2. Extraits des règles *French Context* modifié en .json pour MedSpaCy


```
{
    "context_rules": [
        {
            "literal": "\u00e0 cause de",
            "max_scope": 30,
            "category": "HYPOTHETICAL",
            "direction": "TERMINATE"
        },
        {
            "literal": "a \u00e9cart\u00e9 le patient",
            "max_scope": 30,
            "category": "NEGATED_EXISTENCE",
            "direction": "FORWARD"
        },
        {
            "literal": "freres",
            "max_scope": 30,
            "category": "FAMILY",
            "direction": "FORWARD"
        },
        {
            "literal": "pas d'ant\u00e9c\u00e9dent",
            "max_scope": 30,
            "category": "HISTORICAL",
            "direction": "PSEUDO"
        },
        {
            "literal": "il y a \\> 0 mois",
            "max_scope": 30,
            "category": "HISTORICAL",
            "direction": "BACKWARD"
        },
        {
            "literal": "peut \u00eatre refus\u00e9",
            "max_scope": 30,
            "category": "POSSIBLE_EXISTENCE",
            "direction": "FORWARD"
        },… }
```

Annexe 3. Guide d'annotation pour l'identification de modalités linguistiques de concepts dans les documents médicaux

Guide d'annotation pour l'identification de modalités linguistiques de concepts dans les documents médicaux

Objectif :

Ce guide vise à fournir des instructions pour l'identification et l'étiquetage des modalités linguistiques de concepts présents dans les documents médicaux préalablement identifiés par un algorithme de Reconnaissance d'Entités Nommées (NER).

Instructions générales :

Interface d'annotation : Vous utiliserez l'outil d'étiquetage Doccano pour effectuer vos annotations.
Tâche : Votre tâche consiste à lire attentivement chaque document et déterminer s'il y a une présence d'une modalité ou non pour chaque concept identifié "Positif". Si nécessaire, vous devrez remplacer cette étiquette par le/les labels présentés ci-dessous.
Assignation d'un label : Pour assigner un label, il faut cliquer sur le concept "Positif" puis cliquer sur le bon label correspondant. Celui-ci sera traitée lors d'une tache future. A la fin de l'annotation d'un document, merci d'appuyer sur la croix en haut à gauche pour que celui-ci soit noté "checked".
Utilisation de la souris : Il est possible d'utiliser le clavier comme raccourci pour assigner des labels, cependant, il est demandé de privilégier l'utilisation de la souris car il est possible d'effacer accidentellement le concept (la ré-identification est possible mais doit être parfaite, notamment pour les concepts de plusieurs mots).
Concepts particuliers : Si il vous paraît insensé de labéliser un concept, mettez l'étiquette 'non_pertinent' ou si celui-ci fait partie de l'entête ou d'un document rapporté supplémentaire.

Conseils pour l'annotation :
Contexte : Assurez-vous de considérer tout le contexte fourni pour comprendre correctement le sens du concept médical.

Présentation des modalités

De base, une entité est définie comme positive, certaine, récente, expérimentée par le patient et rapporté par le rédacteur du document médical. L'étiquette "Affirmé" représentera ce concept.

Modalité de Négation

Négation : Indique que l'entité identifiée est explicitement nié, absente ou que l'interprétation, le sens de de cette entité est inversé.

Modalité de Certitude

Incertain : Indique explicitement un doute sur la présence de l'entité identifiée.

Modalité de Temporalité

Historique : Indique que l'entité identifiée a débuté il y a plus de 2 semaines par rapport à la date de la visite ou de l'évènement documenté dans le document médical ou que l'entité est explicitement décrite comme faisait partie des antécédents médicaux du patient.
ATTENTION POUR EVALUATEUR 1, EVALUATEUR 2 ET EVALUATEUR 3 UTILISEZ CETTE DEFINITION : Indique les entité explicitement identifiée comme faisant partie des antécédents médicaux et précédé par "antécédent médical" ou un synonyme.

Conditionnel : Indique que l'entité réfère à un scénario hypothétique.

Modalité d'Expérimentateur

Non patient : Indique si l'entité identifiée est expérimentée ou réfère à un autre individu que le patient lui-même (explicitement décrit).

Assertion de Discours rapporté

Discours rapporté : Indique que l'entité identifiée a été rapporté par le une autre personne que le rédacteur du document médical ou présenté comme ayant été dites par quelqu'un d'autre que le rédacteur du document médical. Cette personne doit être décrite.

Une entité peut avoir une ou plusieurs assertions (uniquement une par catégorie) ou aucune et rester "Affirmé"

Exemple :

L'entité est noté entre [crochets]
Le patient a une [infection] - Positive
Le patient n'a pas d'[infection] - Négation
Le patient a peut-être une [infection] - Incertain
ATCD : Le patient a eu une [infection] - Historique
Le patient a eu une [infection] il y a 1 mois- Historique
Si le patient a une [infection] - Conditionnel
Le frère du patient a une [infection] - Non patient
Le patient dit avoir une [infection] - Discours rapporté

Assistance :

Si vous avez des questions ou des préoccupations, n'hésitez pas à me contacter +
mettre un commentaire pour les documents pour lesquels vous auriez des questions grâce au bouton en dessus du texte.
Merci de votre contribution à ce projet !
Matisse DECILAP

Annexe 4. Temps d'exécution pour traiter le jeu d'exploration selon la chaine de traitement utilisé

Chaine de traitement utilisé	Nombre de documents	Nombre d'entités	Temps d'exécution total (N docs/s)
EDS-NLP + MedSpaCy (Toutes modalités)	10 000	205 877	1min42s (98 docs/s)
MedSpaCy (Toutes modalités)	10 000	205 877	1m02s (161 docs/s)
EDS-NLP (Toutes modalités)	10 000	205 877	1m26 (116 docs/s)
EDS-NLP (Négation)	10 000	205 877	59s (169 docs/s)
EDS-NLP (Hypothèse)	10 000	205 877	58s (172 docs/s)
EDS-NLP (Famille)	10 000	205 877	59s (169 docs/s)
EDS-NLP (Historique)	10 000	205 877	1m04s (156 docs/s)
EDS-NLP (Discours rapporté)	10 000	205 877	58s (172 docs/s)
EDS-NLP (Négation + hypothèse)	10 000	205 877	1m05s (154 docs/s)
EDS-NLP (Négation + hypothèse + Famille)	10 000	205 877	1m05s (154 docs/s)

Annexe 5 : Performances

Annexe 5.1 : Performances de EDS-NLP et MedSpaCy stratifié sur le type sémantique pour chaque modalité linguistique à l'échelle de l'entité

Modalité	Type sémantique	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	Acquired Abnormality	32	1,00	0,44	0,62	1,00	1,00	1,00
Negation	Anatomical Abnormality	93	1,00	0,79	0,88	1,00	1,00	1,00
Negation	Antibiotic	6	0,00	0,00	0,00	0,00	0,00	0,00
Negation	Cell or Molecular Dysfunction	1	0,00	0,00	0,00	0,00	0,00	0,00
Negation	Congenital Abnormality	61	1,00	0,80	0,89	0,50	1,00	0,67
Negation	Diagnostic Procedure	961	0,61	0,23	0,34	0,70	0,55	0,61
Negation	Disease or Syndrome	2025	0,94	0,71	0,81	0,78	0,82	0,80
Negation	Finding	786	0,93	0,73	0,81	0,74	0,82	0,78
Negation	Gene or Genome	8	0,00	0,00	0,00	0,00	0,00	0,00
Negation	Injury or Poisoning	274	0,89	0,73	0,80	0,87	0,87	0,87
Negation	Mental or Behavioral Dysfunction	171	0,93	0,57	0,70	0,64	0,64	0,64
Negation	Neoplastic Process	213	1,00	0,57	0,73	0,88	0,78	0,82
Negation	Pathologic Function	751	0,96	0,78	0,86	0,80	0,89	0,84
Negation	Pharmacologic Substance	121	1,00	0,43	0,60	0,83	0,71	0,77
Negation	Sign or Symptom	786	0,95	0,84	0,90	0,83	0,86	0,85
Negation	Therapeutic or Preventive Procedure	1263	0,79	0,52	0,63	0,76	0,69	0,72
Incertain	Acquired Abnormality	32	0,50	0,50	0,50	0,50	1,00	0,67
Incertain	Anatomical Abnormality	93	1,00	0,20	0,33	0,00	0,00	0,00
Incertain	Antibiotic	6	0,00	0,00	0,00	0,00	0,00	0,00
Incertain	Cell or Molecular Dysfunction	1	0,00	0,00	0,00	0,00	0,00	0,00
Incertain	Congenital Abnormality	61	1,00	0,10	0,18	0,00	0,00	0,00
Incertain	Diagnostic Procedure	961	0,67	0,05	0,10	0,33	0,03	0,06
Incertain	Disease or Syndrome	2025	0,74	0,36	0,49	0,45	0,31	0,36
Incertain	Finding	786	0,90	0,19	0,31	0,50	0,13	0,21
Incertain	Gene or Genome	8	0,50	1,00	0,67	0,00	0,00	0,00
Incertain	Injury or Poisoning	274	0,64	0,41	0,50	0,64	0,39	0,48
Incertain	Mental or Behavioral Dysfunction	171	1,00	0,38	0,55	1,00	0,60	0,75
Incertain	Neoplastic Process	213	0,50	0,25	0,33	0,67	0,40	0,50
Incertain	Pathologic Function	751	0,70	0,29	0,41	0,55	0,37	0,44
Incertain	Pharmacologic Substance	121	0,00	0,00	0,00	0,00	0,00	0,00
Incertain	Sign or Symptom	786	0,43	0,05	0,10	0,29	0,04	0,07
Incertain	Therapeutic or Preventive Procedure	1263	0,33	0,02	0,03	0,00	0,00	0,00
Historique	Acquired Abnormality	32	0,38	1,00	0,55	0,38	1,00	0,55
Historique	Anatomical Abnormality	93	0,27	0,91	0,42	0,22	0,89	0,35
Historique	Antibiotic	6	1,00	1,00	1,00	1,00	1,00	1,00
Historique	Cell or Molecular Dysfunction	1	0,00	0,00	0,00	0,00	0,00	0,00
Historique	Congenital Abnormality	61	0,31	1,00	0,47	0,38	1,00	0,56
Historique	Diagnostic Procedure	961	0,04	0,42	0,08	0,05	0,75	0,10
Historique	Disease or Syndrome	2025	0,10	0,78	0,19	0,09	0,76	0,16
Historique	Finding	786	0,23	0,67	0,34	0,06	0,87	0,12
Historique	Gene or Genome	8	0,00	0,00	0,00	0,00	0,00	0,00
Historique	Injury or Poisoning	274	0,10	0,83	0,17	0,06	0,75	0,11
Historique	Mental or Behavioral Dysfunction	171	0,05	0,75	0,09	0,06	0,80	0,12
Historique	Neoplastic Process	213	0,18	0,86	0,30	0,15	0,83	0,25
Historique	Pathologic Function	751	0,09	0,64	0,16	0,06	0,77	0,12
Historique	Pharmacologic Substance	121	0,00	0,00	0,00	0,00	0,00	0,00
Historique	Sign or Symptom	786	0,02	0,40	0,03	0,04	0,50	0,07
Historique	Therapeutic or Preventive Procedure	1263	0,09	0,63	0,16	0,05	0,76	0,10

Annexe 5.1 : Suite

Modalité	Type sémantique	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Conditionnel	Acquired Abnormality	32	0,00	0,00	0,00	0,00	0,00	0,00
Conditionnel	Anatomical Abnormality	93	0,00	0,00	0,00	0,00	0,00	0,00
Conditionnel	Antibiotic	6	0,00	0,00	0,00	0,00	0,00	0,00
Conditionnel	Cell or Molecular Dysfunction	1	0,00	0,00	0,00	0,00	0,00	0,00
Conditionnel	Congenital Abnormality	61	1,00	0,90	0,95	1,00	1,00	1,00
Conditionnel	Diagnostic Procedure	961	0,23	0,18	0,20	0,17	0,45	0,24
Conditionnel	Disease or Syndrome	2025	0,53	0,18	0,27	0,45	0,48	0,46
Conditionnel	Finding	786	0,73	0,23	0,35	0,73	0,38	0,50
Conditionnel	Gene or Genome	8	0,00	0,00	0,00	0,00	0,00	0,00
Conditionnel	Injury or Poisoning	274	0,25	0,06	0,10	0,25	0,20	0,22
Conditionnel	Mental or Behavioral Dysfunction	171	1,00	0,13	0,22	0,00	0,00	0,00
Conditionnel	Neoplastic Process	213	0,25	0,08	0,13	0,25	0,20	0,22
Conditionnel	Pathologic Function	751	0,63	0,25	0,36	0,37	0,70	0,48
Conditionnel	Pharmacologic Substance	121	0,00	0,00	0,00	1,00	0,20	0,33
Conditionnel	Sign or Symptom	786	0,65	0,60	0,62	0,61	0,78	0,68
Conditionnel	Therapeutic or Preventive Procedure	1263	0,36	0,15	0,21	0,20	0,19	0,20
Nonpatient	Acquired Abnormality	32	0,00	0,00	0,00	0,00	0,00	0,00
Nonpatient	Anatomical Abnormality	93	0,71	0,71	0,71	0,71	0,83	0,77
Nonpatient	Antibiotic	6	0,00	0,00	0,00	0,00	0,00	0,00
Nonpatient	Cell or Molecular Dysfunction	1	0,00	0,00	0,00	0,00	0,00	0,00
Nonpatient	Congenital Abnormality	61	0,56	0,83	0,67	0,78	0,88	0,82
Nonpatient	Diagnostic Procedure	961	0,43	0,23	0,30	0,43	0,43	0,43
Nonpatient	Disease or Syndrome	2025	0,75	0,52	0,61	0,68	0,50	0,58
Nonpatient	Finding	786	0,70	0,49	0,58	0,42	0,65	0,51
Nonpatient	Gene or Genome	8	0,50	0,33	0,40	1,00	1,00	1,00
Nonpatient	Injury or Poisoning	274	0,14	0,17	0,15	0,29	0,33	0,31
Nonpatient	Mental or Behavioral Dysfunction	171	0,75	0,50	0,60	0,25	0,14	0,18
Nonpatient	Neoplastic Process	213	0,78	0,74	0,76	0,75	0,84	0,79
Nonpatient	Pathologic Function	751	0,43	0,45	0,44	0,38	0,44	0,41
Nonpatient	Pharmacologic Substance	121	0,00	0,00	0,00	0,00	0,00	0,00
Nonpatient	Sign or Symptom	786	0,00	0,00	0,00	0,00	0,00	0,00
Nonpatient	Therapeutic or Preventive Procedure	1263	0,18	0,11	0,14	0,36	0,19	0,25
Discours rapporté	Acquired Abnormality	32	0,00	0,00	0,00			
Discours rapporté	Anatomical Abnormality	93	0,00	0,00	0,00			
Discours rapporté	Antibiotic	6	0,00	0,00	0,00			
Discours rapporté	Cell or Molecular Dysfunction	1	0,00	0,00	0,00			
Discours rapporté	Congenital Abnormality	61	0,00	0,00	0,00			
Discours rapporté	Diagnostic Procedure	961	0,00	0,00	0,00			
Discours rapporté	Disease or Syndrome	2025	0,71	0,10	0,17			
Discours rapporté	Finding	786	0,42	0,12	0,18			
Discours rapporté	Gene or Genome	8	0,00	0,00	0,00			
Discours rapporté	Injury or Poisoning	274	0,00	0,00	0,00			
Discours rapporté	Mental or Behavioral Dysfunction	171	0,00	0,00	0,00			
Discours rapporté	Neoplastic Process	213	0,00	0,00	0,00			
Discours rapporté	Pathologic Function	751	0,25	0,03	0,06			
Discours rapporté	Pharmacologic Substance	121	0,67	0,14	0,24			
Discours rapporté	Sign or Symptom	786	0,57	0,22	0,32			
Discours rapporté	Therapeutic or Preventive Procedure	1263	0,00	0,00	0,00			

Annexe 5.2 : Performances de EDS-NLP et MedSpaCy stratifié sur la fréquence d'apparition du concept au sein du corpus pour chaque modalité linguistique à l'échelle de l'entité

Modalité	Fréquence du concept au sein du corpus	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	[0%-1%]	1114	0,91	0,68	0,78	0,79	0,79	0,79
Negation	]1%-5%]	2427	0,93	0,67	0,78	0,79	0,79	0,79
Negation	]5%-10%]	1231	0,96	0,74	0,84	0,80	0,86	0,83
Negation	]10%-20%]	1126	0,91	0,68	0,78	0,78	0,85	0,82
Negation	>20%	1654	0,89	0,67	0,76	0,78	0,79	0,79
Incertain	[0%-1%]	1114	0,72	0,39	0,51	0,36	0,28	0,32
Incertain	]1%-5%]	2427	0,65	0,27	0,38	0,53	0,24	0,33
Incertain	]5%-10%]	1231	0,81	0,18	0,29	0,44	0,14	0,21
Incertain	]10%-20%]	1126	0,86	0,16	0,27	0,50	0,10	0,17
Incertain	>20%	1654	0,58	0,06	0,11	0,33	0,05	0,08
Historique	[0%-1%]	1114	0,12	0,77	0,21	0,11	0,82	0,19
Historique	]1%-5%]	2427	0,11	0,79	0,19	0,09	0,85	0,16
Historique	]5%-10%]	1231	0,08	0,63	0,14	0,07	0,75	0,12
Historique	]10%-20%]	1126	0,08	0,70	0,14	0,06	0,83	0,10
Historique	>20%	1654	0,15	0,63	0,24	0,06	0,60	0,10
Conditionnel	[0%-1%]	1114	0,54	0,18	0,27	0,31	0,44	0,37
Conditionnel	]1%-5%]	2427	0,51	0,27	0,35	0,45	0,47	0,46
Conditionnel	]5%-10%]	1231	0,50	0,19	0,27	0,57	0,62	0,59
Conditionnel	]10%-20%]	1126	0,39	0,18	0,24	0,36	0,26	0,30
Conditionnel	>20%	1654	0,58	0,26	0,36	0,46	0,59	0,52
Nonpatient	[0%-1%]	1114	0,53	0,52	0,52	0,63	0,58	0,61
Nonpatient	]1%-5%]	2427	0,63	0,58	0,60	0,67	0,61	0,64
Nonpatient	]5%-10%]	1231	0,53	0,31	0,39	0,53	0,39	0,45
Nonpatient	]10%-20%]	1126	0,94	0,38	0,54	0,82	0,45	0,58
Nonpatient	>20%	1654	0,83	0,45	0,59	0,23	0,32	0,27
Discours rapporté	[0%-1%]	1114	0,56	0,07	0,12			
Discours rapporté	]1%-5%]	2427	0,46	0,07	0,12			
Discours rapporté	]5%-10%]	1231	0,30	0,09	0,13			
Discours rapporté	]10%-20%]	1126	0,33	0,08	0,13			
Discours rapporté	>20%	1654	0,50	0,06	0,11			

Annexe 5.3 : Performances de EDS-NLP et MedSpaCy stratifié sur la fréquence d'apparition du concept au sein du corpus pour chaque modalité linguistique à l'échelle de l'entité

Modalité	Nombre d'occurrence du concept	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	1	3104	0,94	0,75	0,84	0,79	0,83	0,81
Negation	2	1458	0,93	0,70	0,80	0,80	0,85	0,83
Negation	3	762	0,85	0,67	0,75	0,77	0,85	0,81
Negation	4	596	0,83	0,40	0,54	0,86	0,69	0,77
Negation	5	405	0,84	0,48	0,61	0,68	0,74	0,71
Negation	6	228	1,00	0,48	0,65	0,81	0,54	0,65
Negation	7	189	0,78	0,47	0,58	0,56	0,45	0,50
Incertain	1	3104	0,72	0,25	0,37	0,46	0,20	0,28
Incertain	2	1458	0,72	0,20	0,32	0,53	0,18	0,27
Incertain	3	762	0,94	0,25	0,40	0,65	0,20	0,31
Incertain	4	596	0,67	0,20	0,31	0,17	0,07	0,10
Incertain	5	405	0,58	0,29	0,39	0,47	0,32	0,38
Incertain	6	228	0,40	0,15	0,22	0,40	0,13	0,20
Incertain	7	189	0,00	0,00	0,00	0,00	0,00	0,00
Historique	1	3104	0,12	0,69	0,20	0,09	0,80	0,16
Historique	2	1458	0,13	0,87	0,23	0,08	0,89	0,15
Historique	3	762	0,09	0,77	0,16	0,05	0,63	0,10
Historique	4	596	0,09	0,54	0,16	0,06	0,64	0,12
Historique	5	405	0,12	0,55	0,19	0,08	1,00	0,14
Historique	6	228	0,07	0,75	0,13	0,02	0,33	0,04
Historique	7	189	0,06	0,67	0,10	0,06	0,67	0,10
Conditionnel	1	3104	0,42	0,21	0,28	0,35	0,44	0,39
Conditionnel	2	1458	0,62	0,27	0,38	0,58	0,66	0,62
Conditionnel	3	762	0,56	0,36	0,44	0,49	0,65	0,56
Conditionnel	4	596	0,67	0,15	0,24	0,44	0,29	0,35
Conditionnel	5	405	0,77	0,26	0,39	0,46	0,46	0,46
Conditionnel	6	228	0,67	0,15	0,25	0,67	0,22	0,33
Conditionnel	7	189	0,17	0,07	0,10	0,33	0,20	0,25
Nonpatient	1	3104	0,65	0,49	0,56	0,59	0,56	0,58
Nonpatient	2	1458	0,63	0,53	0,58	0,53	0,51	0,52
Nonpatient	3	762	0,46	0,35	0,40	0,42	0,36	0,38
Nonpatient	4	596	0,65	0,42	0,51	0,61	0,64	0,62
Nonpatient	5	405	0,80	0,67	0,73	0,70	0,58	0,64
Nonpatient	6	228	1,00	0,88	0,93	0,57	0,44	0,50
Discours rapporté	1	3104	0,48	0,12	0,19			
Discours rapporté	2	1458	0,50	0,09	0,16			
Discours rapporté	3	762	0,11	0,02	0,04			
Discours rapporté	4	596	0,00	0,00	0,00			
Discours rapporté	5	405	1,00	0,04	0,08			

Annexe 5.4 : Performances de EDS-NLP et MedSpaCy stratifié sur le type sémantique pour chaque modalité linguistique à l'échelle du document

Modalité	Type sémantique	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	Acquired Abnormality	19	1	0,5	0,67	1	1	1
Negation	Anatomical Abnormality	69	1	0,89	0,94	1	1	1
Negation	Congenital Abnormality	32	1	0,75	0,86	0,33	1	0,5
Negation	Diagnostic Procedure	516	0,82	0,24	0,38	0,36	0,31	0,33
Negation	Disease or Syndrome	1283	0,94	0,76	0,84	0,77	0,85	0,81
Negation	Finding	568	0,94	0,78	0,85	0,76	0,87	0,81
Negation	Injury or Poisoning	120	0,92	0,79	0,85	0,92	0,85	0,88
Negation	Mental or Behavioral Dysfunction	117	1	0,71	0,83	0,6	0,67	0,63
Negation	Neoplastic Process	100	1	0,63	0,77	1	1	1
Negation	Pathologic Function	414	0,97	0,82	0,89	0,77	0,9	0,83
Negation	Pharmacologic Substance	92	1	0,42	0,59	1	0,83	0,91
Negation	Sign or Symptom	500	0,96	0,86	0,91	0,81	0,85	0,83
Negation	Therapeutic or Preventive Procedure	604	0,79	0,63	0,7	0,77	0,7	0,73
Incertain	Acquired Abnormality	19	1	1	1	1	1	1
Incertain	Anatomical Abnormality	69	1	0,33	0,5	0	0	0
Incertain	Congenital Abnormality	32	1	0,25	0,4	0	0	0
Incertain	Diagnostic Procedure	516	0,67	0,25	0,36	0,33	0,1	0,15
Incertain	Disease or Syndrome	1283	0,77	0,34	0,47	0,54	0,33	0,41
Incertain	Finding	568	0,67	0,08	0,14	0,33	0,06	0,1
Incertain	Gene or Genome	5	1	1	1	0	0	0
Incertain	Injury or Poisoning	120	0,4	0,4	0,4	0,4	0,4	0,4
Incertain	Mental or Behavioral Dysfunction	117	1	0,25	0,4	1	0,5	0,67
Incertain	Neoplastic Process	100	0,67	0,33	0,44	0,67	0,5	0,57
Incertain	Pathologic Function	414	0,67	0,25	0,36	0,44	0,36	0,4
Incertain	Sign or Symptom	500	0,75	0,11	0,19	0,25	0,04	0,07
Incertain	Therapeutic or Preventive Procedure	604	0,5	0,04	0,07	0	0	0
Historique	Acquired Abnormality	19	0,43	1	0,6	0,43	1	0,6
Historique	Anatomical Abnormality	69	0,19	0,83	0,3	0,11	1	0,2
Historique	Antibiotic	5	1	1	1	1	1	1
Historique	Congenital Abnormality	32	0	0	0	0,2	1	0,33
Historique	Diagnostic Procedure	516	0,02	0,17	0,03	0,08	0,71	0,15
Historique	Disease or Syndrome	1283	0,09	0,81	0,16	0,1	0,85	0,17
Historique	Finding	568	0,25	0,63	0,36	0,07	0,91	0,13
Historique	Gene or Genome	5	0	0	0	0	0	0
Historique	Injury or Poisoning	120	0,13	0,8	0,23	0,07	0,67	0,12
Historique	Mental or Behavioral Dysfunction	117	0,07	0,75	0,13	0,05	0,67	0,09
Historique	Neoplastic Process	100	0,19	0,89	0,31	0,1	1	0,17
Historique	Pathologic Function	414	0,09	0,69	0,16	0,06	0,75	0,11
Historique	Pharmacologic Substance	92	0	0	0	0	0	0
Historique	Sign or Symptom	500	0,03	0,67	0,05	0,01	0,2	0,03
Historique	Therapeutic or Preventive Procedure	604	0,1	0,67	0,18	0,06	0,75	0,12

Annexe 5.4 : Suite

Modalité	Type sémantique	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Conditionnel	Acquired Abnormality	19	0	0	0	0	0	0
Conditionnel	Congenital Abnormality	32	1	0,75	0,86	1	1	1
Conditionnel	Diagnostic Procedure	516	0,08	0,13	0,1	0	0	0
Conditionnel	Disease or Syndrome	1283	0,38	0,16	0,22	0,46	0,55	0,5
Conditionnel	Finding	568	0,67	0,24	0,35	0,67	0,6	0,63
Conditionnel	Mental or Behavioral Dysfunction	117	1	0,25	0,4	0	0	0
Conditionnel	Neoplastic Process	100	0	0	0	0	0	0
Conditionnel	Pathologic Function	414	0,5	0,25	0,33	0,17	0,4	0,24
Conditionnel	Pharmacologic Substance	92	0	0	0	1	0,2	0,33
Conditionnel	Sign or Symptom	500	0,73	0,68	0,7	0,65	0,81	0,72
Conditionnel	Therapeutic or Preventive Procedure	604	0,2	0,08	0,11	0	0	0
Nonpatient	Anatomical Abnormality	69	1	0,67	0,8	0	0	0
Nonpatient	Congenital Abnormality	32	0,25	1	0,4	0,75	1	0,86
Nonpatient	Diagnostic Procedure	516	0,75	0,33	0,46	0,75	0,6	0,67
Nonpatient	Disease or Syndrome	1283	0,67	0,52	0,58	0,72	0,57	0,63
Nonpatient	Finding	568	0,71	0,48	0,58	0,38	0,67	0,48
Nonpatient	Gene or Genome	5	0	0	0	1	1	1
Nonpatient	Injury or Poisoning	120	0,2	1	0,33	0,2	0,5	0,29
Nonpatient	Mental or Behavioral Dysfunction	117	0,67	0,5	0,57	0	0	0
Nonpatient	Neoplastic Process	100	0,8	0,74	0,77	0,8	0,87	0,83
Nonpatient	Pathologic Function	414	0,5	0,45	0,48	0,5	0,42	0,45
Nonpatient	Sign or Symptom	500	0	0	0	0	0	0
Nonpatient	Therapeutic or Preventive Procedure	604	0,22	0,2	0,21	0,44	0,27	0,33
Discours rapporté	Anatomical Abnormality	69	0	0	0			
Discours rapporté	Diagnostic Procedure	516	0	0	0			
Discours rapporté	Disease or Syndrome	1283	0,71	0,11	0,19			
Discours rapporté	Finding	568	0,63	0,18	0,28			
Discours rapporté	Injury or Poisoning	120	0	0	0			
Discours rapporté	Neoplastic Process	100	0	0	0			
Discours rapporté	Pathologic Function	414	0,25	0,08	0,12			
Discours rapporté	Pharmacologic Substance	92	0,5	0,08	0,13			
Discours rapporté	Sign or Symptom	500	0,6	0,28	0,38			
Discours rapporté	Therapeutic or Preventive Procedure	604	0	0	0			

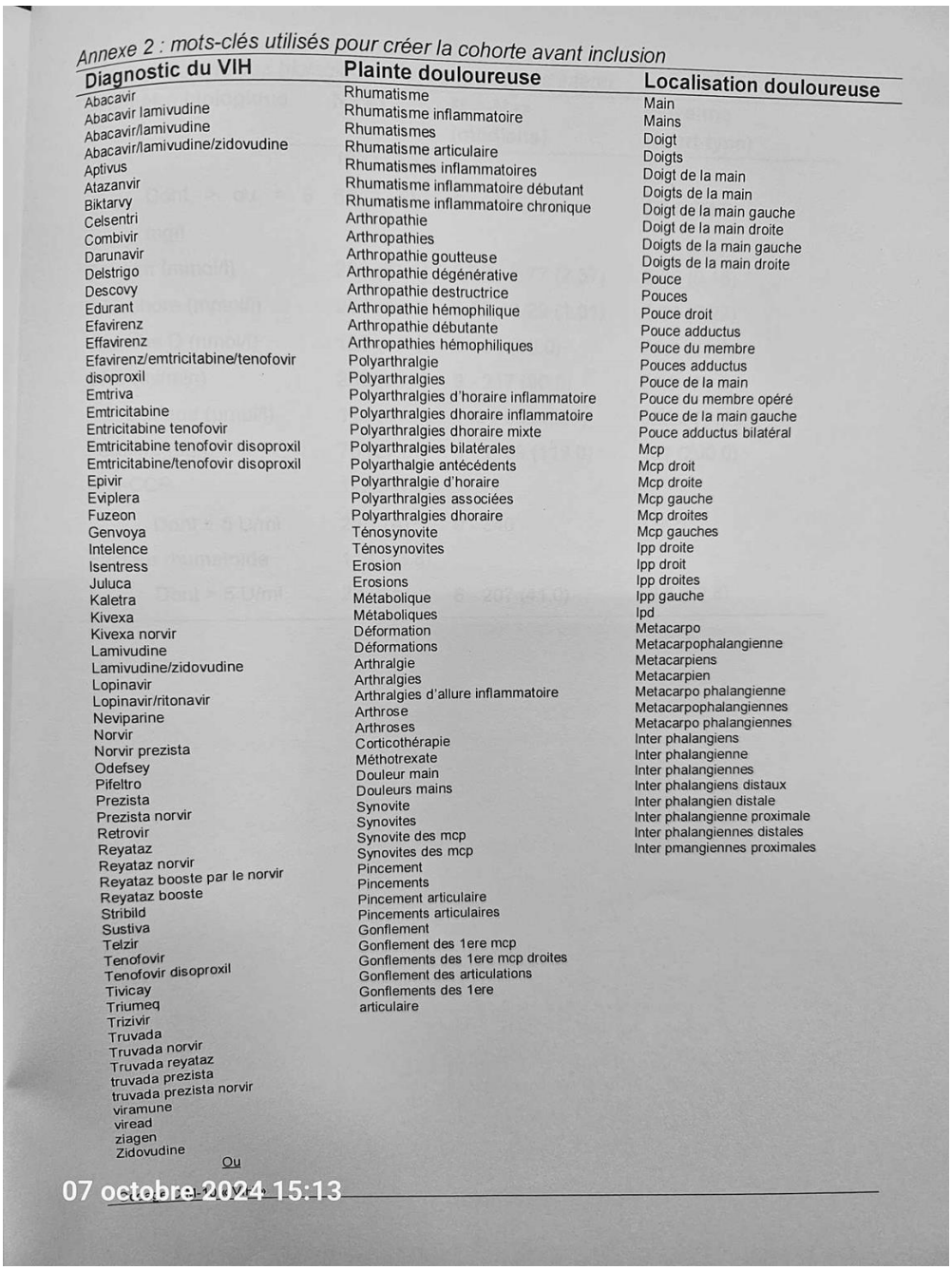
Annexe 5.5 : Performances de EDS-NLP et MedSpaCy stratifié sur la fréquence d'apparition du concept au sein du corpus pour chaque modalité linguistique à l'échelle du document

Modalité	Fréquence du concept au sein du corpus	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	[0%-1%]	599	0,97	0,77	0,86	0,75	0,86	0,80
Negation	]1%-5%]	1457	0,94	0,72	0,81	0,79	0,82	0,80
Negation	]5%-10%]	722	0,93	0,72	0,81	0,79	0,80	0,80
Negation	]10%-20%]	689	0,94	0,76	0,84	0,79	0,82	0,81
Negation	>20%	978	0,94	0,80	0,87	0,77	0,89	0,83
Incertain	[0%-1%]	599	0,82	0,30	0,44	0,36	0,17	0,23
Incertain	]1%-5%]	1457	0,73	0,24	0,36	0,59	0,21	0,31
Incertain	]5%-10%]	722	0,67	0,17	0,27	0,44	0,14	0,22
Incertain	]10%-20%]	689	0,54	0,18	0,27	0,46	0,22	0,30
Incertain	>20%	978	0,83	0,28	0,42	0,39	0,19	0,26
Historique	[0%-1%]	599	0,14	0,64	0,23	0,11	0,82	0,19
Historique	]1%-5%]	1457	0,09	0,56	0,15	0,08	0,80	0,15
Historique	]5%-10%]	722	0,11	0,84	0,20	0,06	0,71	0,12
Historique	]10%-20%]	689	0,14	0,85	0,25	0,09	0,85	0,16
Historique	>20%	978	0,11	0,76	0,19	0,07	0,85	0,12
Conditionnel	[0%-1%]	599	0,24	0,13	0,17	0,18	0,30	0,22
Conditionnel	]1%-5%]	1457	0,46	0,26	0,34	0,46	0,45	0,46
Conditionnel	]5%-10%]	722	0,52	0,33	0,41	0,35	0,50	0,41
Conditionnel	]10%-20%]	689	0,47	0,21	0,29	0,41	0,47	0,44
Conditionnel	>20%	978	0,48	0,19	0,27	0,52	0,65	0,58
Nonpatient	[0%-1%]	599	0,72	0,62	0,67	0,59	0,79	0,68
Nonpatient	]1%-5%]	1457	0,44	0,30	0,36	0,42	0,33	0,37
Nonpatient	]5%-10%]	722	0,64	0,62	0,63	0,64	0,62	0,63
Nonpatient	]10%-20%]	689	0,72	0,69	0,71	0,64	0,67	0,65
Nonpatient	>20%	978	0,69	0,47	0,56	0,56	0,64	0,60
Discours rapporté	[0%-1%]	599	0,44	0,14	0,21			
Discours rapporté	]1%-5%]	1457	0,73	0,11	0,19			
Discours rapporté	]5%-10%]	722	0,25	0,04	0,07			
Discours rapporté	]10%-20%]	689	0,55	0,21	0,30			
Discours rapporté	>20%	978	0,17	0,06	0,08			

Annexe 5.6 : Performances de EDS-NLP et MedSpaCy stratifié sur la fréquence d'apparition du concept au sein du corpus pour chaque modalité linguistique à l'échelle du document

Modalité	Nombre d'occurrence du concept	N total	EDS-NLP			MedSpaCy		
			Rappel	Précision	F1	Rappel	Précision	F1
Negation	1	3104	0,94	0,75	0,84	0,79	0,83	0,81
Negation	2	729	0,93	0,76	0,83	0,74	0,91	0,82
Negation	3	254	0,92	0,92	0,92	0,50	1,00	0,67
Incertain	1	3104	0,72	0,25	0,37	0,46	0,20	0,28
Incertain	2	729	0,75	0,13	0,22	0,50	0,11	0,18
Incertain	3	254	1,00	0,25	0,40	1,00	0,20	0,33
Historique	1	3104	0,12	0,69	0,20	0,09	0,80	0,16
Historique	2	729	0,10	0,93	0,18	0,04	1,00	0,07
Historique	3	254	0,03	1,00	0,06	0,00	0,00	0,00
Historique	4	149	0,09	0,50	0,15	0,00	0,00	0,00
Conditionnel	1	3104	0,42	0,21	0,28	0,35	0,44	0,39
Conditionnel	2	729	0,54	0,30	0,39	0,62	0,73	0,67
Conditionnel	3	254	0,60	0,75	0,67	0,80	1,00	0,89
Nonpatient	1	3104	0,65	0,49	0,56	0,59	0,56	0,58
Nonpatient	2	729	0,58	0,61	0,59	0,42	0,62	0,50
Nonpatient	3	254	0,33	0,50	0,40	0,33	1,00	0,50
Nonpatient	4	149	0,50	0,50	0,50	0,50	1,00	0,67
Discours rapporté	1	3104	0,48	0,12	0,19			
Discours rapporté	2	729	0,00	0,00	0,00			

Annexe 6. Mots-clés utilisés pour la cohorte ArthroVIH



Annexe 2 : mots-clés utilisés pour créer la cohorte avant inclusion

Diagnostic du VIH	Plainte douloureuse	Localisation douloureuse
Abacavir	Rhumatisme	Main
Abacavir lamivudine	Rhumatisme inflammatoire	Mains
Abacavir/lamivudine	Rhumatismes	Doigt
Abacavir/lamivudine/zidovudine	Rhumatisme articulaire	Doigts
Aptivus	Rhumatismes inflammatoires	Doigt de la main
Atazanvir	Rhumatisme inflammatoire débutant	Doigts de la main
Biktarvy	Rhumatisme inflammatoire chronique	Doigt de la main gauche
Celsentri	Arthropathie	Doigt de la main droite
Combivir	Arthropathies	Doigts de la main gauche
Darunavir	Arthropathie goutteuse	Doigts de la main droite
Delstrigo	Arthropathie dégénérative	Pouce
Descovy	Arthropathie destructrice	Pouces
Edurant	Arthropathie hémophilique	Pouce droit
Efavirenz	Arthropathie débutante	Pouce adductus
Effavirenz	Arthropathies hémophiliques	Pouce du membre
Efavirenz/emtricitabine/tenofovir disoproxil	Polyarthralgie	Pouces adductus
Emtriva	Polyarthralgies	Pouce de la main
Emtricitabine	Polyarthralgies d'horaire inflammatoire	Pouce du membre opéré
Entricitabine tenofovir	Polyarthralgies dhoraire inflammatoire	Pouce de la main gauche
Emtricitabine tenofovir disoproxil	Polyarthralgies dhoraire mixte	Pouce adductus bilatéral
Emtricitabine/tenofovir disoproxil	Polyarthralgies bilatérales	Mcp
Epivir	Polyarthalgie antécédents	Mcp droit
Eviplera	Polyarthralgie d'horaire	Mcp droite
Fuzeon	Polyarthralgies associées	Mcp gauche
Genvoya	Polyarthralgies dhoraire	Mcp droites
Intelence	Ténosynovite	Mcp gauches
Isentress	Ténosynovites	Ipp droite
Juluca	Erosion	Ipp droit
Kaletra	Erosions	Ipp droites
Kivexa	Métabolique	Ipp gauche
Kivexa norvir	Métaboliques	Ipd
Lamivudine	Déformation	Metacarpo
Lamivudine/zidovudine	Déformations	Metacarpophalangienne
Lopinavir	Arthralgie	Metacarpiens
Lopinavir/ritonavir	Arthralgies	Metacarpien
Neviparine	Arthralgies d'allure inflammatoire	Metacarpo phalangienne
Norvir	Arthrose	Metacarpophalangiennes
Norvir prezista	Arthroses	Metacarpo phalangiennes
Odefsey	Corticothérapie	Inter phalangiens
Pifeltro	Méthotrexate	Inter phalangienne
Prezista	Douleur main	Inter phalangiennes
Prezista norvir	Douleurs mains	Inter phalangiens distaux
Retrovir	Synovite	Inter phalangien distale
Reyataz	Synovites	Inter phalangienne proximale
Reyataz norvir	Synovite des mcp	Inter phalangiennes distales
Reyataz booste par le norvir	Synovites des mcp	Inter pmangiennes proximales
Reyataz booste	Pincement	
Stribild	Pincements	
Sustiva	Pincement articulaire	
Telzir	Pincements articulaires	
Tenofovir	Gonflement	
Tenofovir disoproxil	Gonflement des 1ere mcp	
Tivicay	Gonflements des 1ere mcp droites	
Triumeq	Gonflement des articulations	
Trizivir	Gonflements des 1ere articulaire	
Truvada		
Truvada norvir		
Truvada reyataz		
truvada prezista		
truvada prezista norvir		
viramune		
viread		
ziagen		
Zidovudine		

Ou

Octobre 2024

07 octobre 2024 15:13

Annexe 7. Evaluation des capacités de filtrage suivant différentes stratégies avec EDS-NLP et MedSpaCy pour la cohorte ArthroVIH

Filtrage avec EDS-NLP	Sans filtrage		Négation			Hypothèse			Non Patient			Négation + Hypothèse			Négation + Non Patient			Négation + Hypothèse + Non Patient		
	N	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%
Patients avec les 3 critères d'inclusion	1949	1854	-95	-5%	1923	-26	-1%	1947	-2	0%	1820	-129	-7%	1852	-97	-5%	1818	-131	-7%
Séjours avec les 3 critères d'inclusion	3154	2734	-420	-13%	3042	-112	-4%	3146	-8	0%	2630	-524	-17%	2726	-428	-14%	2622	-532	-17%
Documents avec les 3 critères d'inclusion	2776	2374	-402	-14%	2689	-87	-3%	2768	-8	0%	2289	-487	-18%	2367	-409	-15%	2282	-494	-18%
Patients avec ≥ 1 concept	2316	2314	-2	0%	2315	-1	0%	2316	0	0%	2313	-3	0%	2314	-2	0%	2313	-3	0%
Avec concept de diagnostic VIH	2225	2209	-16	-1%	2216	-9	0%	2222	-3	0%	2200	-25	-1%	2206	-19	-1%	2197	-28	-1%
Avec concept de plainte douloureuse	2098	2024	-74	-4%	2089	-9	0%	2097	-1	0%	2005	-93	-4%	2023	-75	-4%	2004	-94	-4%
Avec concept de localisation douloureuse	2204	2184	-20	-1%	2193	-11	0%	2204	0	0%	2172	-32	-1%	2184	-20	-1%	2172	-32	-1%
Séjours avec ≥ 1 concept	68132	66449	-1683	-2%	67496	-636	-1%	68100	-32	0%	65931	-2201	-3%	66417	-1715	-3%	65900	-2232	-3%
Avec concept de diagnostic VIH	53310	52364	-946	-2%	52899	-411	-1%	53284	-26	0%	52028	-1282	-2%	52338	-972	-2%	52003	-1307	-2%
Avec concept de plainte douloureuse	23183	20396	-2787	-12%	22539	-644	-3%	23142	-41	0%	19806	-3377	-15%	20362	-2821	-12%	19772	-3411	-15%
Avec concept de localisation douloureuse	13798	13265	-533	-4%	13593	-205	-1%	13769	-29	0%	13077	-721	-5%	13238	-560	-4%	13050	-748	-5%
Document avec ≥ 1 concept	112036	108587	-3449	-3%	110608	-1428	-1%	111974	-62	0%	107402	-4634	-4%	108533	-3503	-3%	107349	-4687	-4%
Avec concept de diagnostic VIH	79851	78091	-1760	-2%	79016	-835	-1%	79802	-49	0%	77408	-2443	-3%	78049	-1802	-2%	77367	-2484	-3%
Avec concept de plainte douloureuse	36110	32258	-3852	-11%	35016	-1094	-3%	36061	-49	0%	31207	-4903	-14%	32217	-3893	-11%	31167	-4943	-14%
Avec concept de localisation douloureuse	19747	18979	-768	-4%	19420	-327	-2%	19712	-35	0%	18700	-1047	-5%	18947	-800	-4%	18668	-1079	-5%
Total entités retrouvés	416139	392194	-23945	-6%	403969	-12170	-3%	415632	-507	0%	381992	-34147	-8%	391805	-24334	-6%	381617	-34522	-8%
Diagnostic VIH	302591	292158	-10433	-3%	295124	-7467	-2%	302318	-273	0%	285843	-16748	-6%	291931	-10660	-4%	285628	-16963	-6%
Plainte douloureuse	74578	63170	-11408	-15%	70810	-3768	-5%	74419	-159	0%	60113	-14465	-19%	63080	-11498	-15%	60024	-14554	-20%
Localisation douloureuse	38970	36866	-2104	-5%	38035	-935	-2%	38895	-75	0%	36036	-2934	-8%	36794	-2176	-6%	35965	-3005	-8%

Filtrage avec MedSpaCy	Sans filtrage		Négation			Hypothèse			Non Patient			Négation + Hypothèse			Négation + Non Patient			Négation + Hypothèse + Non Patient		
	N	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%	N	Dif	%
Patients avec les 3 critères d'inclusion	1949	1893	-56	-3%	1921	-28	-1%	1947	-2	0%	1865	-84	-4%	1892	-57	-3%	1864	-85	-4%
Séjours avec les 3 critères d'inclusion	3154	2899	-255	-8%	3070	-84	-3%	3146	-8	0%	2818	-336	-11%	2891	-263	-8%	2810	-344	-11%
Documents avec les 3 critères d'inclusion	2776	2525	-251	-9%	2706	-70	-3%	2766	-10	0%	2455	-321	-12%	2516	-260	-9%	2446	-330	-12%
Patients avec ≥ 1 concept	2316	2315	-1	0%	2315	-1	0%	2316	0	0%	2314	-2	0%	2315	-1	0%	2314	-2	0%
Avec concept de diagnostic VIH	2225	2219	-6	0%	2204	-21	-1%	2222	-3	0%	2200	-25	-1%	2216	-9	0%	2198	-27	-1%
Avec concept de plainte douloureuse	2098	2044	-54	-3%	2092	-6	0%	2097	-1	0%	2035	-63	-3%	2043	-55	-3%	2034	-64	-3%
Avec concept de localisation douloureuse	2204	2196	-8	0%	2197	-7	0%	2204	0	0%	2189	-15	-1%	2196	-8	0%	2189	-15	-1%
Séjours avec ≥ 1 concept	68132	67474	-658	-1%	67654	-478	-1%	68101	-31	0%	66990	-1142	-2%	67442	-690	-1%	66959	-1173	-2%
Avec concept de diagnostic VIH	53310	53203	-107	0%	52899	-411	-1%	53285	-25	0%	52797	-513	-1%	53176	-134	0%	52771	-539	-1%
Avec concept de plainte douloureuse	23183	21163	-2020	-9%	22751	-432	-2%	23126	-57	0%	20725	-2458	-11%	21116	-2067	-9%	20678	-2505	-11%
Avec concept de localisation douloureuse	13798	13583	-215	-2%	13685	-113	-1%	13789	-9	0%	13472	-326	-2%	13574	-224	-2%	13464	-334	-2%
Document avec ≥ 1 concept	112036	110731	-1305	-1%	110926	-1110	-1%	111976	-60	0%	109614	-2422	-2%	110670	-1366	-1%	109557	-2479	-2%
Avec concept de diagnostic VIH	79851	79637	-214	0%	79032	-819	-1%	79801	-50	0%	78810	-1041	-1%	79586	-265	0%	78762	-1089	-1%
Avec concept de plainte douloureuse	36110	33430	-2680	-7%	35391	-719	-2%	36047	-63	0%	32703	-3407	-9%	33378	-2732	-8%	32650	-3460	-10%
Avec concept de localisation douloureuse	19747	19434	-313	-2%	19576	-171	-1%	19736	-11	0%	19264	-483	-2%	19423	-324	-2%	19254	-493	-2%
Total entités retrouvés	416139	405408	-10731	-3%	408181	-7958	-2%	415534	-605	0%	397672	-18467	-4%	404854	-11285	-3%	397135	-19004	-5%
Diagnostic VIH	302591	300576	-2015	-1%	297657	-4934	-2%	302267	-324	0%	295751	-6840	-2%	300256	-2335	-1%	295441	-7150	-2%
Plainte douloureuse	74578	66678	-7900	-11%	72081	-2497	-3%	74332	-246	0%	64279	-10299	-14%	66479	-8099	-11%	64083	-10495	-14%
Localisation douloureuse	38970	38154	-816	-2%	38443	-527	-1%	38935	-35	0%	37642	-1328	-3%	38119	-851	-2%	37611	-1359	-3%

Annexe 8. Description de l'atteinte des patients filtré selon les différentes stratégies de filtrage avec EDS-NLP et MedSpaCy

Stratégie de Filtrage EDS-NLP*	Atteinte Articulaire	Atteinte Neurologique	Atteinte Orthopédique	Atteinte Infectieuse	Atteinte Dermatologique	Atteinte Doigt à Ressaut
Négation	2	6	4	4	2	0
Hypothèse	0	1	0	1	2	0
Non patient	1	0	0	0	0	0
Négation + Hypothèse	2	8	4	5	4	2
Négation + Non patient	3	6	4	4	2	0
Négation + Hypothèse + Non patient	3	8	4	5	4	2

** Il n'avait été retrouvé aucune erreur de filtrage chez les patients avec une atteinte vasculaire.*

Stratégie de Filtrage MedSpaCy*	Atteinte Articulaire	Atteinte Neurologique	Atteinte Orthopédique	Atteinte Infectieuse	Atteinte Dermatologique
Négation	1	3	4	2	2
Hypothèse	0	1	1	0	0
Non patient	0	0	1	0	0
Négation + Hypothèse	1	5	5	2	2
Négation + Non patient	1	3	4	2	2
Négation + Hypothèse + Non patient	1	5	5	2	2

** Il n'avait été retrouvé aucune erreur de filtrage chez les patients avec une atteinte vasculaire ou un Doigt à Ressaut.*

Annexe 9 : Impact de la filtration

Annexe 9.1 : EDS-NLP

Type sémantique	Sans filtrage		Négation			Hypothèse			Non Patient			Négation + Hypothèse			Négation + Non Patient			Négation + Hypothèse + Non Patient		
	Médiane	[25%-75%]	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif
Total (Distribution globale)*	94	[28-199]	82	[25-178]	-12	[3-221]	89	[27-189]	-5	[-1-10]	93	[28-197]	-1	[-2]	78	[23-167]	-16	[-5-32]	82	[24-174]	-12	[-4-25]	77	[23-166]	-17	[-5-33]
Disease or Syndrome	16	[5-40]	13	[4-33]	-3	[-1-7]	15	[4-37]	-1	[-1-3]	16	[5-39]	[-1]		12	[3-30]	-4	[-2--10]	13	[4-32]	-3	[-1-8]	12	[3-29]	-4	[-2--11]
Therapeutic or Preventive Procedure	14	[4-32]	13	[3-30]	-1	[-1--2]	14	[4-32]			14	[4-32]			13	[3-30]	-1	[-1--2]	13	[3-30]	-1	[-1--2]	13	[3-29]	-1	[-1--3]
Finding	10	[3-24]	9	[2-21]	-1	[-1--3]	10	[2-23]		[-1--1]	10	[3-24]			9	[2-21]	-1	[-1--3]	9	[2-21]	-1	[-1--3]	8	[2-20]	-2	[-1--4]
Diagnostic Procedure	11	[4-22]	10	[3-21]	-1	[-1--1]	11	[4-21]		[-1]	11	[4-22]			10	[3-21]	-1	[-1--1]	10	[3-21]	-1	[-1--1]	10	[3-20]	-1	[-1--2]
Sign or Symptom	7	[2-15]	5	[1-12]	-2	[-1--3]	6	[1-15]	-1	[-1]	7	[2-15]			5	[1-12]	-2	[-1--3]	5	[1-12]	-2	[-1--3]	5	[1-12]	-2	[-1--3]
Pathologic Function	7	[2-15]	5	[1-11]	-2	[-1--4]	6	[2-14]	-1	[-1]	7	[2-15]			5	[1-10]	-2	[-1--5]	5	[1-11]	-2	[-1--4]	4	[1-10]	-3	[-1--5]
Pharmacologic Substance	4	[1-8]	3	[1-7]	-1	[-2]	3	[1-7]	-1	[-2]	4	[1-8]	[-1]		3	[1-7]	-1	[-2]	3	[1-7]	-1	[-2]	3	[1-7]	-1	[-2]
Injury or Poisoning	3	[1-6]	2	[0-5]	-1	[-1--1]	2	[1-5]	-1	[-1]	3	[1-6]			2	[0-4]	-1	[-1--2]	2	[0-5]	-1	[-1--1]	2	[0-4]	-1	[-1--2]
Mental or Behavioral Dysfunction	1	[0-4]	1	[0-3]		[-1]	1	[0-4]		[]	1	[0-4]			1	[0-3]		[-1]	1	[0-3]		[-1]	1	[0-3]		[-1]
Neoplastic Process	1	[0-3]	0	[0-2]	-1	[-2]	0	[0-3]	-1	[-1]	1	[0-3]		[-1]	0	[0-2]	-1	[-2]	0	[0-2]	-1	[-2]		[0-2]	-1	[-2]
Anatomical Abnormality	1	[0-2]	1	[0-2]		[-2]	1	[0-2]		[-2]	1	[0-2]		[-2]	1	[0-2]		[-2]	1	[0-2]		[-2]	1	[0-2]		[-2]
Congenital Abnormality	1	[0-1]	1	[0-1]		[-3]	1	[0-1]		[-3]	1	[0-1]		[-3]	1	[0-1]		[-3]	1	[0-1]		[-3]	1	[0-1]		[-3]

Annexe 9.1 : MedSpaCy

Type sémantique	Sans filtrage		Négation		Hypothèse			Non Patient			Négation + Hypothèse			Négation + Non Patient			Négation + Hypothèse + Non Patient		
	Médiane	[25%-75%]	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif	Médiane	[25%-75%]	Dif
Total (Distribution globale)*	94	[28-199]	86	[26-183]	-8	[-2;-16]	92	[27-194]	-2	[-1;-5]	93	[28-197]	-1	[-2]	84	[25-178]	-10	[-3;-21]	85	[26-182]	-7	[-2;-17]	83	[25-178]	-11	[-3;-21]
Disease or Syndrome	16	[5-40]	14	[4-35]	-2	[-1;-5]	15	[5-39]	-1	[-;-1]	16	[5-40]			13	[4-34]	-3	[-1;-6]	14	[4-35]	-2	[-1;-5]	13	[4-34]	-3	[-1;-6]
Therapeutic or Preventive Procedure	14	[4-32]	13	[3-31]	-1	[-1;-1]	14	[4-32]			14	[4-32]			13	[3-30]	-1	[-1;-2]	13	[3-31]	-1	[-1;-1]	13	[3-30]	-1	[-1;-2]
Finding	10	[3-24]	9	[2-22]	-1	[-1;-2]	10	[3-24]			10	[3-24]			9	[2-22]	-1	[-1;-2]	9	[2-22]	-1	[-1;-2]	9	[2-22]	-1	[-1;-2]
Diagnostic Procedure	11	[4-22]	11	[3-21]	-1	[-1;-1]	11	[4-21]	[-;-1]		11	[4-22]			11	[3-21]		[-1;-1]	11	[3-21]		[-1;-1]	11	[3-21]		[-1;-1]
Sign or Symptom	7	[2-15]	6	[1-13]	-1	[-1;-2]	6	[1-15]	-1	[-1;-]	7	[2-15]			5	[1-13]	-2	[-1;-2]	5	[1-13]	-2	[-1;-2]	5	[1-12]	-2	[-1;-3]
Pathologic Function	7	[2-15]	6	[2-12]	-1	[-;-3]	7	[2-14]	[-;-1]		7	[2-15]			5	[2-12]	-2	[-;-3]	5	[2-12]	-2	[-;-3]	5	[2-12]	-2	[-;-3]
Pharmacologic Substance	4	[1-8]	4	[1-7]		[-;2]	3	[1-7]	-1	[-;2]	4	[1-8]	[-;-1]		3	[1-7]	-1	[-;2]	4	[1-7]		[-;2]	3	[1-7]	-1	[-;2]
Injury or Poisoning	3	[1-6]	2	[0-5]	-1	[-1;-1]	3	[1-6]			3	[1-6]			2	[0-5]	-1	[-1;-1]	2	[0-5]	-1	[-1;-1]	2	[0-5]	-1	[-1;-1]
Mental or Behavioral Dysfunction	1	[0-4]	1	[0-3]		[-;1]	1	[-4]			1	[0-4]			1	[0-3]		[-;1]	1	[0-3]		[-;1]	1	[0-3]		[-;1]
Neoplastic Process	1	[0-3]	0	[0-3]	-1	[-;1]	0	[-3]	-1	[-;-1]	1	[0-3]	[-;-1]		0	[0-2]	-1	[-;2]	0	[0-2]	-1	[-;2]	0	[0-2]	-1	[-;2]
Anatomical Abnormality	1	[0-2]	1	[0-2]		[-;2]	1	[-2]		[-;2]	1	[0-2]	[-;-1]		1	[0-2]		[-;2]	1	[0-2]		[-;2]	1	[0-2]		[-;2]
Congenital Abnormality	1	[0-1]	1	[0-1]		[-;3]	1	[-1]		[-;3]	1	[0-1]	[-;-3]		1	[0-1]		[-;3]	1	[0-1]		[-;3]	1	[0-1]		[-;3]

Titre : Identification des patients dans un Entrepôt de Données de Santé Hospitalier : impact de la détection de contexte sur l'extraction des concepts médicaux

Résumé : Les entrepôts de données de santé hospitaliers (EDSH) permettent une exploitation secondaire des données médicales à des fins de recherche clinique et épidémiologique. Cependant, l'analyse des données non structurées demeure une tâche complexe, nécessitant l'application de techniques avancées de traitement automatique du langage naturel (TAL), notamment pour la détection de contexte. Cette thèse évalue l'impact de la prise en compte du contexte sur l'identification des concepts médicaux au sein de l'EDSH du CHU de Bordeaux. Deux bibliothèques, EDS-NLP et MedSpaCy, ont été comparées quant à leur capacité à détecter la négation, l'hypothèse, ainsi que les notions d'historique et de non-patient. L'évaluation des performances révèle que EDS-NLP présente un meilleur rappel, tandis que MedSpaCy affiche une meilleure précision, bien que les performances globales restent limitées. Une validation en conditions réelles a été réalisée sur la cohorte ArthroVIH, montrant une réduction du bruit dans l'identification des patients de 3 à 7 %, avec une précision avoisinant 80 %. Toutefois, ces résultats ne permettent pas encore d'automatiser intégralement la sélection des patients dans un EDSH pour une étude clinique. Une utilisation prudente des algorithmes à base de règles est recommandée.

Mots clés : Entrepôt de données de santé hospitalier, traitement du langage naturel, détection de contexte, identification de patient, utilisation secondaire des données

Title : Patient identification in a Hospital Health Data Warehouse: impact of context detection on medical concept extraction

Abstract : Hospital health data warehouses (EDSH) enable the secondary use of medical data for clinical and epidemiological research. However, analyzing unstructured data remains a complex task, requiring advanced natural language processing (NLP) techniques, particularly for context detection. This thesis assesses the impact of context detection on the identification of medical concepts within the EDSH of Bordeaux University Hospital. Two libraries, EDS-NLP and MedSpaCy, were compared for their ability to detect negation, hypothesis, as well as historical and non-patient contexts. Performance evaluation shows that EDS-NLP achieves higher recall, while MedSpaCy provides greater precision, though overall performance remains modest. A real-world validation was conducted on the ArthroVIH cohort, demonstrating a 3 to 7% reduction in noise in patient identification, with an accuracy of approximately 80%. However, these results do not yet allow for the full automation of patient selection in an EDSH for clinical research. It is recommended to adopt a cautious approach when using rule-based algorithms.

Keywords : Hospital health data warehouse, natural language processing, context detection, phenotyping, secondary use of data