



Institut National des Langues et Civilisations Orientales

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours : *Ingénierie Multilingue*

par Elvira QUESADA MARAÑÓN

# Prédiction d'une maladie rare : l'amyloïdose cardiaque

Directeur de mémoire : Cyril Grouin

Année universitaire 2018/2019

# Table des matières

# Table des figures

# Section 1

# Introduction

## 1.1 Contexte de l'étude

Dans les années 50, un mathématicien et cryptologue britannique, Alan Turing, a publié l'article *Computing Machinery and Intelligence*, qui introduisait le test de Turing. Le but de ce test était de montrer si un ordinateur se comporte ou non comme un être humain. Ceci consistait à laisser une personne discuter à l'aveugle avec un autre humain et un ordinateur, et cette personne devait déterminer lequel de ses interlocuteurs était une machine. L'année 1950 est donc considérée comme l'année fondatrice de l'Intelligence Artificielle.

Cependant, la première fois que le terme « *artificial intelligence* » a été utilisé officiellement était lors d'un projet de recherche proposé par John McCarthy, professeur de mathématiques à l'Université Darmouth et créateur du LISP, en collaboration avec d'autres collègues comme Marvin Minsky (Université de Harvard), Nathaniel Rochester (IBM) et Claude Shannon (Bell Telephone Laboratories), qui ont proposé à la Fondation Rockefeller de mener à bien un projet dont le but était de « *...to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it* » (Minsky 1958). Cet atelier, qui a eu lieu pendant l'été de 1956, est généralement reconnu comme la date de naissance de ce nouveau domaine.

Plusieurs activités mentales humaines comme le fait d'écrire des pro-

grammes d'ordinateur, faire des mathématiques, raisonner, comprendre le langage ou même conduire une voiture sont considérées comme des tâches qui demandent d'intelligence. Dans les dernières décennies, plusieurs systèmes informatiques capables de réaliser ces tâches ont été construits. En particulier, des systèmes qui peuvent diagnostiquer des maladies, programmer la synthèse de composants chimiques complexes, analyser des circuits électroniques, comprendre le langage humain ou écrire des petits programmes. On pourrait dire que ces systèmes possèdent dans une plus ou moins grande mesure d'un degré d'intelligence artificielle (Nilsson 1980).

Actuellement, l'Intelligence Artificielle (IA) est un domaine de recherche en plein essor. Ses applications concernent tous les secteurs des activités humaines, notamment celui de la médecine, domaine dans lequel se focalise notre projet. Grâce à l'IA, la réalisation de certaines tâches comme les opérations assistées, le suivi de patients, les prothèses intelligentes ou les traitements personnalisées sont devenues désormais possibles. Plusieurs approches et méthodes comme le traitement automatique du langage, la construction d'ontologies, la fouille de données ou l'apprentissage automatique ont permis à ces avancées (Charlet et al. 2018).

De même, d'autres domaines ont fait son apparition à partir de ces avances technologiques. Par exemple, la médecine prédictive essaie d'effectuer des prédictions de maladies et l'évolution de celles-ci. Par ailleurs, la médecine de prédiction se focalise sur les recommandations de traitements personnalisés. L'IA contribue aussi à l'aide à la prise de décisions par rapport au diagnostic, ils existent des robots compagnons, en particulier pour les personnes âgées ou fragiles, ainsi qu'au développement de la chirurgie par ordinateur ou la prévention de maladies, comme l'anticipation d'une épidémie ou la pharmacovigilance (Charlet et al. 2018).

Il n'y a pas longtemps, pendant les années 1950, l'intelligence artificielle est née avec l'objectif de reproduire certaines tâches humaines par les machines, dont le but était d'imiter notre cerveau. L'approche la plus ancienne de l'IA s'appuie sur le fait que nous raisonnons en appliquant des règles logiques,

comme la déduction, la classification ou la hiérarchisation. Les systèmes conçus sur ce principe appliquent différentes méthodes, comme l'élaboration de modèles d'interaction entre automates, de modèles syntaxiques ou linguistiques (traitement automatique du langage) ou la construction d'ontologies (représentation de connaissances) (Charlet et al. 2018).

Plus tard, dans les années 1980, cette approche connue comme « symbolique », a permis de développer des outils capables de reproduire les mécanismes cognitifs d'un expert, d'où leur qualificatif de « systèmes experts ». Les exemples les plus significatifs sont ceux de Mycin (identification d'infections bactériennes) ou Sphinx (détection d'ictères), qui s'appuient sur un ensemble de connaissances médicales dans un domaine spécifique et une formalisation des raisonnements des spécialistes qui lient toutes leurs connaissances pour aboutir à un diagnostic (Charlet et al. 2018).

Cependant, les systèmes actuels, connus comme des systèmes d'aide à la décision, de gestion de connaissances ou d'e-santé, sont plus élaborés et sophistiqués, notamment parce qu'ils bénéficient de meilleurs modèles de raisonnement et de meilleures techniques pour décrire les connaissances médicales, des patients, et des activités médicales (Charlet et al. 2018). Pourtant, cette mécanique algorithmique n'a pour but que de soutenir ou épauler les médecins, pas de les remplacer.

Prenons le cas du projet européen Desiree, auquel participent des équipes du Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé (LIMICS, unité Inserm 1142) et de l'Assistance Publique - Hôpitaux de Paris (Charlet et al. 2018). Ce projet s'appuie sur cette approche symbolique pour aider les médecins dans le traitement et le suivi de patientes atteintes de cancer du sein. La plate-forme Desiree comprend des recommandations de bonnes pratiques fondées sur une ontologie. Le système peut aussi apprendre à partir de cas résolus ou de raisonnements par expérience. La base est enrichie continûment, donc le système peut ainsi évoluer.

Tandis que l'approche symbolique s'appuie sur les connaissances, l'approche dite numérique raisonne sur les données. Les systèmes cherchent dans les don-

nées disponibles pour en extraire des connaissances. Née à partir du connexionisme et des réseaux de neurones pendant les années 1980, elle se développe encore aujourd'hui grâce à l'augmentation de puissance des ordinateurs et l'énorme quantité de données que l'on possède, ce que l'on connaît comme *big data* (CHARLET et al. 2018).

La plupart de ces systèmes s'appliquent par apprentissage automatique et les algorithmes de *deep learning*, qui s'inspirent du fonctionnement cérébral et apprennent de tâches par « essais et erreurs » avant de gagner en autonomie.

Nous pouvons alors grouper les systèmes d'IA dans deux catégories. La première inclut les techniques de *machine learning* qui analysent les données structurées. La deuxième catégorie inclut les méthodes que l'on connaît comme *natural language processing* (NLP), qui consistent à extraire des informations à partir de données non structurées et qui servent ensuite à enrichir les données structurées. L'objectif principal de ces traitements est de transformer les textes en données structurées qui soient lisibles par une machine, et qui puissent être traitées plus tard par les techniques de *machine learning*.

En conclusion, l'intelligence artificielle est en pleine expansion et de nombreux chercheurs explorent différentes voies pour améliorer la performance de ces systèmes, ainsi que leur adéquation aux pratiques médicales visées. Ces voies de recherche portent en particulier sur le traitement de données, très hétérogènes, leur structuration et leur anonymisation. Le travail décrit dans ce mémoire porte sur ces différentes tâches.

## 1.2 Présentation du projet

Ce travail naît du projet de recherche réalisé par Sopra Steria et l'entreprise pharmaceutique Pfizer, en collaboration avec la Fundación San Juan de Dios et l'hôpital San Juan de Dios de León. Le projet vise à améliorer la prédiction du diagnostic de l'amyloïdose cardiaque, une maladie rare, à partir des informations médicales extraites des bases de données de ce hôpital.

Étant donné la nature du projet en ce qui concerne le caractère privée des

données, il y a eu énormément des difficultés pour accéder aux données dû aux différents contrats à signer par les différentes entités, de façon à respecter le règlement général sur la protection de données. Du fait de ce retard, on a dû créer nous mêmes notre propre corpus afin de mener à bien toutes les expériences possibles pour accomplir des tâches que nous rapprochent à la prédiction de cette maladie. Il est vrai que les données des hôpitaux seraient plus exhaustifs et complets du fait qu'il correspondent aux rapports cliniques et résultats de différents tests réalises aux patients lors de leur hospitalisation, mais nous avons dû changer notre conception du travail afin de nous adapter aux circonstances. Pour cette raison, on a finalement décidé de continuer avec l'objectif de prédire l'amyloïdose cardiaque mais avec des données différentes. Étant donné qu'il s'agit d'une maladie rare, nous avons eu des difficultés pour construire un corpus qui permette de prédire cette maladie. Finalement, nous avons décidé d'utiliser des forums comme ressources où l'on pourrait non seulement trouver des patients atteints d'amyloïdose qui parlent de la maladie et leurs symptômes mais aussi des patients atteints d'autres maladies cardiaques rares. De cette façon, nous pourrions entraîner un modèle qui prédite si un patient est atteint d'amyloïdose ou non. Bien que le but et le concept du projet ont été modifiés légèrement, l'objectif est toujours de prédire le diagnostic de l'amyloïdose.

### 1.2.1 Les amyloses

Ce projet porte sur la prédiction d'une maladie rare, l'amyloïdose cardiaque, qui appartient à une famille de maladies connues comme amyloses. Les amyloses sont des pathologies rares liées aux dépôts dans différents organes de substance amyloïde qui prend la forme de bâtonnets rigides (D'HÉMATOLOGIE, IFRAH et MAYNADIÉ 2018).

Tout au long de notre vie, les cellules se renouvellent grâce à leur ADN, qui « code » pour la fabrication de molécules appelées protéines, indispensables au bon fonctionnement de notre corps puisqu'elles fournissent la structure et la fonction de presque tous les processus biologiques. Une fois les protéines

fabriquées, elles adoptent une forme particulière, celle qui leur permet de remplir une fonction. Lorsqu'elles sont correctement repliées, tout fonctionne bien. Cependant, si elles se plient mal, ce qu'on appelle défaut de repliement, cela entraîne des problèmes de santé. Notre corps est normalement capable d'identifier ces protéines mal repliées dites anormales et de les détruire. Néanmoins, dans certaines maladies, le corps produit trop de protéines anormales qu'il ne peut plus traiter, il fini par être dépassé et n'arrive pas à les éliminer. Ces protéines mal repliées s'appellent protéines amyloïdes. Lorsqu'elles s'accumulent, elles se collent les unes aux autres et prennent la forme de fibrilles rigides linéaires. Ces fibrilles s'accumulent elles aussi dans les organes et forment des plaques amyloïdes qui empêchement leur bon fonctionnement. Si les organes sont infiltrés par ces plaques on parle d'amylose. Alors, l'amylose est une maladie liée au repliement anormal des protéines dans le corps (Georgin-Lavialle et al. 2017).

L'amylose est une maladie peu fréquente, ayant une incidence d'environ 3 sur 5 personnes par million et par année. Elle est plus fréquente dans les hommes et l'âge moyen d'apparition ou de développement est de 65 ans (Ruiz-Mori et al. 2018).

Les symptômes rencontrés sont souvent vagues et peu spécifiques. À cause de la variété des symptômes initiaux et de la rareté de la maladie, mal connue par les médecins à vrai dire, le diagnostic est habituellement retardé. Selon (Georgin-Lavialle et al. 2017), une étude menée par l'association de patients américains a montré que « le temps entre les premiers symptômes et le diagnostic était d'environ un an et que les patients avaient vu en moyenne cinq médecins avant que le diagnostic ne soit fait. » Il est donc essentiel que le diagnostic soit réalisé le plus rapidement possible de façon que l'aggravation progressive des organes atteints puisse être évitée, en particulier la cardiaque. Le seul moyen de faire le diagnostic de certitude d'amylose aujourd'hui est de réaliser une biopsie (Georgin-Lavialle et al. 2017). Dans le laboratoire, le tissu de la biopsie est coloré à l'aide d'une substance qui s'appelle Rouge Congo, qui colore les dépôts d'amylose.

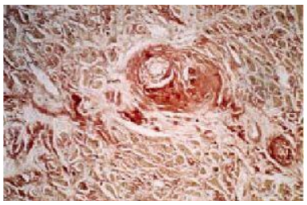


FIGURE 1.1 – L'amylose est reconnue sur les biopsies grâce à la coloration au Rouge Congo. Amyloïdose cardiaque : rouge Congo

Ils existent plusieurs types d'amylose. Elles sont normalement classées en fonction du nom de la protéine précurseuse. La nomenclature utilisée est simple, elle consiste à utiliser la lettre « A » suivie de l'abréviation du nom de la protéine. De plus, les symptômes et les organes touchés dépendent du type d'amylose, ainsi que les traitements proposés aux patients (GEORGIN-LAVIALLE et al. 2017).

Nous pouvons proposer trois types principaux : ATTR, connue également comme amylose à la Transthyrétine, amylose familial ou héréditaire ou amylose sénile ; amylose AL, L pour « Light Chain » ; et amylose AA, pour Sérum Amyloïde A. Cependant, on parle aussi d'amyloses localisées lorsqu'elles ne concernent qu'un organe (dans la grande majorité des cas des amyloses AL) et des amyloses cardiaques.

Nous effectuons par la suite une brève présentation des amyloses les plus repandues (GEORGIN-LAVIALLE et al. 2017).

### Les amyloses héréditaires rares

Dans ce cas, la protéine de structure anormale qui s'accumule est causée par une mutation génétique. La plus fréquente est liée aux mutations de la protéine transthyrétine qui cause essentiellement une atteinte des nerfs périphériques et du cœur. Il y a aussi plusieurs autres formes d'amyloses héréditaires causées par des mutations d'autres protéines comme la chaîne alpha du fibrinogène ou celles de la famille des aplopoprotéines (GEORGIN-LAVIALLE et al. 2017).

Ces formes sont très rares et n'ont pas de caractéristiques cliniques spécifiques, ce qui est particulier c'est le caractère familial de la maladie, avec

présence d'autres cas d'amyloses dans la famille. Cette répartition est dû au mode de transmission génétique de la maladie qui est « dominant ». S'il existe un cas d'amylose dans une famille, il faut penser en premier lieu à l'amylose de la transthyrétine, mais aussi à ces formes plus rares. Les amyloses héréditaires rares touchent d'abord les reins. Les symptômes cliniques sont tardifs : gonflement des chevilles, élévation de la pression artérielle, et symptômes généraux provoqués par l'insuffisance rénale (fatigue, pâleur, essoufflement, démangeaison). Il y a aussi d'autres anomalies biologiques présentes, à savoir la présence de protéine en excès dans l'urine et l'élévation de la créatinine sanguine (Georgin-Lavialle et al. 2017).

En ce qui concerne les traitements, il faut d'abord contempler le traitement de l'atteinte rénale, souvent par la dialyse et la transplantation rénale.

### Les amyloses héréditaires à Transthyrétine

L'amylose héréditaire à transthyrétine est due à une anomalie d'une région de l'ADN. Ce gène sert à fabriquer la protéine transthyrétine ou TTR. L'amylose à TTR touche principalement les nerfs (neuropathie) et le cœur, mais aussi rarement les yeux et les reins. Il s'agit normalement d'une maladie très grave et invalidante étant donné le manque de traitement et de prise en charge spécifique (Georgin-Lavialle et al. 2017).

Plus de 500 cas ont été diagnostiqués en France (Georgin-Lavialle et al. 2017). Elle concerne l'adulte entre 20 et 88 ans, étant l'âge de début très variable : vers 30 ans chez les patients originaires du Nord du Portugal, où des très nombreuses familles sont atteintes, ou plus tardif, vers 60 ans, pour les sujets d'origine non portugaise.

Les manifestations des amyloses héréditaires à Transthyrétine peuvent être :

— Manifestations neurologiques : dysfonctionnement des nerfs périphériques sensitifs et moteurs. Il peut s'agir de perte de sensibilité des pieds, douleurs spontanées et intenses (à type de « brûlures ») ; faiblesses de pieds ou plus rarement des mains pouvant gêner la marche, difficultés à la marche ou troubles d'équilibre ; des troubles végétatifs peuvent éga-

lement survenir par lésions des nerfs commandant les systèmes digestif, sexuel, cardiovasculaire et urinaire avec des nausées, lenteur de digestion, perte d'appétit, constipation, diarrhées, sensations de vertige ou perte de connaissance au lever (Georgin-Lavialle et al. 2017).

— Manifestations en rapport avec la cardiopathie : il peut s'agir de troubles conductifs (ralentissement du rythme cardiaque, arrêt cardiaque), augmentation de l'épaisseur et de la rigidité cardiaque, essoufflement à l'effort, atteinte des nerfs du système cardiovasculaire (chute de tension au lever, perte des variations normales de la fréquence cardiaque, anomalies en scintographie à la MIBG [1]) (Georgin-Lavialle et al. 2017).

— Manifestations oculaires : il peut se montrer par des dépôts vitréens (par la perception de « corps flottants » ou de « mouches volantes »), par le glaucome, ou par une sécheresse oculaire, qui se manifeste par une sensation de picotement (Georgin-Lavialle et al. 2017).

Les traitements de l'amylose ATTR visent principalement à empêcher la formation de nouveaux dépôts d'amylose en stabilisant la protéine TTR ou en bloquant sa production (Georgin-Lavialle et al. 2017).

— Transplantation hépatique ou greffe de foie : l'objectif est de supprimer le principal organe producteur de protéine TTR.

— Médicament « anti-amyloïdes ». Le Tafamidis ®(Vyndaqel®) est le seul médicament autorisé actuellement par les autorités françaises. Il s'agit d'un stabilisateur de la transthyrétine mutée qui permet de ralentir la progression de la maladie.

### Les amyloses AA

Dans ce type d'amylose, le foie produit en excès la protéine appelée sérum amyloïde A. Les organes les plus touchés sont les reins, le tube digestif, le foie

1. La scintigraphie à la MIBG (métaiodobenzylguanidine) est un examen d'imagerie en médecine nucléaire. Elle associe une petite quantité de matière radioactive à une substance appelée métaiodobenzylguanidine (MIBG) pour détecter certains types de tumeurs dans le corps. (Bombardieri, Gianmmarile et Aktolun 2010)

et les glandes salivaires.

Les amyloses AA (Georgin-Lavialle et al. 2017) sont souvent liées à d'autres maladies inflammatoires comme des maladies articulaires (rhumatismales), maladies inflammatoires du tube digestif, des tumeurs et hémopathies et à l'obésité.

Le but pour les amyloses AA est de réduire l'excès circulant dans le sang de la protéine SAA. Pour arrêter la production de cette protéine, il faut interrompre l'inflammation, donc soigner la maladie qui l'entraîne. Les traitements dépendent donc de la maladie et peuvent comporter des antibiotiques (en cas de maladie infectieuse), des corticoïdes ou des biothérapies (Georgin-Lavialle et al. 2017).

### Les amyloses AL

L'amylose AL (ou immunoglobulinique) est une maladie rare liée aux dépôts dans différents organes d'une partie (la chaîne légère) d'un anticorps anormal dit monoclonal sous forme de fibrilles. Il est estimé qu'il y a de 500 à 700 nouveaux cas en France tous les ans et que l'âge moyen du diagnostic est d'environ 65 ans (Georgin-Lavialle et al. 2017).

L'organe le plus souvent touché est le rein, chez plus de 2/3 patients. Le patient peut avoir des œdèmes au niveau de jambes ou plus disséminés, et épreuve aussi une insuffisance rénale qui rend parfois nécessaire la pratique d'une dialyse. L'atteinte du cœur entraîne de fatigue et essoufflement, d'abord à l'effort puis au repos. D'autres conséquences sont l'apparition des rythmes anormaux, trop rapides ou trop lents, voire l'arrêt cardiaque.

Environ 20% des patients ont une atteinte neurologique qui se manifeste par des sensations anormales au niveau de pieds puis des jambes, des cuisses et des mains. Elle peut aussi entraîner des troubles digestifs et une baisse de la tension artérielle quand on passe de la position couchée à la position debout entraînant une sensation de malaise et éventuellement de chutes (Georgin-Lavialle et al. 2017).

Les traitements visent d'abord à éliminer les cellules qui fabriquent les anti-

corps monoclonaux par des protocoles de chimiothérapie, le plus souvent dérivés de ceux de myélome (Georgin-Lavialle et al. 2017). Pour que la maladie améliore, il faut que le niveau de chaînes légères baissent le plus possible.

#### Les amyloses localisées

La plupart des amyloses se caractérisent par l'atteinte de plusieurs tissus et organes, d'où qu'on les appelle multisystémiques. Cependent, certaines formes d'amylose ne concernent qu'un organe, normalement il s'agit des cas d'amyloses AL (Georgin-Lavialle et al. 2017).

Les symptômes sont en rapport avec les organes atteints :

— Dysphonie, avec une voix qui se modifie progressivement, toux, essoufflement.

— Présence de sang dans les urines, difficultés à la miction, douleur du bas ventre pour les amyloses urinaires.

### 1.2.2 Amyloïdose cardiaque

Étant donné que notre projet concerne l'amyloïdose cardiaque, nous allons consacrer une section à expliquer de façon plus détaillée cette type de maladie.

On parle d'amyloïdose cardiaque lorsque le cœur ne fonctionne plus correctement à cause de l'accumulation de fibrilles amyloïdes. Cette infiltration altère peu à peu la fonction du myocarde (Georgin-Lavialle et al. 2017). Plusieurs protéines peuvent être en cause et vont donner son nom à la maladie, par exemple, amylose cardiaque de type AL. On ne peut pas parler à priori « d'une » amylose cardiaque au singulier, puisqu'il existe plusieurs maladies pouvant conduire à la création de protéines fibrillaires dans le cœur.

Les principaux types d'amylose atteignant le cœur sont l'amylose AL, l'amylose sénile et certaines amyloses héréditaires, dont l'amylose à TTR. Ainsi, l'amylose cardiaque peut être isolée ou s'intégrer dans une maladie atteignant plusieurs autres organes, et donc être associée à des nombreux symptômes, d'où les difficultés du diagnostic (Georgin-Lavialle et al. 2017).

Les manifestations sont non spécifiques, pouvant comporter des symptômes d'insuffisance cardiaque (dyspnée) et/ou des troubles conductifs et rythmiques (Damy 2014). Même si les symptômes sont difficiles à observer, ils existent plusieurs anomalies qui peuvent être considérées comme évocatrices d'amylose sur des examens complémentaires cardiologiques (Damy 2014).

- Les dépôts dans le muscle cardiaque sont responsables d'un épaississement du cœur appelé « hypertrophie myocardiaque », qui conséquemment produit l'apparition d'une rigidité entraînant la limitation de la capacité du cœur à se remplir. Plus tard, le cœur a du mal a se contracter, ce qui est à l'origine d'une baisse du débit cardiaque et occasionne différents symptômes dits « d'insuffisance cardiaque » : essoufflement à l'effort puis au repos, fatigue à l'effort puis au repos, rétention d'eau au niveau de jambes principalement, prise de poids, et palpitations (Georgin-Lavialle et al. 2017).
- Les dépôts dans le circuit électrique perturbent le rythme du cœur, qui devient trop rapide ou trop lent, et peut même conduire à un arrêt cardiaque. Cette atteinte nécessite souvent d'un simulateur cardiaque (« pacemaker »), qui permet de contrôler le rythme, ou d'un défibrillateur. Les symptômes résultant de ces perturbations peuvent être : malaises, syncopes, palpitations, etc.(Georgin-Lavialle et al. 2017)

L'amylose cardiaque est recherchée dans deux types de situations :

1. Devant les symptômes précédemment mentionnés. Certains examens permettent de montrer les atteintes cardiaques caractéristiques d'amylose cardiaque. L'épaississement cardiaque ou hypertrophie myocardiaque est mise en évidence par l'écho-cardiographie trans-thoracique et l'IRM cardiaque. L'atteinte du rythme est montrée par un électrocardiogramme standard (ECG - enregistrement des signaux électriques du cœur) et prolongé, aussi appelé Holter-Electrocardiogramme (Holter ECG). Si ces atteintes sont prouvées, il faut alors confirmer le dépôt amyloïde par une biopsie. Dans ce cas, le cœur est alors le premier ou le seul organe atteint par l'amylose (Georgin-Lavialle et al. 2017).

2. Chez un patient déjà atteint par une amylose. Si celui présente aussi des symptômes évocateurs de l'atteinte cardiaque, on utilise encore l'échocardiographie, l'IRM cardiaque, l'ECG et le Holter ECG (Georgin-Lavialle et al. 2017).

Le traitement proposé pour l'amyloïdose cardiaque dépend du type d'amylose et des symptômes. En général, la prise en charge comprend deux types de traitements : ceux qui visent à améliorer les symptômes et prévenir les complications et ceux qui visent à stopper la formation des fibrilles amyloïdes (Georgin-Lavialle et al. 2017).

- Les traitements visant à améliorer les symptômes ne sont pas propres à cette maladie. L'insuffisance cardiaque est traitée par des diurétiques et un régime pauvre en sel. Les troubles du rythme cardiaque peuvent être traités ou prévenus par un stimulateur cardiaque (« pacemaker »).
- Les traitements anti-amyloïdes sont adaptés au type d'amylose.

Comme l'on a déjà mentionné, le corps médical, même les cardiologues, connaissent encore mal ces pathologies et le diagnostic d'amylose est souvent réalisé tardivement. Aujourd'hui, un essai clinique est en cours à l'hôpital Henri Mondor avec pour objectif de « tester un nouveau traitement capable d'enrayer le développement de la maladie » (Adams et al. 2009). Même s'il empêche l'aggravation de la maladie, le traitement ne permet pas de supprimer les dégâts déjà produits sur le cœur, d'où l'intérêt à développer des méthodes pour la prévenir.

D'autre part, le suivi est essentiel dans la prise en charge de patients avec une amylose. Les objectifs sont de surveiller la progression de l'atteinte cardiaque, de s'assurer de l'efficacité et de l'absence de toxicité des médicaments mis en place et de détecter précocement les atteintes d'autres organes (Georgin-Lavialle et al. 2017). Tout cela implique la répétition des examens d'imagerie (echographie, IRM) et des examens biologiques (dosages des chaînes légères libres dans le sang, marqueurs cardiaques : BNP, NT-proBNP ; marqueurs d'atteinte rénale : protéinurie, micro-albuminurie, bandelette urinaire).

Le pronostic de l'amyloïdose cardiaque dépend considérablement du type d'amylose en cause et des autres organes atteints. Quoiqu'il en soit, l'amylose cardiaque constitue une atteinte grave. Cependant, si elle est identifiée et prise en charge à temps, sa progression peut être ralentie.

En France, plus de 500 cas ont été diagnostiqués. Cette maladie concerne l'adulte (de 20 à 88 ans) et l'âge de début est très variante (ADAMS et al. 2009).

## 1.3 Anonymisation

D'abord, il faut étudier le signifié du terme « anonymisation » et pourquoi son rôle est tellement important dans le cadre du traitement des langues implémenté au domaine médical.

En langue générale, le dictionnaire Larousse propose la définition de « rendre anonyme » pour le verbe anonymiser, et celle de « fait d'anonymiser ; son résultat » pour le nom « anonymisation ». De son côté, la Real Academia Española fournit une définition pour le verbe « anonimizar » ; par contre, le substantif n'est pas recueilli par cette ressource.

*Anonimizar : expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad.* (Exprimer une donnée relative à une entité ou personne, tout en supprimant les références à son identité.)

Si l'on applique ce terme au corpus, l'anonymisation suppose généralement le fait de masquer des informations qui identifient un individu, tout en conservant le reste d'informations présentes dans le document. L'un des objectifs principal de notre travail concerne cette problématique.

L'objectif de la tâche d'anonymisation est évident : masquer les informations qui permettent l'identification d'un individu, même si cette tâche elle-même devient plus compliquée lorsqu'on cherche quels sont les indices pour anonymiser. Toutefois, il faut se poser plusieurs questions. Quel est le traitement que l'on va appliquer à ces informations ? Par quoi est-ce qu'on va remplacer les informations ciblées ? Une première possibilité consisterait à remplacer les informations par une suite de caractères à notre choix, par exemple, plusieurs

caractères « X ». On pourrait aussi les remplacer par des balises XML, qui décriront le type d'information qu'a été anonymisé à cet endroit-là, c'est-à-dire, remplacer par exemple tous les noms rencontrés dans le corpus par la balise « <nom> ». D'une part, la première option semble la plus facile mais aussi la plus pauvre puisqu'elle suppose une perte d'information qui peut se révéler important concernant plusieurs aspects comme la catégorie sémantique de l'information anonymisée, les écarts temporels entre les dates, la distinction de plusieurs intervenants, etc. D'autre part, il faudrait obtenir un taux de rappel de 1 pour les deux méthodes, puisque tous les cas de faux négatifs « sauteront aux yeux », cela veut dire, si jamais un prénom ou une date ne sont pas repérés et que le modèle ne les remplace pas, on s'apercevra rapidement.

Une deuxième solution consisterait à remplacer les informations du document par d'autres vraisemblables de la même catégorie, c'est-à-dire, un nom par un autre nom ou une date par une autre date. Dans ce cas, il faudrait faire attention à remplacer un nom par le même nom dans la totalité du document et conserver les écarts temporels réels entre les dates. Si jamais on connaît la corrélation des échanges, on parlerait donc de « pseudonymisation », une méthode qui nous permet de discriminer les différents individus présents dans les documents ; on saura toujours de qui on parle. Par ailleurs, une autre méthode consisterait à implémenter une combinaison de ces deux dernières solutions : utiliser des balises XML avec un identifiant numérique unique pour chaque occurrence pour les informations dites nominales, par exemple, « <nom1> » ou « <prénom1> », et des dates fictives mais qui conserveraient les écarts.

On vient d'introduire précédemment un nouveau concept, celui de « pseudonymisation ». Analysons à quoi cela correspond et quelle est la différence avec la notion d'anonymisation.

Selon le CCIN (Commission de Contrôle des informations nominatives) (NOMINATIVES 2019), « l'anonymisation est une technique consistant à supprimer tout caractère identifiant à un ensemble de données ». D'après la norme ISO 29100, il s'agit du « processus par lequel des informations personnellement identifiables (IPI) sont irréversiblement altérées de telle façon que le sujet des

IPI ne puisse plus être identifié directement ou indirectement, que ce soit par le responsable du traitement des IPI seul ou en collaboration avec une quelconque autre partie » (MANAGEMENT et GOVERNANCE 2018). La caractéristique principale de l'anonymisation est donc l'irréversibilité de la perte du caractère identifiable d'individus.

En revanche, la pseudonymisation ou « anonymisation réversible » (NOMINATIVES 2019) « consiste à remplacer un attribut par un autre dans un enregistrement. La personne physique est donc toujours susceptible d'être identifiée indirectement ». Par exemple, même si on codifie le nom d'un individu, cela n'empêcherait pas son individualisation s'il est possible d'avoir accès à d'autres attributs comme son sexe, son adresse ou sa date de naissance. La pseudonymisation donc réduit « le risque de corrélation directe entre des informations nominatives » mais n'efface pas le caractère nominative des informations exploitées. Par conséquent, et comme exprime le CCIN, « la pseudonymisation n'est pas une forme atténuée d'anonymisation, mais une simple mesure de sécurité ».

En plus, ces deux méthodes ont des objectifs distincts. Le choix de l'une ou de l'autre procède moins d'un choix technique que du besoin de conserver ou non les informations nominatives. La pseudonymisation s'utilise dans des situations qui nécessitent ou envisagent un retour en arrière des informations nominatives codées aux informations nominatives originales. Par contre, l'anonymisation « ne s'inscrit pas dans une démarche ultérieure de ré-identification et elle n'a pas vocation à permettre un tel retour en arrière ».

Le règlement nº 2016/679 (CONSEIL DE L'UNION EUROPÉENNE 2016), dit règlement général sur la protection des données (GDPR), est un règlement de l'Union européenne qui constitue le texte de référence en matière de protection de données de caractère personnel. Ce règlement renferme la différence entre les termes « anonymisation » et « pseudonymisation ». Dans l'article 4 (5), ils parlent de la pseudonymisation comme « *. . . the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided*

*that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified natural person* ».

La directive 95/46/EC fait référence à l'anonymisation dans l'article 26. Cet article affirme que pour qu'une information soit anonymisée, elle doit être dépourvue des éléments suffisants tel que l'objet des données ne soit plus identifiable (PARTY 2014).

En conclusion, la pseudonymisation et l'anonymisation répondent à des buts différents : conserver ou non le caractère personnel des informations, cela implique le fait que le procès soit réversible ou non.

Mesytre et al. (MEYSTRE et al. 2010) distinguent aussi entre les termes « *anonymization* » et « *de-identification* », qui s'utilisent souvent de façon interchangeable. Cependent, la désidentification signifie que les identifiants explicites sont cachés ou enlevés, tandis que l'anonymisation implique que les données ne peuvent pas être liés pour permettre d'identifier les patients, c'est-à-dire, les données desidentifiées sont loin d'être anonymes. « *Scrubbing* » est aussi employé souvent comme un synonyme d'anonymisation.

Aux États Unis, l'emploi de données dérivées de dossiers médicaux est régi par les régulations HIPAA (Privacy Rule of Health Insurance Portability and Accountability Act) (MURPHY et al. 2007) et doivent respecter certains exigences en matière de sécurité. Plus particulièrement, pour que des données identifiables récupérées des dossiers médicaux puissent être utilisés pour faire de la recherche, elles doivent être débarrassées de toute information identifiable et, en plus, le patient aurait dû être notifié avant que les données soient introduites dans le dossier médical. Même ayant été notifié, les données identifiables ne peuvent pas être libérées qu'avec l'approbation du Institutional Review Board (IRB) (MURPHY et al. 2007).

Après avoir analysé les différents concepts existants autour de l'anonymisation, nous pouvons conclure que le processus de masquage des informations identifiantes lors de notre projet corresponde au concept de « pseudonymisation » parce qu'une fois que l'on aurait réussi à diagnostiquer des patients,

il sera nécessaire de pouvoir les identifier afin de vérifier s'ils sont vraiment atteints d'amyloïdose. Désormais, nous parlerons de « désidentification » ou « pseudonymisation ».

Finalement, en ce qui concerne les informations à anonymiser, il faut définir les catégories d'informations que l'on souhaite masquer. En général, on peut distinguer plusieurs types d'informations susceptibles de permettre l'identification d'un individu et qui doivent obligatoirement être l'objet du processus de désidentification : les informations nominatives comme les noms, les prénoms ou les noms de lieux, et les informations numériques comme la date de naissance, d'admission, d'opération, le numéro de téléphone, le numéro de la sécurité sociale, etc. En ce qui concerne les informations qui doivent être enlevées des documents médicaux, l'HIPAA indique dix-huit éléments considérés comme *protected health information* (PHI) (Kayaalp et al. 2015), qui nécessitent d'être desidentifiés afin de protéger la vie privée du patient. Ces catégories (Meystre et al. 2010) comprennent entre autres des noms et prénoms, emplacements géographiques (plus spécifiques qu'un état y compris des adresses de rues, villes, codes postales ou équivalents sauf pour ceux qui ont moins de 20000 habitants), éléments de dates sauf les âges (sauf celles au-dessus de 89 ans), numéro de sécurité sociale, âges au dessus de 90 ans, numéros de téléphone et fax, numéro de dossier médical, adresse mail, numéros de compte bancaire, adresses IP, identifiants biométriques, etc.

On pourrait même réaliser un repérage d'entités nommées associées aux catégories pré-définies pour ensuite procéder à l'anonymisation de ces entités.

Comme nous avons mentionné précédemment, la disponibilité de données cliniques nous permet désormais d'en extraire des connaissances et les implémenter au sein de différentes tâches de la médecine. Cependant, bien que l'approche numérique dont nous avons parlé auparavant peut aboutir à des grandes performances dans le domaine de la médecine, elle a besoin impérativement de données parfaitement propres et bien annotées. Étant donné que les données cliniques ne sont pas recueillies avec comme objectif de les implémenter dans des algorithmes d'intelligence artificielle, elles posent beaucoup de difficultés

lors de leur exploitation. Le principal problème est que la plupart des données médicales stockent à la fois des données personnelles avec lesquelles l'on ne peut pas travailler puisque cela attenterait à la confidentialité des données des patients. Il faut toujours respecter les lois, plus spécifiquement le règlement sur les données personnelles (Conseil de l'Union Européenne 2016) mentionné précédemment, entrée en vigueur en mai 2018. Il existe également une loi en Espagne qui vise la protection des données personnelles et la garantie des droits numériques. Tant en France comme en Espagne, les données de santé doivent être anonymisées pour qu'ils puissent être accessibles par les chercheurs.

Les données dont nous disposons ont été extraites à partir de grandes bases de données, desquelles on a pris les informations qui étaient considérées pertinentes pour le projet de recherche. La plupart de ces données étaient des données structurées et on a utilisé un identifiant unique personnel pour croiser tous les tables et obtenir une base de données qui recueille toutes les données disponibles. Dans ces données, il n'y a aucune information personnelle du patient, seulement le code identifiant de celui-ci, à travers lequel il est impossible d'identifier à chaque individu.

Cependant, une partie de la base de données fournie correspond aux données en format texte libre et celle-ci contient des données personnelles, donc il est indispensable que nous effectuons un processus de masquage de données personnelles. Cette tâche a ainsi deux objectifs : donner la meilleure couverture possible de désidentifications en même temps que l'on préserve au maximum la qualité du document pour qu'il puisse être utilisé dans un processus de recherche futur.

Finalement, les données personnelles identifiables que l'on nous a demandé de désidentifier correspondent aux noms et prénoms, numéros de téléphones, numéro d'identification personnelle et numéro de la sécurité sociale.

## 1.4 Propriétés des informations médicales

Comme nous avons expliqué auparavant, le premier problème que l'on rencontre lors de l'exploitation de données médicales c'est que le 80% des informations sur les patients se trouvent sous forme textuelle (Charlet et al. 2018). Par exemple, les comptes rendus d'hospitalisation ou les rapports d'imagerie médicale. D'après (Charlet et al. 2018), il est donc essentiel de « mettre en oeuvre des logiciels de traitement automatique des langues pour analyser ces textes et en extraire des informations sur les patients ». Ces logiciels peuvent s'enrichir plus tard soit des approches symboliques soit des approches fondées sur les réseaux de neurones ou d'algorithmes de *machine learning*. De plus, les algorithmes d'apprentissage non-supervisé, c'est-à-dire, qui n'apprennent pas préalablement sur des échantillons, « susparencitent des espoirs » dans le domaine, puisqu'ils permettent en effet de recouper de façon très rapide une grande quantité de données avec l'objectif de définir des structures cachées et de déterminer des catégories d'intérêt pour la tâche visée. Parmi les ambitions de cet apprentissage se trouvent celles d'identifier des facteurs de risque, personnaliser les traitements ou prédire des maladies, comme c'est notre cas. Par contre, il ne faut pas oublier que, même si ces algorithmes semblent être assez performants, ils ne garantissent pas une efficacité à 100% et nécessitent un travail de contrôle a posteriori réalisé par des humains.

## 1.5 Prédiction

Dans le domaine de la santé, le *big data*, ou données massives, correspond à l'ensemble des données socio-démographiques et de santé, disponibles auprès de différentes sources et qui sont collectées pour diverses raisons. L'exploitation de ces données suppose de nombreux intérêts comme l'identification de facteurs de risque de maladie, l'aide au diagnostic, au choix et au suivi de l'efficacité de traitements, etc. Par contre, elle soulève des nombreux défis techniques et humains, et pose autant de questions éthiques (Charlet et al. 2018).

Les deux problèmes majeurs que l'on rencontre dans le domaine de la recherche de la prédiction de maladies sont l'exploitation des énormes volumes de données en plus de la présence de données personnelles.

En ce qui concerne la première des difficultés, cette énorme quantité de données désormais disponible soulève des défis techniques, concernant leur stockage et aussi les capacités d'exploitation. D'autre part, les informations que l'ont été collectées sont en effet hétérogènes à cause de leur nature (génomique, physiologique, biologique, clinique, sociale...) et leur format (texte, valeurs numériques, signaux, images...).

Pour que leur traitement et leur exploitation puissent être menés à terme, ces informations doivent être intégrées dans des bases de données de manière structurée. En France s'utilisent des standards tel i2b2 (Informatics for Integrating Biology and the Bedside), développé à Boston et financé par les National Institutes of Health, il est désormais exploité au CHU de Rennes, à Bordeaux et à l'Hôpital européen Georges Pompidou de Paris, en plus d'autres 200 institutions dans le monde (Murphy et al. 2007). C'est une plate-forme d'analyse de données clinique open-source qui transforme les données de patients récupérées à partir des EHR (Electronic Health Record) en un format optimisé pour plusieurs types et étapes de recherche.

Ce système a été utilisé par exemple pour identifier et quantifier le risque d'infarctus du myocarde (Murphy et al. 2007). Grâce à ces standards, les hôpitaux et les centres des soins ont plus des capacités à l'heure de compiler toutes les données collectées dans des entrepôts de données biomédicales, interrogeables par les chercheurs via des interfaces web.

D'autre part, l'on trouve les défis éthiques du *big data*, notamment dans le domaine de la santé. Au regard du droit français, les données de santé constituent des données personnelles dites « sensibles ». Qu'est-ce que cela veut dire ? Elles méritent une protection accrue étant donné leur nature. Lors d'un essai clinique, un consentement est nécessaire avant le recueil de données de santé (Thiebaut et Inserm 2016). Cependant, lorsque les chercheurs mènent à bien un projet, ils utilisent des données des patients qui probablement n'ont pas été

prévenus pour l'utilisation de celles-ci avec l'objectif de faire de la recherche. Évidemment, cela pose des problèmes éthiques relatifs au souhait des patients de partager ou non ces données avec des tiers, ainsi que sur la préservation de l'anonymat. D'autres questions se posent comme qui doit gérer les données et sous quelles conditions. Le risque de divulgation de données privées et d'autres problématiques relatives aux données personnelles appartenant aux données médicales font l'objet d'avis de la part des comités d'éthique, dont le Comité consultatif national d'éthique en France.

Les pouvoirs publics se sont aussi saisi de la question : la loi de modernisation du système de santé français promulguée le 26 janvier 2016 prévoit l'ouverture des données agrégées de santé à des fins de recherche, d'étude ou d'évaluation d'intérêt public, tout citoyen, professionnel de santé ou organisme participant du fonctionnement du système de santé et aux soins. Cette ouverture est assortie à une condition fondamentale : les données ne doivent pas permettre l'identification des personnes concernées.

Pour y avoir accès, tout organisme de recherche ou d'étude qui voudrait mener un projet doit soumettre ce dernier à l'Institut national des données de santé.

# Section 2

# État de l'art

## 2.1 Les corpus de textes médicaux

Au niveau médical, nous pouvons distinguer trois types principaux de documents qui ont des contenus tout à fait opposés dû à leur finalité : les articles scientifiques, les forums de santé et sites de spécialité, et les documents cliniques qui composent habituellement un dossier médical. Notre travail repose sur les deux derniers types de documents. Il faut remarquer que tous s'opposent autant en contenu qu'en termes d'accès. Bien que l'on puisse facilement récupérer des articles scientifiques, et avec un peu plus de difficultés des messages de forums, les documents cliniques nécessitent de traitements et comportent de contraintes juridiques sur leur utilisation et diffusion. C'est la raison pour laquelle un processus d'anonymisation devient indispensable dans ce domaine (Grouin 2013). Du fait qu'initialement notre projet était basé sur des dossiers cliniques, nous nous intéressons aussi à ce type de documents médicaux.

Ces documents cliniques contiennent des données relatives aux patients et sont composés souvent de plusieurs types de documents en même temps, qui sont rédigés lors des différentes étapes de la vie du patient, comme un séjour hospitalier, des consultations externes, des visites aux urgences, etc. Ce dossier contient donc des informations nominatives, en particulier l'identité du patient (nom, prénom, date de naissance, numéro d'identification. . . ) mais aussi celles des personnes à prévenir et des professionnels de la santé en charge du patient.

Dans l'ensemble des documents médicaux d'un patient on trouvera informations cliniques essentielles pour la recherche scientifique, comme les motifs d'hospitalisation, la recherche d'antécédents et de facteurs de risques, les conclusions de l'évaluation clinique initiale, le type de prise en charge prévu, les prescriptions effectuées à l'entrée, les prescriptions médicales, les examens complémentaires, les comptes rendus d'hospitalisation, etc. Ces documents comportent un corpus de textes qui constituent une ressource cruciale en traitement automatique des langues.

Grâce aux corpus élaborés à partir des documents médicaux, on peut soit développer d'outils soit les adapter afin de mieux apprendre les spécificités linguistiques d'une nouvelle thématique ou d'un nouveau domaine. Nous pouvons aussi les employer dans l'entraînement de systèmes qui reposent sur des méthodes d'apprentissage statistique, qui construisent des modèles basés sur les observations faites en corpus de façon qu'ils soient capables de modéliser des tâches comme la classification, l'indexation, etc. Il faut souligner que ces systèmes nécessitent une grande quantité d'observations pour réaliser leur apprentissage. Finalement, ces corpus servent aussi à évaluer les systèmes. Nous pouvons ainsi confronter différents systèmes sur un même jeu de données afin de vérifier quelles méthodes et approches se révèlent les plus performantes.

En outre, il faut tenir compte du fait que nous sommes face à un corpus qui emploie une langue de spécialité, plus spécifiquement, la langue médicale, et qu'il faudra adapter notre travail à ses particularités. Une langue de spécialité « naît du besoin que ressentent les spécialistes de communiquer entre eux de façon concise et sans ambiguïté » (Rouleau 1995). Il s'agit donc de la façon dont s'expriment les gens qui travaillent dans un même domaine ou sous-domaine de l'activité humaine.

À l'instar d'autres langues de spécialité, la langue médicale se caractérise par la présence très fréquente de termes techniques. De plus, elle repose sur des bonnes pratiques pour la rédaction et dispose de ses propres codes et caractéristiques ainsi qu'un vocabulaire que lui est propre. Si nous allons implémenter des outils génériques de traitement du langage au traitement de la langue mé-

dicale, ce traitement aura besoin néanmoins des ressources dédiées, comme par exemple, des lexiques spécifiques, adaptation des règles de reconnaissance d'entités nommées, etc.

Comme toute langue de spécialité, la langue médicale a donc ses propres particularités, nous voudrions en remarquer trois : sa création, son évolution et sa composition.

La maîtrise du vocabulaire médical repose sur la connaissance des racines grecques et latines (ROULEAU 1995). L'étymologie pourrait donc être d'aide pour le repérage de termes médicaux. Peut-être qu'on pourrait créer des lexiques ou des dictionnaires avec ces racines afin d'étiqueter automatiquement ces termes.

En plus, la langue médicale est toujours en croissance dû à la création constante de termes. Les locuteurs utilisent cette stratégie lorsqu'ils ont besoin d'un « moyen économique de nommer la réalité », c'est-à-dire, de donner un nom à une réalité pour laquelle il n'existe encore un terme qui la recueille. Il est évident que le langage et presque toutes les langues de spécialité sont en constante évolution, mais nous avons décidé de remarquer ce phénomène du fait qu'il faudra tenir compte de cette évolution pour intégrer les nouveaux constructions et tenir compte des nouveaux termes dans notre système de repérage d'entités nommées du domaine médical et de désidentification. L'exemple expliqué par Rouleau (ROULEAU 1995) montre bien ce phénomène. Considérons le terme inhibiteur calcique, qui a été créé pour désigner un composé qui inhibe le passage à travers la membrane cellulaire des ions calcium grâce à des unités fonctionnelles, les canaux calciques. L'économie qui proportionne ce terme est indéniable. Un autre exemple est celui du terme « greffé ». Le fait de devoir utiliser une expression plus longue comme « receveur de greffe » à chaque fois qu'un médecin doit parler de ce type de patient a donné lieu à l'expression « greffé » ou « transplanté ». Du point de vue linguistique, ce serait l'organe qui est greffé et non le patient ; par conséquent, utiliser ce terme pour désigner le sujet qui a reçu une greffe est contraire au sens désiré. Cependant, les médecins ont suivi la formation d'autres termes fortement utilisés appartenant aussi bien

à la langue médicale qu'à la langue générale comme ménopausée, handicapé, accidenté, etc. Ces derniers phénomènes supposent sans doute des difficultés dans notre travail dans le cadre du traitement automatique des langues.

Une autre caractéristique de la langue médicale est la construction des termes formés d'un substantif suivi d'un ou plusieurs adjectifs, par exemple : « infarctus pulmonaire », « abdomen aigu » ou « leucémie lymphoïde chronique ». Il faudra tenir compte de ces termes composés ou « expressions figées » comme ceux-ci pour prévenir leur présence dans notre corpus afin de bien les repérer.

Toutes ces caractéristiques pourraient être conceptualisés sous forme de règles ou lexiques afin d'aider au repérage des termes médicaux désignant des symptômes, anomalies, maladies, etc., qui seront utilisés plus tard pour la classification dans la prédiction de la maladie.

## 2.2 L'intelligence artificielle et la recherche médicale

Le succès obtenu suite à l'application du *data mining* dans des domaines avec une forte visibilité comme l'*e-business*, le marketing et le *retail* a donné lieu à l'application de ces méthodes dans d'autres industries. Parmi ces secteurs se trouve celui de la santé. L'environnement médical est encore très « riche » en information mais « pauvre » en connaissances (Ansari, Sharma et Soni 2011). Le « *data mining* médical » a un fort potentiel dans la tâche d'exploration de patrons cachés dans les jeux de données du domaine médical (Ansari, Sharma et Soni 2011). Cependant, il se nécessite de plus de recherche afin de bien comprendre l'impact clinique de ces travaux.

Grâce à la récente augmentation de la disponibilité des données cliniques, les chercheurs ont implémenté différentes méthodes d'IA sur un large éventail de tâches cliniques, dès l'identification et le diagnostic à la prédiction (Wiens et Shenoy 2017).

Cela ne veut pas dire que les médecins seront remplacés par des machines

dans un futur proche, mais l'IA peut certainement aider et soutenir les médecins afin de prendre des meilleurs décisions ou même remplacer le jugement humain dans certains domaines d'activités de la santé, par exemple, la radiologie (Jiang et al. 2017).

Même si la littérature manifeste que l'IA aboutira vers un changement du paradigme de la santé grâce à la disponibilité croissante de données cliniques et le progrès rapide de techniques d'analytique (Jiang et al. 2017), il faut toujours prendre ces déclarations avec précaution.

### 2.2.1 Motivations

Les bénéfices de l'IA appliquée au domaine de la santé sont nombreux. En fait, ses avantages ont été longuement débattues dans la littérature médicale. L'IA peut utiliser des algorithmes sophistiqués pour « apprendre » des *features* à partir d'un grand volume de données de santé pour ensuite utiliser les connaissances obtenues pour soutenir la pratique clinique. Elle peut également se servir des capacités d'apprentissage et auto-correction avec l'objectif d'améliorer sa précision basée sur le *feedback*. Les systèmes d'IA peuvent aussi aider les médecins en fournissant des informations médicales actualisées extraites des revues, des manuels, et des pratiques cliniques afin de renseigner à propos des soins les plus adéquats pour un patient en particulier. En outre, ils peuvent contribuer à réduire les erreurs de diagnostic et de thérapie considérées comme inévitables dans la pratique clinique humaine (Jiang et al. 2017). Finalement, les systèmes d'IA extraient l'information utile issue d'une large population de patients pour aider à réaliser des inférences en temps réel des risques sanitaires et des résultats de prédiction.

De fait, la littérature médicale s'intéresse notamment à la prédiction dans le domaine de la santé. Quelle sera la souche de grippe prédominante dans la prochaine saison grippale ? Combien de vaccins contre la grippe est-ce qu'il faut préparer pour répondre aux demandes ? (Crown 2015) Par ailleurs, à partir d'un *dataset* de patients qui décrit leurs caractéristiques démographiques et les détails d'admission, on pourrait essayer de prédire le résultat spécifique de

la réadmission après 30 jours (Wiens et Shenoy 2017).

### 2.2.2 Les données cliniques

Avant que les systèmes d'IA puissent être mis en place dans les applications de santé, elles nécessitent être entraînés avec des données générées à partir des activités cliniques comme le dépistage, le diagnostic ou l'affectation du traitement (Jiang et al. 2017). Ces données cliniques peuvent être des données démographiques, des notes cliniques, des enregistrements électroniques de dispositifs médicaux, des examens physiques ou des laboratoires cliniques et images. Plus particulièrement, dans la phase de diagnostic, une grande partie de la littérature d'IA analyse des données issues d'imagerie diagnostique, du dépistage génétique et d'électrodiagnostic médical. Par exemple, Li *et al.* (Jiang et al. 2017) ont étudié les anomalies des expressions génétiques dans des longues ARN non-codantes pour diagnostiquer le cancer gastrique.

De plus, les notes d'examens physiques et les résultats des analyses de laboratoire comprennent les deux autres sources de données les plus importants. Il faut distinguer entre images, données génétiques et données électrophysiologiques (EP), puisque ces dernières contiennent des grandes parties de texte narrative non-structuré, comme les notes cliniques, qui ne sont pas directement analysables. En conséquence, les applications d'IA se concentrent d'abord dans la conversion de textes non-structurés en rapports médicaux électroniques (EMR) lisibles par une machine.

Le graphique 2.1 montre les types de données cliniques considérés par la littérature de l'IA. Cette comparaison est obtenue à travers la recherche de techniques de diagnostic dans la littérature d'IA dans la base de données de PubMed (Jiang et al. 2017).

Par ailleurs, dans le passé, les chercheurs visaient généralement à apprendre des modèles qui puissent généraliser sur tous les hôpitaux ou établissements de santé. Ces modèles réussissaient en moyenne mais résultaient peu performants lorsqu'ils étaient appliqués à d'autres institutions spécifiques. Cette limitation est due aux différences dans la façon dont les données sont recueillies et en-

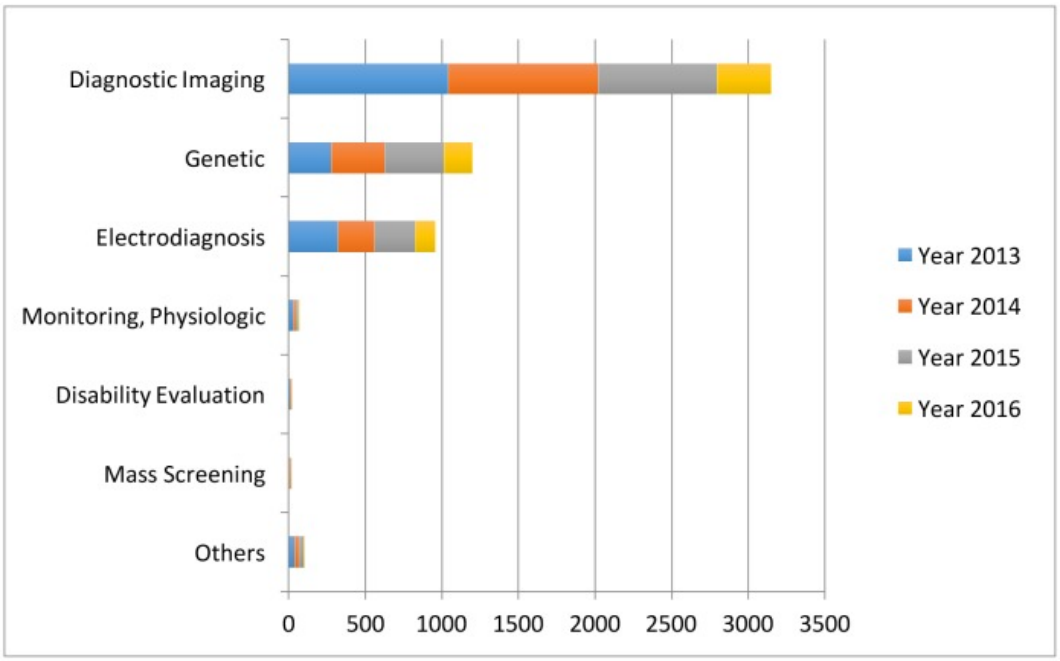



FIGURE 2.1 – Les données cliniques dans la littérature de l'intelligence artificielle

registrées (WIENS et SHENOY 2017). Il faudrait mieux chercher des méthodes généralisables que puissent être utilisées pour générer des modèles spécifiques pour les institutions, plutôt que chercher des modèles qui généralisent dans tous les hôpitaux. De cette manière, les institutions pourraient entraîner des modèles spécifiques à leurs données issues de leurs pratiques et leurs populations de patients.

### 2.2.3 Applications

Même si la littérature d'IA dans le domaine de la santé est davantage plus riche, la recherche se concentre principalement autour quelques types de maladies : le cancer, les maladies du système nerveux, et les maladies cardiovasculaires (JIANG et al. 2017). Pour le cancer, Esteva *et al.* (JIANG et al. 2017) ont analysé des images cliniques pour identifier sous-types de cancer de peau. Dans le champ de la neurologie, Bouton *et al.* (2016) ont développé un système d'IA pour aider aux patients de tétraplégie récupérer le contrôle du mouvement (JIANG et al. 2017). Finalement, en neurologie, Dilsziann et Siegel (JIANG et al.

2017) ont abordé l'éventuelle application de l'IA pour diagnostiquer les maladies cardiaques à travers des images cardiaques.

La concentration autour ces trois maladies n'est pas complètement inattendue. Tous ces trois maladies représentent les principales causes de décès ; pour cette raison, le diagnostic précoce est essentiel afin de prévenir la détérioration de l'état de santé des patients. D'ailleurs, le diagnostic précoce peut être éventuellement atteint grâce à l'amélioration de procédures d'analyses d'images, génétiques, électrophysiologiques ou de rayonnement électromagnétique, étant cette tâche le point fort de l'IA (Jiang et al. 2017).

En plus de ces trois domaines de spécialité, l'IA a été aussi bien implémentée à d'autres maladies. Deux exemples très récents sont Long *et al.* (Jiang et al. 2017), qui ont analysé les données des images oculaires pour diagnostiquer les cataractes héréditaires, et Gulshan *et al.* (2016), qui ont détecté la rétinopathie diabétique à travers des images de fonds rétiniens (Jiang et al. 2017).

Comme l'on a déjà exposé, les maladies cardiaques représentent le type de maladies le plus exploré par l'IA du fait de son importance en nombre de décès. En effet, l'Organisation Mondiale de la Santé a estimé que 12 millions de morts ont lieu mondialement chaque année dû aux maladies cardiaques (Ansari, Sharma et Soni 2011). Un système intelligent de prédiction de maladies cardiaques (Intelligent Heart Disease Prediction System, IHDPS) proposé par Sellappan Palaniappan *et al.* (Dangare et Apte 2012) a été développé à partir de 909 rapports avec 15 attributs (facteurs) obtenus de la base de données Cleveland Heart Disease. Naïve Bayes semble être le plus efficace car il a le pourcentage le plus élevé de prédictions correctes (86,53%) pour les patients ayant une maladie cardiaque. Par contre, les arbres de décision ont des meilleurs résultats (89%) lorsqu'on prédit les patients qui n'ont pas une maladie cardiaque (Ansari, Sharma et Soni 2011). Un autre système intelligent pour la prédiction des maladies cardiaques a été développé avec la base de données de Cleveland Heart Disease (Dangare et Apte 2012). Cette fois-ci, les auteurs ont ajouté deux attributs de plus pour obtenir des résultats plus précis : obesité et tabaquisme, considérés comme deux facteurs de grande im-

portance pour les maladies cardiaques. En conclusion, les résultats montraient que les réseaux de neurones offrent des résultats plus précis en comparaison avec les arbres de décision et Naïve Bayes.

La tâche principale des sciences médicales est la prévention et le diagnostic de maladies. Dans notre projet, l'accent sera mis sur la première tâche, la prévention de l'amyloïdose cardiaque.

### 2.2.4 Méthodes

Les systèmes d'IA peuvent se classer dans les suivantes catégories : les techniques de *machine learning*, les techniques plus récentes de *deep learning* et les méthodes de *natural language processing* (TAL).

Nous intégrons ci-dessus (figure 2.2) un schéma proposé dans l'article *Artificial Intelligence in healthcare : past, present and future* (Jiang et al. 2017), qui montre le parcours du processus depuis la génération, suivi de l'enrichissement avec les données obtenues après la partie du TAL, puis l'analyse de *machine learning* jusqu'à la prise d'une décision médicale.

Comme on pouvait s'y atteindre, la littérature liée à l'implémentation du *machine learning* et du TAL dans le domaine de la santé est énormément large, mais quelles sont leurs utilités et avantages en santé ? En premier lieu, mieux prévenir et prendre en charge les maladies. Les données récoltées à long terme sur des larges populations permettent d'identifier des facteurs de risque pour certaines maladies comme le cancer, le diabète, l'asthme, etc. En outre, le *big data*, le *data analytics* et le TAL permettent le développement de systèmes d'aide au diagnostic et d'outils qui rendent possible la personnalisation de traitements. Entre autres, Watson d'IBM analyse en quelques minutes le résultat de séquençage génomique de patients atteints de cancer, compare les données obtenues à celles déjà disponibles, et propose ainsi une stratégie thérapeutique personnalisée. Ils permettent également de vérifier l'efficacité d'un traitement. Par exemple, on peut contrôler qu'une vaccination a bien fonctionné au bout d'une heure seulement à partir d'un goutte de sang. D'ailleurs, si l'on dispose de nombreuses informations sur l'état de santé des individus dans une région

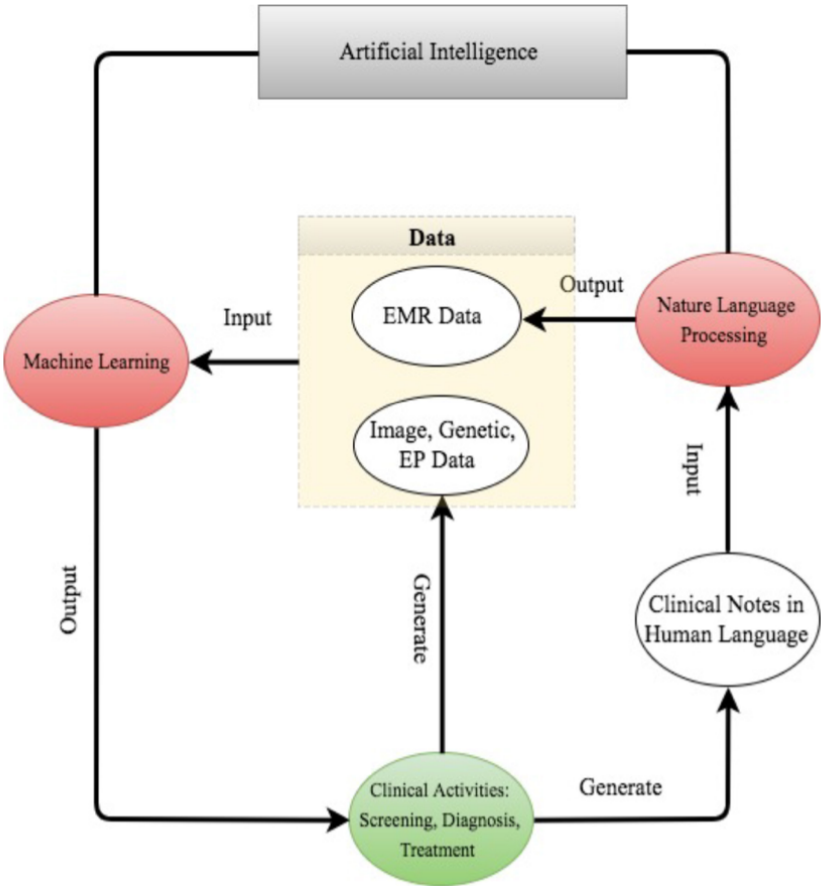



FIGURE 2.2 – Schéma depuis la génération de données cliniques jusqu'à l'enrichissement du traitement automatique des langues et l'analyse de données avec du machine learning.

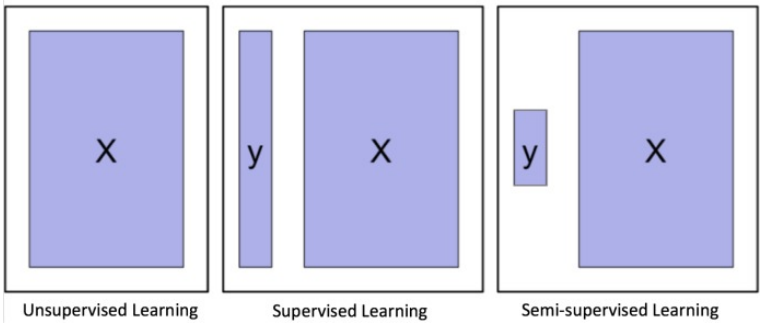



FIGURE 2.3 – Illustration graphique de l'apprentissage supervisé, l'apprentissage non-supervisé et l'apprentissage semi-supervisé

donnée, on pourrait repérer l'élévation de l'incidence de maladies ou de comportements à risque, et d'alerter les autorités sanitaires rapidement (CHARLET et al. 2018).

### Apprentissage automatique et deep learning

L'apprentissage automatique construit des algorithmes qui analysent les données pour en extraire des *features*. Les *inputs* de ces algorithmes peuvent être les traits des patients ou des résultats médicaux. Les traits de patients incluent ce qu'on appelle « *baseline data* », comme par exemple l'âge, le genre, les antécédents, et d'autres informations comme les résultats d'examens, des symptômes cliniques, médication, etc.

Les algorithmes de *machine learning* peuvent se diviser en même temps en deux groupes, en fonction du fait qu'on introduit les résultats (*outputs*) ou pas : *supervised learning* et *unsupervised learning*. L'apprentissage non-supervisé est notamment utilisé pour l'extraction de *features*, tandis que l'apprentissage supervisé est plus approprié pour la modélisation prédictive à travers la création de « relations » entre les traits des patients (*inputs*) et le résultat attendu (*output*). Plus récemment, l'apprentissage semi-supervisé a été proposé comme une méthode hybride entre l'apprentissage supervisé et l'apprentissage non-supervisé, et qui est approprié dans les cas où certains sujets manquent d'étiquette ou d'*output* (JIANG et al. 2017).

Le clustering et le PCA (*principal component analysis*) sont les princi-

pales méthodes d'apprentissage non supervisé. Les algorithmes de clustering rassemblent les sujets avec des traits similaires dans un même groupe nommé *cluster*, c'est-à-dire, l'algorithme retourne des *cluster labels* afin de maximiser et minimiser la similarité des patients entre les *clusters*. Le PCA tente de réduire la dimensionnalité d'un ensemble de données tout en conservant autant d'informations que possible.

En revanche, l'apprentissage supervisé considère les résultats de sujets ainsi que les traits, et suit un processus d'apprentissage afin de pouvoir déterminer les meilleurs *outputs* associés aux *inputs* qui sont plus proches des résultats en moyenne. Le résultat peut être, par exemple, la probabilité d'un événement clinique, la valeur attendue du niveau d'une maladie ou le temps de survie prévu.

En comparaison avec l'apprentissage non-supervisé, l'apprentissage supervisé fournit des résultats cliniquement plus pertinents ; c'est la raison pour laquelle la plupart d'applications d'IA en médecine réposent sur ce type de méthodes. Cependant, l'apprentissage non supervisé peut faire partie de la phase de pré-traitements afin de réduire la dimensionnalité ou identifier des groupes, ce que, par conséquence, fait que la suite, la phase d'apprentissage supervisé, soit plus efficace.

La régression linéaire, la régression logistique, naïve Bayes, les arbres de décision, K-means, random forest, SVM et les réseaux de neurones sont des exemples d'algorithmes d'apprentissage automatique. Le graphique 2.4 montre la popularité des différents techniques d'apprentissage automatique dans le domaine de la santé, ce qui révèle que les plus populaires sont le SVM et les réseaux de neurones.

D'autre part, pour que les décisions proposées par l'algorithme soient acceptables ou légitimes, voire pour être écartées car jugées non pertinentes, celles-ci doivent pouvoir être comprises, et donc, explicables. Cependant, certaines des approches numériques dont on a parlé précédemment se ressemblent à une boîte noire, incapable de justifier ses décisions : personne sait que fait l'algorithme ou pourquoi. Et pourtant, dans les cas de prédiction de maladies,

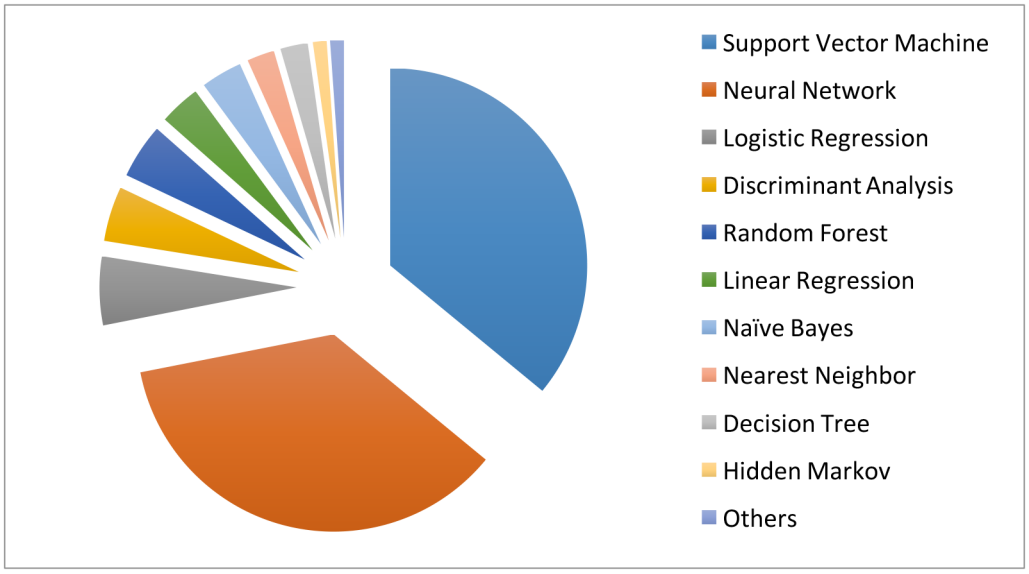



FIGURE 2.4 – Algorithmes de machine learning utilisés dans la littérature médicale. Les résultats sont obtenus après la recherche des algorithmes de machine learning employés au sein du domaine de la santé dans PubMed (JIANG et al. 2017)

il serait énormément utile et intéressant de connaître les variables ou les motifs sur lesquels s'appuie l'algorithme pour fournir le modèle. C'est pour cela que les projets tentent aujourd'hui de combiner les deux approches, symbolique et numérique, le but étant de bénéficier à la fois du raisonnement de l'un et des performances de l'autre (CHARLET et al. 2018).

Aujourd'hui, les ANN (artificial neural network) sont devenus l'outil le plus utilisé pour le diagnostic de maladies. La classification est un outil important pour l'aide à la décision dans les diagnostics médicaux. Les réseaux de neurones fournissent un outil performant qui aide les médecins à analyser, modéliser et comprendre les données médicales complexes. Dans des nombreux cas, les réseaux de neurones sont utilisés dans le cadre de la médecine pour des problèmes de classification ; c'est-à-dire, la tâche consiste à assigner un patient à un petit groupe de classes à partir de caractéristiques mesurées. Lors du processus de diagnostic, les médecins généralement utilisent leurs connaissances, expériences et capacités pour diagnostiquer une maladie. Le processus de diagnostic commence habituellement avec les plaintes du patient, le docteur ensuite essaie

d'apprendre plus à l'occasion de la consultation du patient, ainsi qu'à travers l'évaluation de certains indicateurs comme la pression artérielle ou la température corporelle. Le diagnostic est donc déterminé en prenant en considération l'ensemble de conditions du patient. Selon le diagnostic, le traitement est défini (KHAN et al. 2001).

Par exemple, (KHAN et al. 2001) ont implémenté les réseaux de neurones pour le diagnostic médical et la détection précoce de maladies, particulièrement la néphrite aigüe et les maladies cardiaques. Chaque patient était classé en deux catégories : infecté ou non infecté.

Mirtskhulava *et al.* (MIRTSKHULAVA et al. 2015) ont utilisé les réseaux de neurones pour le diagnostic d'infarctus. Dans leur analyse, les variables d'entrée $X_{i1}, \ldots, X_{ip}$ sont des symptômes liés à l'infarctus (étant p = 16), comme la confusion aiguë, les problèmes de mobilité, la paresthésie, etc. Le résultat $Y_1$ est binaire : $Y_i = 0$ ou $Y_i = 1$, et indique si le patient i a ou pas un infarctus. Le paramètre d'intérêt obtenu est la probabilité d'infarctus, $a_i$.

D'autres techniques similaires ont été implémentées pour diagnostiquer le cancer, comme (KHAN et al. 2001), qui ont utilisé les PCs estimés à partir de 6567 gènes. Les résultats sont les catégories de tumeur. (DHEEBA, SINGH et TAMIL 2014) utilisent aussi les réseaux de neurones pour prédire le cancer du sein, avec des informations d'images tomographiques comme *inputs* et des indicateurs de tumeur comme *output*. (HIRSCHAUER, ADELI et BUFORD 2015) utilisent un modèle de réseau de neurones plus sophistiqué pour diagnostiquer la maladie de Parkinson en fonction des symptômes motrices et non-motrices et des neuroimages.

Dans le cadre du deep learning, les réseaux de neurones récurrents peuvent être employés dans les cas où il existe une relation entre les variables considérées comme *inputs* et d'autres variables qui peuvent être prédites (*outputs*). Les avantages les plus importantes de l'utilisation des ANN c'est que ce type de système résout des problèmes qui sont trop complexes pour les technologies traditionnelles, qui n'ont pas une solution algorithmique ou la solution est trop complexe pour l'utiliser. Ces caractéristiques apparaissent souvent

en médecine. Les ANN ont été utilisés avec succès dans plusieurs domaines de la médecine comme des systèmes de diagnostic, des analyses biomédicales, des analyses d'images et le développement de médicaments. Les réseaux de neurones sont une alternative puissante et servent comme complément aux techniques statistiques lorsque les données sont multivariées ou multidimensionnelles et présentent un degré élevé d'interdépendance entre les facteurs, les données ont du bruit ou sont incomplètes, ou quand il y a plusieurs hypothèses à poursuivre ou on a besoin de réussir des taux élevées (KHAN et al. 2001).

L'étude réalisé par (JIANG et al. 2017) a l'intention de fournir un sondage des techniques actuelles en recherche de connaissances dans des bases de données en utilisant des techniques de *data mining* que sont employées aujourd'hui dans le domaine de la recherche médicale, particulièrement dans la prédiction de maladies cardiaques. Après avoir effectué plusieurs expériences afin de comparer la performance de techniques de *data mining* prédictives dans le même dataset, les conclusions indiquent que les arbres de décision ont des meilleurs résultats et que parfois la classification bayésienne obtient une précision similaire à celle de l'arbre de décision. Par contre, d'autres méthodes prédictives comme KNN, les réseaux de neurones ou la classification basée sur le clustering n'obtiennent pas si des bons résultats. De plus, la précision des arbres de décision et de la classification bayésienne améliore davantage après l'application des algorithmes génétiques qui réduisent la taille actuelle des données afin d'obtenir un sous-ensemble optimal d'attributs suffisants pour la prédiction de la maladie cardiaque. Le résultat du processus détermine si le patient souffre ou non une maladie cardiaque. Pour cela, ils se sont muni de 3000 observations avec 13 attributs différents : sexe, type de douleur thoracique, glycémie à jeun, les résultats electrocardiographiques de repos (restecg), angine de poitrine induite par l'exercice, sous-décalage du segment ST après une preuve d'effort, nombre de vaisseaux principaux colorés par fluoroscopie, examen du coeur au thallium, pression artérielle, cholestérol sérique, fréquence cardiaque maximale réelle, abaissement de ST induit par l'exercice et âge.

### Traitement automatique des langues

Les images, les études électro-physiologiques, et les données génétiques sont compréhensibles par une machine, de manière que les algorithmes de *machine learning* peuvent être « directement » exécutées une fois que les pré-traitements ont été effectués. Cependant, une grande quantité des informations cliniques existent sous forme de texte, comme les examens médicaux, les analyses médicales de laboratoire, les notes opératoires des patients ou les résumés de sortie. Ce sont des données non-structurées et incompréhensibles pour les programmes d'ordinateur.

Dans ce contexte, le TAL vise à extraire l'information utile qui se trouve dans le texte. À partir du traitement du texte, le TAL identifie quelques mots clés concernant des maladies contenues dans les notes cliniques fondées sur des bases de données historiques. Ensuite, il faut sélectionner les mots clés qui enrichissent les données structurées et peuvent aider à la tâche visée.

Le cas de (M., WW. et D. 2000) illustre bien l'utilité de TAL dans ce type de projets. Ils introduisent le TAL pour identifier les concepts importants pour des systèmes d'aide à la décision, c'est-à-dire, les rapports de radiographies thoraciques peuvent aider les systèmes d'assistance antibiotique pour alerter les médecins de la nécessité d'une thérapie anti-infectieuse.

D'autant plus, les pipelines de TAL peuvent aider aux diagnostics de maladies. Prenons le cas de (Castro, Dligach et Finan 2017), qui ont identifié 14 variables associées aux anévrismes cérébraux grâce à l'implémentation du TAL dans les observations cliniques. Les variables obtenues sont employées avec succès pour la classification des patients dits « normaux » et les patients avec la maladie cérébrale, avec des pourcentages de précision de 95% et 86% respectivement pour les preuves d'apprentissage et évaluation. Finalement, (Afzal, Sohn et Abram 2017) ont implémenté le TAL pour extraire des mots clés liés à une maladie qui affecte les artères périphériques à partir des observations cliniques. Ensuite, ces mots clés sont utilisés pour classifier les patients qui ont ou n'ont pas cette maladie. Ce modèle atteint un taux de précision de 90%.

Nous pouvons également mentionner un projet développé en 2010 par les

chercheurs du LIMICS, qui ont conçu un logiciel de traitement automatique des langues dans le cadre du projet Lerudi (*Lecture rapide en urgence du dossier informatique du patient*). Grâce à une ontologie des urgences insérée dans un moteur de fouille du dossier médical du passé, celui-ci est capable de répondre aux besoins des urgentistes afin de les aider à prendre une décision dans quelques minutes (Charlet et al. 2018).

Ce groupe de recherche, le LIMICS, en collaboration avec l'hôpital Trousseau, ont conçu un système d'aide à la décision dans l'analyse d'échographies pour les grossesses extra-utérines (GEU), qui fournit un modèle centré sur les signes du domaine, avec les relations entre les signes des différents types de grossesse extra-utérine, les structures anatomiques et les éléments techniques.

Il faut également souligner le projet international Big data for better outcomes (BD4BO) (País 2018), qui fait partie de l'initiative européenne Innovative Medicines Initiative (IMI), qui a pour but d'utiliser les techniques de *big data* pour intégrer et réutiliser les données qui pourraient donner des réponses aux problèmes plus urgents qui posent certaines pathologies comme la maladie d'Alzheimer, le cancer de la prostate, ou les maladies cardiovasculaires.

### La prédiction de l'amyloïdose cardiaque

L'objectif général de la recherche menée par (Garg et al. 2016) est de développer un outil qu'identifie des possibles patients atteints de n'importe quelle maladie rare ou risque de l'avoir. Ils ont implémenté cette approche initialement avec l'amyloïdose cardiaque ; par contre, cette approche est extensible à toutes les maladies rares. Les principales caractéristiques de cette plate-forme sont la représentation de chaque patient comme un vecteur multidimensionnel et l'utilisation des algorithmes de *machine learning* basés sur des données disponibles concernant de maladies rares afin de les identifier automatiquement.

Ils ont procédé de la suivante façon (figure 2.5). Au début, ils ont créé un dataset de taille raisonnable avec des patients positifs et négatifs pour l'amyloïdose cardiaque à partir des essaies cliniques d'amyloïdose cardiaque dans le Northwestern Medicine. Un total de 73 patients d'amyloïdose cardiaque com-

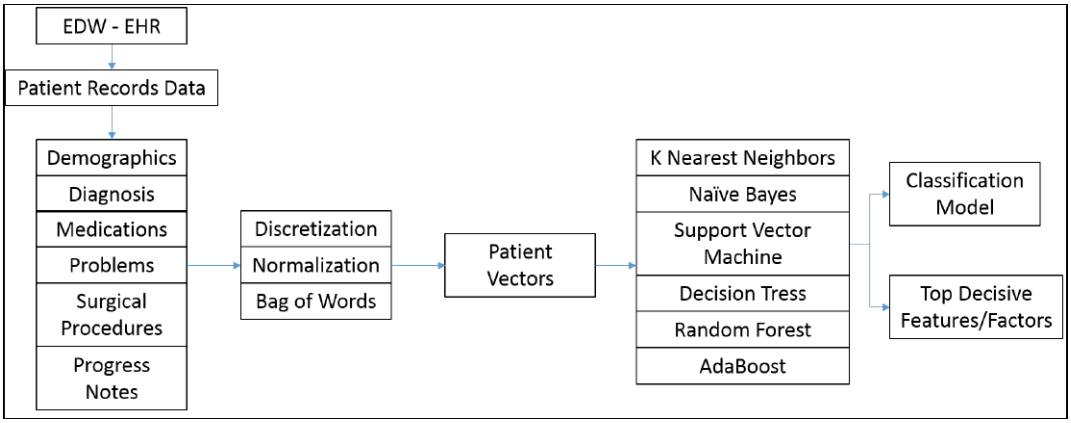



FIGURE 2.5 – Architecture de l'approche

posent les instances positives. Les instances négatifs sont formées par un échantillon aléatoire de 197 patients qui ont visité un cardiologue dans le Northwestern Medicine et qui n'ont pas été diagnostiqués avec amyloïdose cardiaque. Chaque patient était représenté à travers un vecteur par l'usage d'information recueillie des histoires cliniques (*EHR*) dans le Northwestern Medicine Enterprise Data warehouse (NMEDW). Les sections pertinentes ou types de *features* considerées dans l'étude étaient les caractéristiques démographiques (comme le genre, la race, l'ethnicité et l'âge), le diagnostic, la médication, les problèmes, les données chirurgicales et les notes d'évolution. Les valeurs de chaque type de *feature* ont été transformées de façon qu'elles puissent être utiles pour un classifieur de *machine learning*. Par exemple, si un patient est diagnostiqué avec anémie, ils ont créé un vecteur « Anémie » et ils ont assigné la valeur 1 pour ces patients et 0 pour le reste de patients. En ce qui concerne les notes d'évolution, ils ont implementé le modèle Bag of Words. Les notes de patients étaient premièrement pre-processées (mis en minuscules, enlèvement des stopwords, normalisation de valeurs numériques), et ensuite tokenisation en unigrammes et bigrammes (GARG et al. 2016)

Les algorithmes de classification de *machine learning* envisagés par l'étude étaient les KNN, les SVM, les arbres de décision, Random Forest, AdaBoost et Naïve Bayes. En plus de ces algorithmes, ils ont implémenté des méthodes de sélection de *features* afin d'améliorer les modèles de classification. Les *features*

avec peu de variance sont supprimés. Ils ont également appliqué la normalisation L1 pour identifier les « top features » qui jouent un rôle décisif dans la classification du patient dans chaque catégorie (GARG et al. 2016).

Finalement, pour évaluer le système, ils ont réalisé la validation croisée et calculé la F-mesure :

| **Algorithm** | **Configuration** | **Average 10 cross validation F1 score** |
|---|---|---|
| K Nearest Neighbor | K – 1, criterion – 'uniform' | 0.42 |
| Naïve Bayes | | 0.71 |
| SVM | Kernel – rbf, C -100, gamma - 0.001 | 0.93 |
| Decision Tree | Criterion – gini | 0.94 |
| Random Forest | N-estimator - 15, Criterion – gini | 0.94 |
| **AdaBoost** | **N-estimator - 15** | **0.97** |

FIGURE 2.6 – Paramètres optimaux pour chaque modèle et score F1

Les KNN et NB fournissent une F-mesure très baisse en comparaison avec d'autres algorithmes plus complexes puisqu'ils sont sensibles aux *outliers* et ne performent pas une bonne sélection de *features*. Ces algorithmes accordent la même importance à tous les *features*, ce qui conduit à une performance pauvre. Par contre, SVM, DT et les méthodes d'ensemble offrent une F-mesure élevée à cause de leur complexité et la capacité intrinsèque de performer une sélection de *features* adéquate (GARG et al. 2016).

Après réaliser la sélection de *features*, ils ont obtenu (GARG et al. 2016) une liste de ceux étant les plus pertinants pour la détection de l'amyloïdose cardiaque (figure 2.7). Les patients âgés entre 70-90 risquent davantage d'être atteints que les patients d'autres âges. En plus, le fait que d'autres diagnostiques cardiaques comme l'insuffisance cardiaque congestive et la douleur thoracique soient présents dans cette liste montre que l'amyloïdose cardiaque a une comorbidité associée à d'autres problèmes cardiaques.

| |
|---|
| Age range of 80-89 |
| Age range of 70-79 |
| Diagnosis of cardiac arrest |
| Diagnosis of chest pain |
| Diagnosis of congestive heart failure |
| Diagnosis of hypertension |
| Diagnosis of prim open angle glaucoma |
| Diagnosis of shoulder arthritis |
| Medication of Doxercalciferol |
| Medication of Loratadine 10mg |
| Medication of Levothyroxine Sodium 25mcg |
| Medication of 008278-ZZceFAZolin Sodium |
| Medication of Sodium Chloride 0.9% |

FIGURE 2.7 – Top features ou facteurs de la sélection de features

## 2.3 Anonymisation

La problématique de l'anonymisation de documents cliniques émerge dès que ces documents sont utilisés pour d'autres buts que les originaux, comme par exemple que ce soient rassemblés afin de créer des corpus pour faire de la recherche, ou utilisés dans des publications scientifiques (GROUIN 2013). Quoi que ce soit, les informations personnelles contenues dans ces documents ne doivent pas être jamais accessibles par d'autres personnes que ne soient pas le patient ou ses médecins. À cause de cela, un processus d'anonymisation doit être mis en oeuvre, afin que l'identification du patient dont parle le document soit impossible de réaliser.

L'article (ELUSTONDO et al. 2011) réfléchit à propos de toutes les situations possibles que l'on peut rencontrer lors de l'emploi de documents cliniques avec des buts de recherche. Lorsque les données sont anonymes, elles peuvent être utilisées et cédées sans le consentement des individus. Étant donné qu'on peut pas associer l'information avec aucune personne, il n'y a pas des considérations

éthiques à tenir en compte et le traitement de ce type de données reste dehors des exigences établies par la Ley Orgánica 15/1999 de Protección de datos de caracter personal (Protection de données de caractère personnelle, LOPD) (Estado 1999) ou le règlement européen (Conseil de l'Union Européenne 2016). Cependant, lorsque les données contenues dans les dossiers cliniques ne sont pas anonymes, il faut réaliser un processus d'anonymisation qui permette d'obtenir les données cliniques d'intérêt, en les séparant des données d'identification. La loi 41/2002, régulatrice de l'autonomie du patient et des droits et obligations dans le domaine de l'information et documentation clinique, établie que « toute personne a le droit à que le caractère confidentiel de données concernant leur santé soit respecté, que personne puisse accéder à eux sans une autorisation préalable » (article 7.1) et que « l'accès au dossier clinique à des fins de recherche oblige à préserver les données d'identification personnelle du patient, de façon que l'anonymat reste assuré ».

Dans la législation espagnole, même si l'emploi de données à caractère personnel à des fins de recherche est légitime, tout patient a le droit à la confidentialité.

Étant donné la difficulté d'accès aux données cliniques, puisque comme expose (Grouin 2013) le seul fait d'accéder à ce type de documents suppose le non-respect de la vie privée du patient, nous nous sommes posé la suivante question : comment est-ce qu'on va anonymiser des documents auxquels nous n'avons pas d'accès jusqu'à ce qu'ils soient anonymisés ? Une des solutions possibles serait de mettre à disposition des corpus médicaux composés de données réelles mais qui ne contiennent pas des données personnelles. Si, au moins, on réussissait à « masquer » les données personnelles dites sensibles, nous pourrions développer des méthodes d'anonymisation sur des textes réels mais qui n'ont pas d'informations fortement identifiables. Celle-ci était l'approche suivie par le projet Akenaton (Grouin 2013). Ils ont effectué une première passe d'anonymisation reposant sur des méthodes assez simples et mises en oeuvre sur la source elle-même. De cette manière, ils ont supprimé presque toute la totalité des informations identifiables fondamentales : le nom, le prénom et

la date de naissance du patient. Ces informations étaient aisément accessibles puisqu'elles figuraient dans la partie structurée des dossiers des patients et elles avaient été extraites des fichiers de métadonnées associées à chaque document textuel. Ces métadonnées contenaient les données du patient (nom, nom marital, prénom, date de naissance et sexe) et les données de provenance du document. L'algorithme ne devait que remplacer le texte mentionnant ces informations relatives au patient par des balises.

### 2.3.1 Méthodes

#### La désidentification et la reconnaissance d'entités nommées

La tâche de désidentification est souvent envisagée comme un problème de reconnaissance d'entités nommées (REN). Cette problématique du TAL a donné lieu à des systèmes qui reposent soit sur des approches orientées connaissances soit sur des approches guidées par les données (NOUVEL et al. 2013). Le premier type de techniques regroupe celles basées sur des ensembles de règles grammaticales et syntaxiques construites manuellement pour chaque type d'entité nommée ou des transducteurs. Celles-ci ont généralement une grande précision mais nécessitent un coût de développement important. Le deuxième type regroupe les techniques basées sur des modèles statistiques (e.g Modèle de Markov Caché (MMC), Maximum Entropy Model (MEM) ou encore Conditional Random Field (CRF)), qui sont entraînés avec un ensemble de textes dans lesquels les entités nommées à détecter ont déjà été identifiées et classées. Ces dernières méthodes d'apprentissage automatique réussissent de bonnes performances avec un coût d'entrée pas si important comme dans le premier cas, par contre elles nécessitent d'un corpus d'apprentissage déjà annoté. Ils existent différents types de modèles probabilistes qui peuvent être employés pour la REN. Cette tâche est formalisée comme un problème d'étiquetage de séquences dont l'objectif est de trouver la meilleure séquence d'étiquettes pour une suite d'observations données. Cela se modélise comme suit : étant donnée une séquence d'observations (de mots dans notre cas) $X = x_1, x_2, ..., x_n$, il faut trouver la sé-

quence d'étiquettes E = $e_1$, $e_2$, ..., $e_n$ qui maximise la probabilité conditionnelle (HATMI 2015). Nous pouvons distinguer deux familles de modèles probabilistes pour résoudre ce problème, les modèles génératifs et les modèles discriminants.

Les modèles génératifs, comme les Modèles de Markov Cachés (HMM), s'intéressent à maximiser la probabilité jointe de la paire ($X$, $E$). À travers l'application de la règle de Bayes, cette probabilité se décompose en deux : la probabilité conditionnelle de génération des observations P ($X$, $E$) et la probabilité de la séquence d'étiquettes P ($E$).

Par contre, les modèles discriminants considèrent la probabilité conditionnelle. Cela permet de détendre les hypothèses faites sur l'indépendance des observations et prendre une décision globale avec l'objectif de prédire une étiquette. En outre, ils ne nécessitent pas l'estimation de la probabilité de génération des observations. Les champs aléatoires conditionnels (CRF pour *Conditional Random Fields*) se sont imposés comme l'un des modèles discriminants le plus performant pour la tâche d'étiquetage de séquences. Les CRF sont des modèles graphiques non dirigés qui permettent de calculer la probabilité conditionnelle d'une séquence d'étiquettes $E = e_1, e_2, \ldots, e_n$ étant donné une séquence de mots X = $x_1$, $x_2$, ..., $x_n$. Une fois le modèle d'annotation appris, l'application des CRF à des nouvelles données revient alors à trouver la séquence d'étiquettes la plus probable étant donnée une séquence de mots en entrée.

### La désidentification de PHI

Ils existent différentes méthodes possibles en traitement automatique des langues pour effectuer l'anonymisation de documents cliniques sous forme écrite. Meystre *et al.* (MEYSTRE et al. 2010) ont dressé un bilan des méthodes de désidentification fondé sur 200 publications. Une grande partie de ces publications portaient sur des données structurées plutôt que sur de texte narratif. Ils mettent en évidence deux méthodes principales pour la désidentification : le « *pattern matching* » et le *machine learning*. Beaucoup de systèmes combinent les deux approches pour les différents types de PHI, par contre la majo-

rité n'utilise pas le *machine learning* et compte uniquement sur les patrons, les règles et les dictionnaires. Ces ressources sont élaborées manuellement, nécessitent de nombreux mois de travail par des spécialistes du domaine et se caractérisent par une possibilité de généralisation limitée. Presque tout le filtrage par motifs est implémenté à travers des expressions régulières. Les dictionnaires sont construits à partir de plusieurs sources, et visent à distinguer ce qui relève des informations identifiantes de celles qui n'en sont pas. La première classe de dictionnaires énumère des termes qui sont normalement considérés comme PHI, tels que noms propres, locations géographiques, noms d'établissements de soins de santé, et parfois même des noms de patients ou des professionnels de santé de l'établissement dans lequel le système a été développé. Ces listes ont été construites à partir des sources accessibles au public comme l'Indice de certificats de décès de la Sécurité Sociale, des lexiques de noms, ou des listes de villes, départements ou états des États-Unis. La deuxième contient des termes qui ne sont pas considérés habituellement comme des PHI, et incluent des termes générales en anglais ou des termes bio-médicaux. Les termes dites « générales » sont issus de sources publiques comme la liste de mots Atkinson ou le lexique Ispell, et ceux-ci qui appartiennent au domaine bio-médical proviennent pour la plupart du UMLS Metathesaurus (Unified Medical Language System) ou d'autres ressources terminologiques comme MeSH (Medical Subject Headings). Parmi les avantages des méthodes à base de règles et du filtrage par patrons pour la désidentification nous pouvons remarquer qu'ils nécessitent peu de données annotées pour l'entraînement et qu'ils peuvent être rapidement modifiés afin d'améliorer la performance grâce à l'ajout de règles, dictionnaires de termes ou d'expressions régulières. Par contre, les développeurs doivent élaborer des algorithmes très complexes dans l'objectif de tenir compte des différentes catégories de PHI, en plus de l'adaptation pour chaque jeu de données. Pour cette raison, la performance de la reconnaissance de patrons ne peut pas être généralisée à des différents jeux de données. Une autre contrainte de cette approche c'est que les développeurs doivent connaître tous les possibles patrons de PHI qui peuvent apparaître, comme des formats de

dates non attendus ou des patrons de lieux qui utilisent des abréviations non-standardisées. Quant aux systèmes qu'appliquent ces méthodes, nous pouvons remarquer certains mentionnés dans le bilan de Meystre *et al.* (Meystre et al. 2010).

Beckwith (Beckwith et al. 2006) a développé HMS Scrubber, un outil de désidentification implémenté sur des rapports pathologiques. Ceci utilise des dictionnaires de noms propres et des locations pour trouver les informations concernant le patient ainsi que plus de cinquante expressions régulières pour repérer des dates, numéros de téléphone, et numéros de la sécurité sociale dans le corps du rapport. Ce système atteint un rappel de 98% et une précision de 43%. Berman (Berman 2003) propose une méthode alternative aux systèmes traditionnels de désidentification qui consiste à extraire et retirer tous les mots du texte sauf ces appartenant à une liste de mots connus considérés comme non-PHI. Aucun résultat a été indiqué concernant les mesures d'évaluation.

Les applications les plus récentes ont tendance à reposer généralement sur des méthodes de *machine learning* pour classifier les mots comme des PHI ou non-PHI, et même dans différentes classes de PHI à l'intérieur du premier groupe. Les méthodes utilisées vont du Support Vector Machines à Conditional Random Fields, Arbres de Décision et Entropie Maximale. Ces algorithmes de *machine learning* ont besoin d'un large corpus textuel annoté pour l'étape d'entraînement, une ressource qui nécessite aussi de beaucoup de travail réalisé en particulier par des experts du domaine, même si la tâche d'annotation est souvent considérée plus simple que l'ingénierie de connaissances. Presque tous les systèmes basés sur de *machine learning* ajoutent des patrons pour extraire des *features* pour la classification, ou afin de détecter des types spécifiques de PHI qui ont tendance à se répéter, comme les numéros de téléphone ou de la sécurité sociale. Ces systèmes utilisent une diversité de *features* pour ses algorithmes comme des *features* lexicales, par exemple la casse, la ponctuation, les caractères numériques ou la morphologie du mot, des *features* syntactiques comme le *part of speech* ou des *features* sémantiques.

Les principaux avantages des méthodes de désidentification avec du *ma-*

*chine learning* reposent sur le fait qu'ils peuvent apprendre automatiquement à reconnaître des patrons complexes de PHI, et que les développeurs du système n'ont pas besoin de grandes connaissances en ce qui concerne les patrons de ce type d'informations. De plus, les systèmes basés sur de *machine learning* n'ont pas tendance à incrémenter en complexité et la vitesse de calcul ne ralentisse pas au fil du temps, comme c'est le cas des systèmes de reconnaissance de patrons lorsqu'ils sont adaptés progressivement à des nouveaux types de documents. Parmi les inconvénients, ils nécessitent d'une grande quantité de données annotées pour l'entraînement. Par ailleurs, il est parfois difficile de comprendre les erreurs produites, et, même si ces méthodes sont normalement plus généralisables, elles ont besoin de données annotées supplémentaires quand elles sont implémentées à un nouveau jeu de données. Nous pouvons également ressortir certains systèmes qui emploient ces méthodes de *machine learning* dans le bilan de méthodes de désidentification réalisé par Meystre *et al.* (2010). Tous ces systèmes que l'on va décrire ont été qualifiés comme ayant les meilleurs résultats dans le défi de désidentification i2b2.

Aramaki *et al.* (2006) suit une approche qui combine des *features* non-locales (longueur de la phrase, location au sein du document...) avec des *features* locales (mots voisins). La méthodologie suivie consiste à annoter manuellement tous les mots du jeu de données d'entraînement avec les étiquettes PHI ou non-PHI. Ensuite, l'algorithme Conditional Random Fields (CRF) apprend la relation de ces features et les étiquettes à partir du jeu d'entraînement annoté. Toutes les mesures d'évaluation atteignaient le 94% (Meystre et al. 2010). Le système HIDE (Health Information DE-identification) a été développé par Gardner *et al.* (Gardner et Xiong 2008) et traite l'extraction de PHI comme un problème de reconnaissance d'entités nommés (NER), en utilisant aussi les Conditional Random Fields (CRF) pour extraire les attributs identifiants et sensibles. Parmi les *features* de l'algorithme se trouvent le mot précèdent, le mot suivant, la capitalisation, la présence de caractères spéciaux, ou si le token est une chiffre. La précision est de 98,2% (Meystre et al. 2010). D'autre part, Guo *et al.* (2006) envisagent également le problème de désidenti-

fication comme une tâche de reconnaissance d'entités nommés (CUNNINGHAM et al. 2002), laquelle à son tour ils considèrent un problème de classification. Leur système utilise Support Vector Machines (SVM). L'équipe Guo a utilisé le système open source GATE, duquel ils ont employé le système d'extraction d'information ANNIE. Plusieurs *features* pour le repérage de dates, noms de docteurs, noms d'hôpitaux, âges et locations ont été ajouté au classifieur SVM afin de réussir un meilleur taux de rappel pour la détection des PHI. Ce système est un de peu qui n'utilise pas des expressions régulières dans leur méthodologie. Les mesures d'évaluation sont toutes au-dessus de 86%. Les arbres de décision ont été aussi utilisés par Szarbas *et al.* (SZARVAS, FARKAS et BUSA-FEKETE 2007), qui décrivent un système de désidentification qui emploie une approche de NER avec du *machine learning* pour l'identification de PHI dans des dossiers de congés. Ils ont appliqué une méthode d'apprentissage itérative basée sur des arbres de décision qui utilisent l'information contenue dans la partie structurée des rapports, normalement l'en-tête, pour améliorer la reconnaissance de PHI dans le corps du rapport (SZARVAS, FARKAS et KOCSOR 2006). Le système inclut un modèle de classification au niveau de mots avec un jeu de *features* qui inclut des caractéristiques orthographiques (capitalisation, longueur du mot...), informations à propos de la fréquence, information syntagmatique (classes des mots précédents et des suffixes communs avec le token cible), dictionnaires (noms, noms de localisations...) et information à propos du contexte (comme la position à l'intérieur de la phrase ou l'en-tête le plus proche). Il emploie aussi des expressions régulières pour identifier des patrons bien connus de PHI. L'évaluation de la détection de l'ensemble d'informations PHI a donné des taux de rappel, précision et F-mesure supérieurs au 96%. Taira *et al.* (TAIRA, BUI et KANGARLOO 2002) décrivent un système de désidentification qui utilise des modèles statistiques pour enlever les noms de patients dans les rapports médicaux pris des pratiques de pédiatrie en urologie.

L'algorithme s'appuie aussi bien sur des lexiques que sur d'information sémantique de contrainte pour assigner les probabilités qu'un mot soit un nom. Le système est désigné pour n'enlever que les noms de patients de rapports et

il n'est pas applicable à d'autres catégories de PHI. Le système Taira utilise un algorithme de Maximum Entropy basé sur les statistiques afin d'estimer la probabilité d'une référence à un nom d'un patient dans un contexte d'un ensemble de relations logiques prédéfinies. Le lexique est composé de plus de 64000 noms et prénoms, et les restrictions sémantiques apportent les conditions contextuelles que les mots candidats doivent accomplir. Ceci est basé sur l'hypothèse qu'il existe un lien étroit entre certaines classes de mots et certains concepts. Par exemple, le mot « *presented* » est fortement lié avec le concept d'un nom du patient comme dans « *John Smith presented to the clinic today* ». Ces contraintes ont été automatiquement déterminées à partir du corpus d'entraînement annoté.

En général, les méthodes basées sur des dictionnaires obtiennent des meilleurs résultats avec des PHI qui sont rarement présents dans les documents cliniques, mais ils sont plus difficiles à généraliser. Les méthodes de *machine learning* ont tendance à mieux fonctionner, particulièrement avec des PHI qui ne sont pas présents dans les dictionnaires employés. Dans le défi de désidentification de i2b2, les systèmes qui avaient les meilleurs résultats étaient ceux basés sur de *machine learning* avec des *features* des expressions régulières pour toutes les catégories de PHI.

### 2.3.2 Applications

L'Informatics for Integrating Biology and the Bedside (i2b2) (Murphy et al. 2007) est une des initiatives du NIH Roadmap National Centers for Biomedical Computing, dont l'objectif principal est de fournir aux chercheurs du domaine médical les outils nécessaires pour rassembler et gérer les données concernant les projets de recherche médicale. Cette interface i2b2 intègre un module d'anonymisation qui retourne l'ensemble de données dépourvus de données personnelles, ce qui met en évidence l'importance de cette tâche dans la gestion de données cliniques.

L'HIPAA permet que les identifiants soient enlevés afin de créer un dataset « propre », ce que l'institution appelle *Clinical Data Set*. Ce c'est que fait

une des cellules de l'infrastructure. Cette cellule contient le « code book » qui associe les identifiants réels des patients à des chiffres aléatoires (Murphy et al. 2007). La plate-forme Hive de i2b2 est construite sur l'hypothèse que le partage, la diffusion et la mise en commun doivent être l'intention finale de toute recherche.

Les notes cliniques sont ajoutées à travers la cellule de Gestion de l'identité. Les noms, prénoms et chiffres des dossiers médicaux sont débarrassés et retenus par cette cellule. À chaque fois qu'on ajoute des informations provenants de dossiers médicaux au Clinical Research Chart (CRC), celles-ci doivent toujours passer par cette cellule. En plus d'extraire les données d'identification des dossiers médicaux et d'échanger les chiffres identifiables par des codes aléatoires qui seront utilisés pour représenter les patients, la cellule de gestion de l'identité peut aussi associer ce patient en utilisant ces informations à aucun patient déjà présent dans le CRC qui résulte être la même personne (Murphy et al. 2007).

### 2.3.3 Représentation des informations

L'annotation manuelle de documents est une étape nécessaire lors du développement de systèmes automatiques de désidentification. Même si tous les systèmes de désidentification qui suivent une approche d'apprentissage supervisé ont forcément besoin de données d'entraînement manuellement annotées, tous les systèmes nécessitent de documents manuellement annotés pour l'évaluation. Dans le cadre du projet NLM Scrubber (Kayaalp et al. 2015), les auteurs ont utilisé des documents annotés autant pour le développement que pour l'évaluation. Malgré l'utilisation d'outils pour réaliser une annotation semi-automatique, l'annotation manuelle est une activité coûteuse. Lors du développement de NLM-Scrubber, ils ont annoté un grand échantillon de rapports concernant 7571 patients, fournis par le NIH Clinical Center. Par rapport aux informations identifiables visées, ils ont conçu un espace d'annotation en deux dimensions. D'une part, la première dimension désigne les identifiants personnels, plus particulièrement ils distinguent douze catégories : adresse, prénom,

initiales, organisation, profession, télécommunication, date, âge, temps, identifiants numériques et alpha-numériques, contexte d'identification personnel et rôle. D'autre part, la deuxième dimension concerne l'identité individuelle, qui associe un identifiant à une identité. Ils ont défini cinq catégories : Patient (patient), Relative (proche ou membre de la famille), Employer (employeur), Provider (fournisseur) et Other (autres). Ce double contexte d'annotation donne lieu à considérer le mot « John » comme dénotant un prénom personnel et probablement un patient. L'étiquette finale pourrait ressembler à ceci : PersonalName : : : Patient, s'il s'agit du patient, ou PersonalName : : : Provider, en cas que celui-ci soit un des professionnels de santé.

Un aspect important à prendre en compte du projet NLM-Scrubber c'est qu'en plus des 18 catégories déterminées par l'HIPAA, ils ont également envisagé le contexte comme un autre élément d'identification personnel. Ils ont considéré que dû aux subtilités de la langue naturelle, il est possible de rencontrer des contextes dans lesquels la personne puisse être identifiée indirectement et que les étiquettes existantes ne pourvoyaient pas ces situations. Dans ces cas, ils étiquetaient les tokens comme PIC (*Personally Identifying Context*).

En bref, dans ce rapport (Kayaalp et al. 2015), ils ont introduit un schéma d'annotation qui étend les éléments identifiants du HIPAA Privacy Rule, mais avec l'objectif de dessiner un large éventail d'étiquettes d'annotation afin d'établir un consensus ou une base commune pour la communauté.

Un autre projet qui se base sur les règles HIPAA est celui qui s'intéresse à la désidentification du texte libre contenu dans des rapports médicaux présents dans la base de données MIMIC II, une grande base de données annotée composée de signaux cardio-vasculaires et de données cliniques adjointes à ceux derniers et provenant des unités de soins intensifs des États Unis (Neamatullah et al. 2008). Les auteurs ont créé un outil pour la désidentification écrit en Perl utilisable de façon générale dans la plupart de rapports médicaux contenant du texte libre, comme les notes des infirmières, les résumés de départ, les rapports de radiographies, etc. Cet outil utilise des tables lexicales de recherche, expressions régulières, et d'heuristiques simples pour repérer les PHI de l'HI-

PAA. C'est dans ce but qu'ils ont constitué un corpus de notes des infirmières ré-identifié avec des PHI réels remplacés par des données réalistes de substitution. Ce corpus est composé de 2434 notes d'infirmières contenant un total de 334000 mots et de 1779 instances de PHI comprises dans 163 rapports de patients prélevés au hasard. L'algorithme atteint un rappel de 0,967 et une précision de 0,749. Comme l'on a déjà mentionné, l'approche de désidentification à travers le *pattern-matching* (Neamatullah et al. 2008) est applicable de manière générale à toute partie du texte libre de rapports médicaux. L'algorithme effectue la correspondance (*matching*) lexicale avec les tables de consultation, les expressions régulières et les heuristiques qui vérifient le contexte pour identifier et retirer les PHI. L'approche actuelle désidentifie les noms (de patients, visiteurs, médecins...), les lieux (noms d'hôpitaux, noms de bâtiments, noms de villes, les adresses et les codes postales), les dates, numéros de téléphone ou fax, les numéros d'identification des médecins et patients y compris les numéros de sécurité sociale et des rapports médicaux, des adresses mail, ou aucune information par rapport à l'âge concernant des patients au-dessus de 89 ans.

L'algorithme utilise quatre types de dictionnaires dites « de consultation ». Le premier est une table de PHI contenant des noms de patients et personnel hospitalier connus. Étant donné que la base de données MIMIC II inclut tous les noms complets associés à chaque rapport médical, ceci permet le fait d'extraire tous les noms spécifiques de patients et professionnels par correspondance directe. Deuxièmement, une table de possibles PHI de prénoms féminins et masculins et de noms, préfixes, noms d'hôpitaux, locations et états. De plus, une table de déclencheurs de PHI composée de mots et de phrases clés qui précédent ou suivent des termes PHI. Ces déclencheurs de PHI constituent des indices contextuels pour les titres (« *Mr* », « *Dr* »), les noms (« *mother* », « *son* »), les lieux (« *Hospital* », « *Town* ») et les âges (« *age* », « *patient is* »). En dernier lieu, les tables de non-PHI contiennent des dictionnaires de « mots communs » ou des termes UMLS (*Unified Medical Language System*) qui ont tendance à non être des PHI.

Le processus de désidentification (Neamatullah et al. 2008) implique de

parcourir les notes médicales ligne par ligne et de les diviser en mots séparés par des espaces. Ensuite, on identifie les occurrences de PHI en utilisant les dictionnaires et des expressions régulières. Les instances de PHI qui concernent des patrons numériques tels que les adresses de rue, les dates, les numéros de téléphone et de fax sont identifiés à travers des expressions régulières fondées sur des patrons numériques ainsi que sur des mots-clés contextuels. D'autre part, pour la désidentification de tokens non-numériques, comme les noms ou les lieux, s'utilisent les dictionnaires ainsi que les vérifications contextuelles pour repérer des possibles PHI. L'étape finale du processus de désidentification comporte le remplacement de chaque PHI par une étiquette afin d'indiquer la catégorie à laquelle ceci correspond.

Plus récemment, l'outil NeuroNER (Dernoncourt, Lee et Szolovits 2017) a été conçu pour la reconnaissance d'entités nommées tout en reposant sur des réseaux de neurones. Avec l'objectif de permettre aux utilisateurs de créer et modifier des annotations pour un nouveau corpus, NeuroNER se connecte avec l'interface d'annotation basé sur le web BRAT. NeuroNER effectue le flux annotation-entraînement-prédiction, en même temps qu'exploite les capacités de prédictions des réseaux de neurones. De plus, il s'agit d'un projet *open source* disponible en ligne gratuitement. Le moteur de NeuroNER reçoit comme *input* trois jeux de données : le jeu d'entraînement, le jeu de validation et le jeu d'évaluation. Les fichiers contenus par chaque jeu de données doivent être dans le même format que celui utilisé pour l'outil d'annotation BRAT ou le jeu de données de CONLL-2003 NER. En ce qui concerne les annotations, ce programme permet de tout annoter de zéro ou d'améliorer les annotations d'un jeu de données déjà étiqueté.

# Section 3

# Corpus

## 3.1 Introduction

Tout d'abord, nous voudrions mettre l'accent sur le fait que nous avons travaillé sur deux jeux de données différents et que les expériences menées et les résultats obtenus sont donc distincts selon les données utilisées.

Le premier correspond à des données cliniques réelles, rédigées en espagnol et fournies par l'Hôpital San Juan de Dios de León. Reçues tardivement, nous avons dû réfléchir à la création d'un deuxième jeu de données. Celui-ci est un corpus constitué de messages postés sur des forums de santé, rédigés en anglais et que nous avons construit pour mettre en place la méthodologie avant application sur le jeu de données en espagnol.

Par ailleurs, nous avons accompli deux tâches, une première tâche de désidentification, en cherchant un équilibre entre préservation de la vie privée et conservation d'informations utiles pour le diagnostic, et une tâche de classification de l'amyloïdose.

## 3.2 Messages de forums de santé

D'abord, la première tâche à mettre en œuvre est de construire le corpus de messages de patients. Premièrement, nous avons fait une recherche de forums portant sur des maladies cardiaques rares d'où nous pourrions extraire

des messages de patients atteints de ce type de maladies et qui parlent de leurs cas et décrivent leurs symptômes. Bien qu'il s'agit d'une tâche difficile dû à la nature des maladies que l'on cherche, nous avons éventuellement fini par trouver certaines ressources qui nous ont permis de constituer un corpus. Cependant, tous ces messages sont en anglais, ce qui nous a obligé à refaire les traitements de la phase de désidentification, étant donné que dans un début nous avions développé tout ce processus pour des textes cliniques en espagnol.

Le corpus est constitué de 113 messages extraits de différentes ressources : Methodist Debakay Cardiovascular Journal (KELTY et LAWRENCE 2012), Amyloidosis Support Network (https ://www.amyloidosissupportnetwork.org/), National Amyloidosis Centre Patient Forum (https ://amyloidosis.org.uk/forum), Amyloidosis Foundation (http ://amyloidosis.org), Leukaemia Foundation (https ://www.leukaemia.org.au/) et RareConnect (https ://www.rareconnect.org). Toutes ces ressources sont des forums, certains se focalisent seulement sur l'amyloïdose comme maladie rare mais d'autres contiennent des discussions pour une grande variété de maladies rares. L'objectif de mélanger des maladies cardiaques rares avec l'amyloïdose cardiaque est de tenter que l'algorithme apprenne à distinguer quand est-ce qu'il s'agit d'amyloïdose cardiaque et quand est-ce qu'on traite avec une maladie cardiaque rare. Les messages visés correspondent aux patients atteints soit d'amyloïdose soit d'une maladie cardiaque rare. Les messages concernant les patients d'amyloïdose ont été extraits de toutes les ressources, par contre ceux liés à d'autres maladies ont été seulement pris de deux dernières. En tout, le corpus est composé de messages associés à 8 maladies cardiaques rares : *amyloidosis* (89 messages), *Alström syndrome* (7 messages), *Gaucher disease* (3 messages), *Leber hereditary optic neuropathy* (3 messages), *mitochondrial disease* (6 messages), *myotonic dystrophy* (6 messages), *postural orthostatic tachycardia* (2 messages), et *primary carnitive deficiency* (1 message).

D'abord, nous élaborons une première version du corpus en format XML qui recueille tous les messages. Chaque élément contient les suivantes informa-

tions : maladie, message et ressource. La balise « maladie » nous servira pour l'entraînement du modèle, puisqu'elle correspondra à l'étiquette « y » à prédire par celui-ci.

## 3.3 Dossiers cliniques électroniques de l'Hôpital San Juan de Dios de Léon

Malgré le retard dans l'envoi de données, que nous n'avons reçues qu'à la fin octobre et avec lesquelles nous n'avons pas eu encore le temps de réaliser les expériences nécessaires pour mener à bien la prédiction de l'amyloïdose, nous avons accompli un premier analyse de données. Nous intégrons par la suite un bref analyse et une description de celles-ci.

En premier lieu, les données correspondent aux hospitalisations des derniers cinq ans de l'Hôpital San Juan de Dios de León. Le rang analysé concerne les patients sortis de l'hôpital depuis le 10 juillet 2014 au 10 juillet 2019, ayant ceux-ci un âge au-dessus de 65 ans lors de leur sortie de l'hôpital. Les données sont rassemblées dans des fichiers en format Excel et peuvent êtres classifiées dans deux groupes :

- Données identifiantes. Ce fichier contient entre autres l'identifiant du patient, la date de naissance, le sexe, le code postal, les locations, en plus des données administratives, comme l'identifiant de l'épisode clinique, la date d'admission, les circonstances lors de l'entrée, date de sortie, service de sortie, raison de sortie, s'il s'agit d'une réadmission, etc.
- Données cliniques. Ce groupe contient plusieurs fichiers qui rassemblent les diagnostics cliniques, évolutifs médicales et d'infirmerie, rapports de sortie et notes. Ces derniers incluent des données largement pertinents parmi lesquelles nous pouvons remarquer les signes vitaux, les glycémies et glycosuries, des bilans hidroélectriques, traitements et soins, ulcères, régimes alimentaires, médicaments, consultations externes, ordonnances, etc.

Toutes ces informations sont présentées dans les colonnes des différents fi-

chiers Excel sous le nom correspondant aux codes du système informatique de l'hôpital. De cette façon, par exemple, l'identifiant du patient corresponde à la colonne « gidenpac », la date de naissance se trouve dans « fnacipac » et « gcodicie » recueille les codes CIM9 et CIM10, qui font référence à la Classification internationale de maladies (CIM). La CIM, gérée par l'Organisation mondiale de la santé (OMS), a pour but de permettre l'analyse systématique, l'interprétation et la comparaison de données de mortalité et de morbidité. « La CIM est utilisé pour transposer les diagnostics de maladies ou autres problèmes de santé, en codes alphanumériques, ce qui facilite le stockage, la recherche et l'analyse des données et son utilisation en épidémiologie, en planification et gestion sanitaire ou encore à des fins cliniques » (L'HOSPITALISATION 2015). Pour cette raison, nous avons également reçu des grilles de référence afin de pouvoir comprendre à quoi correspond chaque code.

Le nombre total de patients uniques est de 11586, mais on compte 16620 enregistrements des sorties de l'hôpital. Comme l'on a déjà mentionné, tous les patients ont plus de 65 ans, étant l'âge moyen de 87 ans. La distribution de patients est de 9632 femmes et 6986 hommes et le temps moyen d'hospitalisation est de 18 jours.

Nous avons analysé les diagnostics les plus fréquents, recueillis dans le graphique 3.1, en plus des médicaments les plus prescrits, parmi lesquels nous pouvons mentionner Furosemida, Duphalac, glucosamine, paracetamol, oxygen, lavement, Hibor, chlorure de potassium, Lorazepan, etc.

La table 3.2 recueille aussi les motifs d'hospitalisation les plus habituelles, qui montre que les motifs les plus communs correspondent à des insuffisances cardiaques.

Une fois présentées brièvement les données, nous voudrions analyser plus en détail les patients d'amyloïdose cardiaque se trouvant dans notre jeu de données, étant ceux-ci le cible du projet. Pour ce but, nous avons cherché les enregistrements de patients qui contiennent les codes CIM9 et CIM10 correspondant à l'amyloïdose : E85.4, E85.8, E85.9, E85.81, E85.82, E85.89, 277.39, 277.30. Comme résultat, nous avons extrait 26 patients. Nous décrivons ici la

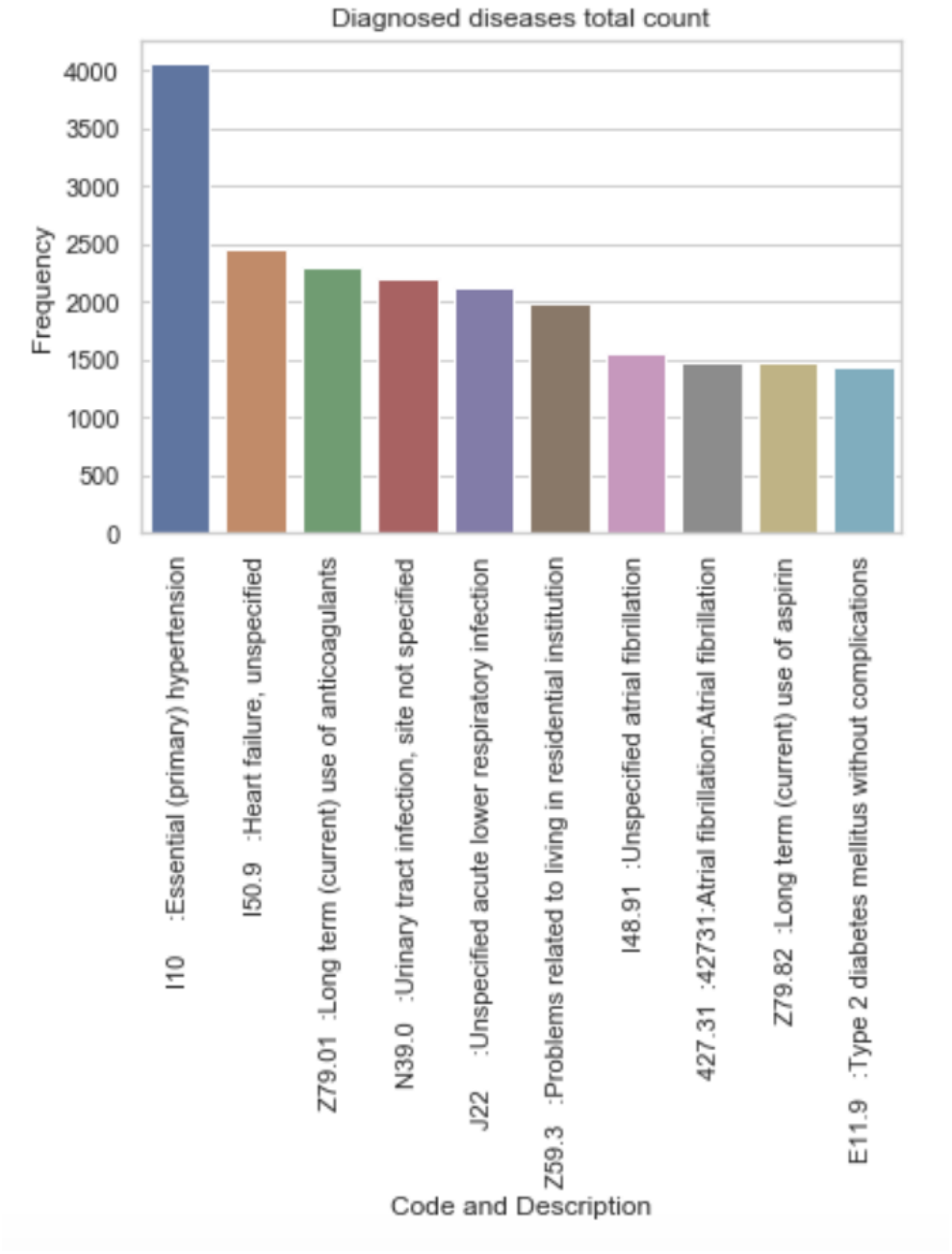
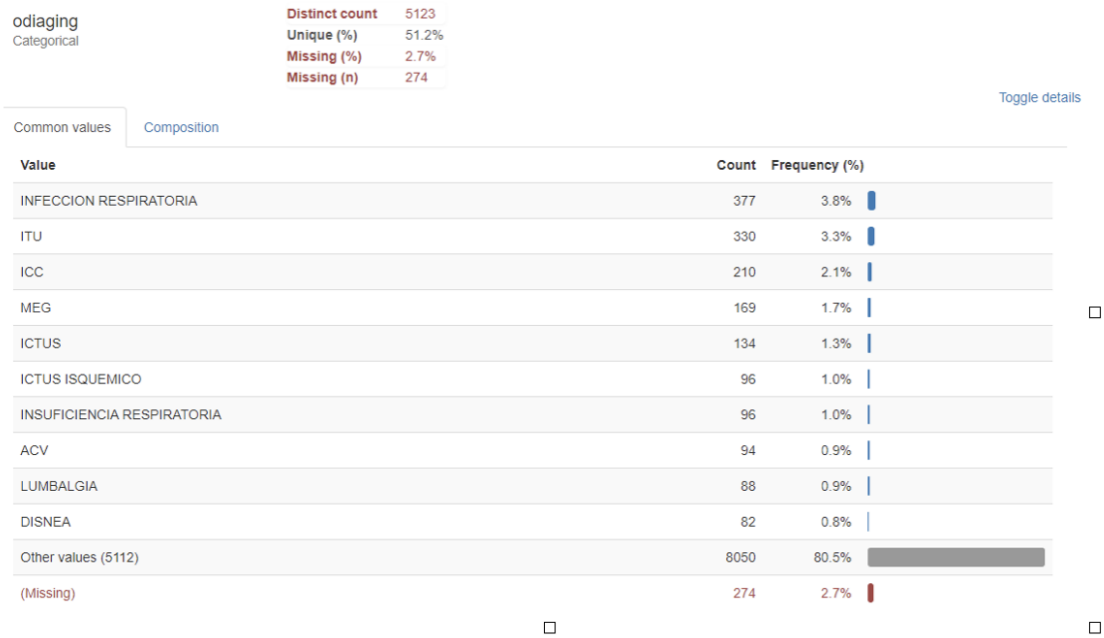



FIGURE 3.1 – Les dix diagnostics les plus fréquents



odiaging
Categorical

| | |
|---|---|
| Distinct count | 5123 |
| Unique (%) | 51.2% |
| Missing (%) | 2.7% |
| Missing (n) | 274 |

Toggle details

Common values | Composition

| Value | Count | Frequency (%) |
|---|---|---|
| INFECCION RESPIRATORIA | 377 | 3.8% |
| ITU | 330 | 3.3% |
| ICC | 210 | 2.1% |
| MEG | 169 | 1.7% |
| ICTUS | 134 | 1.3% |
| ICTUS ISQUEMICO | 96 | 1.0% |
| INSUFICIENCIA RESPIRATORIA | 96 | 1.0% |
| ACV | 94 | 0.9% |
| LUMBALGIA | 88 | 0.9% |
| DISNEA | 82 | 0.8% |
| Other values (5112) | 8050 | 80.5% |
| (Missing) | 274 | 2.7% |

FIGURE 3.2 – Analyse de l'attribut « odiaging » correspondant aux causes d'hospitalisation

suite des identifiants de patients et leurs codes de diagnostics correspondant à l'amyloïdose :

- 'PLGB192704550' : 'E85.4'
- 'LZGR192503050' : 'E85.4'
- 'MRSN193205700' : 'E85.4'
- 'GRRD193301280' : 'E85.4'
- 'PRLV192709130' : 'E85.4'
- 'RMPS192805190' : 'E85.4'
- 'LRRS193209201' : 'E85.4'
- 'RDPN193212200' : 'E85.4'
- 'BHPN192809040' : 'E85.4'
- 'MCPR000000002' : 'E85.4 E85.9'
- 'GRGR192808110' : 'E85.4'
- 'BNCR193006491' : 'E85.4'
- 'VDLN192801430' : 'E85.4'
- 'FLDM192509230' : 'E85.4'

- 'MRRV193208700' : 'E85.4'
- 'LPCR193105680' : 'E85.4'
- 'GNFR192804570' : 'E85.4'
- 'HRVL192705460' : 'E85.4'
- 'RMBL193108230' : 'E85.4'
- 'CRDM194806440' : 'E85.8'
- 'SNMC193501540' : 'E85.9'
- 'SNMN192609660' : 'E85.9'
- 'LVLL192709250' : 'E85.9'
- 'PRSN192704250' : 'E85.9'
- 'MNFR193903191' : '277.39'
- 'TSFR193205290' : '277.30'

Ce code, appelé « gidenpac » dans le système informatique de l'hôpital, est la clé de notre base de données, puisque c'est la seule colonne qui nous permet d'associer tous les tables.

L'âge moyen de patients avec amyloïdose est de 88 ans, qui restent hospitalisés pendant une moyenne de 20 jours. En ce qui concerne la distribution par sexe, il y a 11 femmes et 15 hommes avec amyloïdose. Ils ont dû être hospitalisés entre 2 et 4 fois chacun.

En plus de l'amyloïdose, nous avons observé d'autres diagnostics qu'ont été réalisés à ces patients, le but étant de pouvoir analyser quelles autres maladies, symptômes ou signes peuvent être associées à l'amyloïdose, c'est-à-dire, quelle est la comorbidité de celle-ci.

En dernier lieu, nous intégrons une table qui décrit les causes d'hospitalisation les plus habituelles parmi les patients atteints d'amyloïdose.

Étant donné que l'amyloïdose cardiaque est une maladie rare et que son incidence annuelle, difficile à évaluer, est estimée à 6-10 cas/million, il est absolument compréhensible que l'on trouve que 26 cas d'amyloïdose dans un échantillon de presque 12000 patients. Il faudra tenir compte de ce déséquilibre et du réduit nombre de patients atteints d'amyloïdose lors de la construction de jeux d'entraînement et validation.

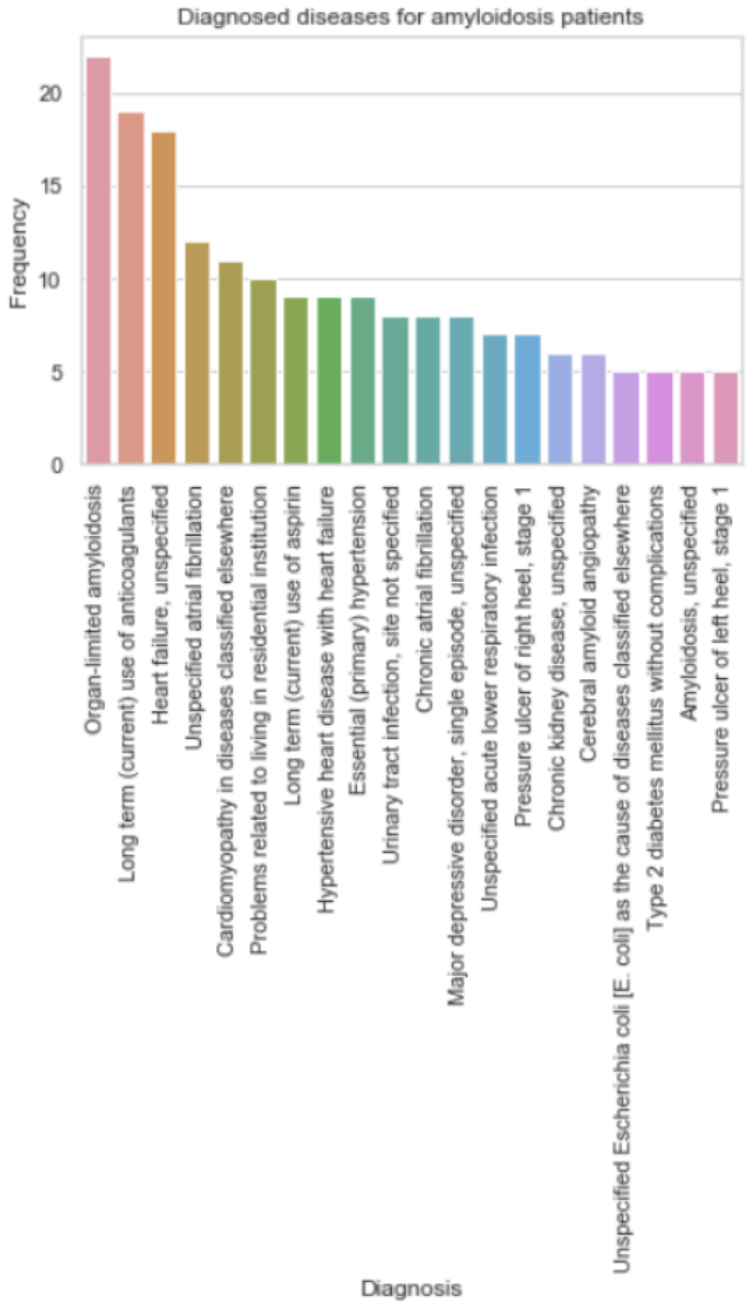



FIGURE 3.3 – Diagnostics réalisés aux patients avec amyloïdose les plus fréquents

odiagalt
Categorical

| | |
|---|---|
| Distinct count | 39 |
| Unique (%) | 90.7% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

Toggle details

Common values | Composition

| Value | Count | Frequency (%) |
|---|---|---|
| ICC | 4 | 9.3% |
| INSUFICIENCIA CARDIACA | 2 | 4.7% |
| HEMATOMA DE GANGLIOS BASALES IZQDOS | 1 | 2.3% |
| INFECCION RESPIRATORIA | 1 | 2.3% |
| SINDROME FEBRIL | 1 | 2.3% |
| ARTRALGIA. FRACTURA L1, L2 CON MAL CONTROL DEL DOLOR | 1 | 2.3% |
| ACVA ISQUÉMICO CON TRANSFORMACIÓN HEMORRÁGICA | 1 | 2.3% |
| ANGOR HEMODINAMICO EN EL COTEXTO DE FIBRILACION AURICULAR E ICC | 1 | 2.3% |
| ERC AGUDIZADA + NEUMONIA LII + IC LEVE | 1 | 2.3% |
| ITU,INFECCION RESPIRATORIA | 1 | 2.3% |
| Other values (29) | 29 | 67.4% |

FIGURE 3.4 – Motifs d'hospitalisations de patients avec amyloïdose les plus fréquents

## 3.4 Conclusion

Enfin, le fait d'avoir travaillé sur deux corpus nous a permis d'implémenter différentes méthodologies, de réaliser plusieurs expériences et nous a obligé à réfléchir et concevoir des tâches identiques mais en s'adaptant aux particularités de chaque jeu de données.

Il faut reconnaître que les deux jeux des données sont complètement différents. L'un est composé de données structurées et non-structurées, l'autre ne contient que du texte libre. Les contenus sont également distincts. La caractéristique principale du deuxième jeu de données est que le texte a été rédigé, dans la plupart de cas, par les patients eux-mêmes, en décrivant de première main leurs expériences, symptômes et douleurs. Pourtant, les données fournies par l'hôpital sont des dossiers cliniques qui rassemblent des notes médicales et d'infirmerie, des résultats de tests ou l'évolution du patient rédigé par le médecin en charge, renferment donc aussi bien des données structurées comme des données non structurées.

Il est vrai que cette dissemblance a entraîné un effort supplémentaire en termes de travail et conception du projet, mais par contre nous trouvons que,

dans les grandes lignes, les deux corpus partagent un point commun essentiel, l'information qu'elles fournissent : la description d'une maladie, plus particulièrement l'amyloïdose.

À cause de cela, malgré ces incontestables différences des deux jeux de données, autant en termes de contenu que de format, le corpus constitué de messages de forums s'avère finalement valable pour réaliser des premières expériences visant les tâches de désidentification et classification entre amyloïdose et non amyloïdose.

# Section 4

# Expériences et discussions

## 4.1 Expériences

### 4.1.1 Désidentification

#### Introduction

Tout d'abord, nous voudrions mettre en relief que dû aux contraintes imposées par l'institution Fundación San Juan de Dios et aux spécificités de chacun des corpus, particulièrement la langue, nous avons développé deux méthodes différentes selon le corpus de travail en question.

Alors que pour le premier corpus de messages notre solution repose principalement sur une tâche d'annotation et d'entraînement d'un modèle CRF, le processus de désidentification de textes cliniques en espagnol repose sur l'approche dont nous avions parlé précédemment dite orientée connaissances. De ce fait, dans le premier cas, nous avons annoté un corpus composé de 113 messages basé sur un jeu d'étiquettes construit visant ce corpus en particulier. Nous rappelons que cette approche basée sur les données réussit des bonnes performances et que le coût d'entrée, bien qu'il n'est pas si important, nécessite d'un corpus annoté avec l'objectif de pouvoir entraîner et évaluer le modèle. En revanche, pour le deuxième corpus, nous avons suivi une deuxième méthodologie de sorte qu'on a créé des règles grammaticales et syntaxiques, des expressions régulières et patrons, et nous avons utilisé des ressources linguistiques comme

des dictionnaires et des listes.

### Désidentification de messages de forums

Commençons par examiner les traitements développés pour la désidentification du corpus de messages de forums. Comme l'on a exposé précédemment, pour ce corpus, nous allons rester sur le deuxième type de méthodes en ce qui concerne la tâche de désidentification, plus particulièrement sur l'algorithme Conditional Random Field (CRF), comme l'ont a déjà fait Aramaki et al. (2006) et Gardner et Xiong (2006).

Le premier pas est de décider quelles seront les étiquettes que nous allons utiliser pour la tâche d'annotation d'informations personnelles. Étant donné qu'il n'existe pas de loi européenne qui définisse quelles sont les données exactes qui ne doivent pas être présentes dans un texte pour que celui-ci soit considéré comme anonyme, nous allons nous en servir de la réglementation HIPAA et les dix-huit éléments définis comme PHI (Personal Health Information) : *names*, *dates*, *ages > 89 years*, *telephone numbers*, *fax numbers*, *electronic mail addresses*, *social security numbers*, *medical record numbers*, *health plan beneficiary numbers*, *account numbers*, *certificate license numbers*, *vehicle identifiers*, *device identifiers and serial numbers*, *Web Universal Resource Locators (URLs)*, *Internet Protocol (IP) address numbers*, et *biometric identifiers*.

Comme on a dû travailler avec des données d'entrée différents à ceux conçus initialement à cause des contraintes déjà mentionnées, cela veut dire qu'il ne s'agit plus de dossiers cliniques mais des messages des patients racontant leur propre cas, il y a donc des informations qui ne sont pas présents ou qui sont peu fréquents dans le corpus construit, comme des numéros de téléphone, des numéros de comptes, des numéros de la sécurité sociale, etc.

Finalement, le jeu d'étiquettes pour l'annotation des données identifiantes comprend : « Account number », « Age 0-15 », « Age 15-40 », « Age 40-60 », « Age 60-80 », « Age 80-100 », « Biometric identifier », « Certificate number », « Date », « Device identifier », « Fax », « Health plan beneficiary », « IP »,

« Location », « Mail », « Medical record », « Name », « Social security », « Telephone », « URL », et « Vehicle Number ». L'étiquette « Age » est étendue puisqu'elle servira autant pour l'annotation des informations identifiantes comme cliniques.

**Traitements de données** Pour l'entraînement de notre modèle de désidentification, nous allons utiliser Wapiti, un outil qui implémente le modèle Conditional Random Fields (CRF) pour la reconnaissance d'entités nommées.

Wapiti reçoit en entrée un fichier en format tabulaire où chaque ligne contient les informations correspondantes à un token en particulier. Le dernier élément de la ligne, c'est-à-dire, la dernière tabulation, correspond à l'étiquette à apprendre et donc à prédire par le modèle.

On intègre ci-dessous les différents traitements réalisés sur les données afin d'aboutir aux « datasets » finals :

- Annotation sur Brat et génération des fichiers *.ann
- Production de tabulaires. Transformation de tous les fichiers du répertoire, qui contient des fichiers *.txt pour chaque message, en fichiers *.tok avec deux colonnes, colonne d'offsets de début et de fin du token et colonne du token. Ensuite, on exécute un script qui tokénise sur l'espace mais découpe également selon la ponctuation. Nous avons réutilisé le script fourni lors du cours Fouille de textes du deuxième année du Master Traitement Automatique des langues (IM).
- Pour chaque token, en plus de l'étiquette créée suivant le format BIO, nous avons ajouté d'autres informations qui peuvent aider à l'apprentissage du modèle CRF : longueur du token, présence de signes de ponctuation, de majuscules ou des chiffres.
- *Part-of-speech.* En plus des caractéristiques produites précédemment, nous avons décidé d'ajouter le *POS* comme une autre caractéristique interne du token. Avec le module TreeTagger Python Wrapper, qui utilise l'outil TreeTagger pour « tagger », nous avons réalisé l'étiquetage morphosyntaxique de chaque token et nous avons ajouté le *POS* comme

une information interne du token en plus de celles citées précédemment.
- Division du corpus en trois jeux de données : entraînement (60%), développement (20%) et test (20%).
- Masquage d'étiquettes cliniques : toutes les étiquettes concernant des données cliniques sont remplacées par des étiquettes BIO « O ». Les différentes étiquettes correspondant à des tranches d'âge sont rassemblées dans une seule étiquette « Age ».

Finalement, nous avons obtenu un fichier qui ressemble à cela :

*A 1 Mm NUL NUL DT O*
*renal 5 mm NUL NUL JJ B-Biopsy*
*biopsy 6 mm NUL NUL NN I-Biopsy*
*was 3 mm NUL NUL VBD O*
*completed 9 mm NUL NUL VVN O*
*in 2 mm NUL NUL RB O*
*November 8 Mm NUL NUL NP B-Date*
*2018 4 mm NUL DIGIT LS I-Date*
*, 1 mm PCT NUL , O*
*which 5 mm NUL NUL WDT O*
*revealed 8 mm NUL NUL VVD O*
*AL 2 Mm NUL NUL NP B-Disorder*
*amyloidosis 11 mm NUL NUL NN I-Disorder*
*. 1 mm PCT NUL SENT O*

### Désidentification de textes cliniques

Par la suite, nous voudrions remarquer que le fait que ce corpus soit construit à partir du texte des messages trouvés sur des forums nous a permis de mener à bien le méthode que l'on souhaitait : l'annotation des données textuelles suivie de l'entraînement d'un algorithme d'apprentissage automatique. Cependant, du fait des exigeantes mesures de sécurité du département d'informatique de l'Hôpital San Juan de Dios, nous n'avons pas pu implémenter la même méthodologie sur ces dernières données. Plus précisément, nous

n'avons pas eu d'accès aux données ayant des données identifiantes, et donc cibles du processus de désidentification. C'est la raison pour laquelle, de manière à mettre en place la désidentification des données cliniques, correspondant à des données non-structurées et qui généralement soit font référence à l'évolution des patients lors de leur hospitalisation soit elles correspondent aux notes de médecins ou d'infirmières, nous avons adopté l'autre des approches les plus utilisées pour cette tâche. Pour réaliser le processus de désidentification, nous nous sommes déplacés au département d'informatique de la Fundación San Juan de Dios, du fait que les données ne pouvaient pas « sortir » de l'institution jusqu'à ce qu'ils étaient débarrassées des données personnelles.

En plus, nous n'avons pas eu l'opportunité de visualiser le corpus. Nous avons dû développer un processus de désidentification qu'un tiers a exécuté sur les données.

Étant donné les conditions imposées par l'institution, nous avons décidé de développer une approche dite orientée connaissances, basée sur des ressources linguistiques, expressions régulières et règles.

La méthode proposée se déroule en une seule étape même si elle est composée de deux tâches : le repérage des données identifiantes dans le texte, et la désidentification par remplacement de ces données.

En premier lieu, il faut déterminer quelles sont les données considérées comme sensibles et qui seront objet du processus de désidentification. Les catégories de données à enlever ont été définies par le « fournisseur » des données, dans ce cas, la Fundación San Juan de Dios. Bien que l'on croît qu'il y a plus des données qui doivent être considérées comme des PHI, les catégories déterminées sont les suivantes : noms et prénoms, identifiant du patient, numéro de téléphone et numéro de la carte d'identité.

Il faut souligner que lors de l'élaboration du processus de désidentification, nous avons suivi les indications du client, même si parfois on aurait décidé d'agir autrement.

Une fois les catégories déterminées, nous nous sommes demandé comment est-ce qu'on allait procéder afin d'extraire chacune des catégories et quel devrait

être le *pipeline* à suivre concernant le traitement automatique du langage.

D'abord, on a créé une fonction qui « normalise » le texte, c'est-à-dire, qui remplace les accents et met des espaces entre les caractères de ponctuation. On a utilisé cette fonction tout au long du processus de TAL pour la désidentification des données.

Afin de parvenir à une bonne performance du processus de désidentification, on a décidé de traiter chacune des catégories séparément et de concevoir des méthodologies selon leurs propriétés et particularités. Une fois décidé quelles sont nos données cibles, il faut décider comment est-ce qu'on va les traiter. Comme l'on a vu précédemment, ils existent plusieurs méthodes, soit remplacer par une étiquette, soit par une chaîne de caractères, soit par un élément vraisemblable, etc. Les types d'informations recherchés peuvent être regroupés en deux classes : d'abord les informations de type chaîne de caractères, telles que les noms et prénoms, et ensuite les informations pouvant être représentées par un patron, telles que les numéros de téléphone ou les numéros d'identité personnelle.

D'après la Fundación San Juan de Dios, les données considérées comme « cibles » de la tâche de désidentification doivent être remplacées par des données vraisemblables.

Bien que nous étions restreints en ce qui concerne les méthodes à implémenter, dans un deuxième temps, il serait possible d'annoter le corpus, plus spécifiquement les données qu'ont été insérées pour remplacer les vraies données personnelles, pour ensuite entraîner un modèle CRF de la même façon que nous avons fait avec l'autre corpus. À partir de là, nous pourrions réutiliser notre modèle CRF pour désidentifier d'autres corpus cliniques. En fait, bien que la tâche d'annotation serait coûteuse en temps, la performance de ce modèle s'avère davantage meilleure que celle des modèles construits antérieurement, étant donné que cette fois-ci on compterait avec un large corpus.

**Pipeline de désidentification** Premièrement, nous nous sommes occupé des noms et prénoms. Pour cette catégorie, nous nous en servons de deux ressources.

D'une part, des listes de noms et prénoms existant en Espagne et fournies par l'Institut National de Statistique. D'autre part, du paquet StanfordNLP de Python. Nous avons implémenté l'outil de reconnaissance d'entités nommées (StanfordNERTagger) afin d'étiqueter le texte et chercher les tokens étiquetés comme « PERSON ». On parcourt la liste de tuples générée, en cherchant des tokens dont l'étiquette est « PERSON », si c'est le cas, on vérifie que le token se trouve dans une des listes. Les tokens appartenant à ces listes seront remplacés aléatoirement par d'autres composants de la liste, soit par un prénom masculin ou un prénom féminin, selon la liste où il se trouvé, soit par un nom. Parfois un token peut se trouver dans les deux listes car il peut être considéré autant un nom comme un prénom. Dans ce cas, une règle qui vérifie les tokens voisins décide s'il s'agit d'un nom ou d'un prénom.

En ce qui concerne les initiales, le processus est un peu plus élaboré. On cherche des tokens composés par une seule lettre et qui soit une majuscule. Pour les repérer, on se base sur deux indices contextuels : le voisinage d'un prénom précédent ou d'un nom qui le suit. Si un mot d'un sel caractère apparaît à proximité immédiate d'un nom, il est considéré comme prénom, par contre s'il est précédé par un prénom, on le considère un nom.

Par ailleurs, étant donné que notre corpus est composé de textes médicaux, il est plus qu'évident que ceux-ci mentionneront souvent des noms de maladies, enfermant ceux-ci parfois des noms propres. Normalement, « Alzheimer » doit être ignoré et laissé dans le texte. Par contre, si ce mot est étiqueté comme étant une entité « PERSON » et qu'il est présent dans la liste de noms, il faudra être vigilant à ne pas les remplacer et effacer conséquemment du contenu médical important. Afin d'éviter le remplacement de noms de maladies, nous avons créé une liste noire qui inclut de noms de maladies, extraits automatiquement à partir de plusieurs sites webs. De cette façon, à chaque fois qu'on trouvé un token « PERSON », il est confronté à la liste noire de noms de maladies pouvant être considérés aussi comme des noms.

Les dernières catégories ont été traitées avec des expressions régulières et patrons. Toutes ces catégories reposent sur des patrons alphanumériques fa-

cilement repérables à travers des expressions régulières. Les identifiants des professionnels de l'hôpital sont composés de neuf chiffres, donc ils sont faciles à trouver et remplacer par une autre suite de chiffres. Pour les numéros de téléphone, on a conçu l'expression régulière « (9|6|7)[0-9]8 » et ils sont remplacés par une autre suite de chiffres. Finalement, les cartes d'identité espagnoles, qui suivent le patron « [0-9]8- ?[A-Z] » ont été également remplacés par d'autres semblables.

### Résultats

Pour réaliser l'entraînement du modèle CRF avec notre corpus, nous avons utilisé Wapiti. Le rappel du modèle est de 98,85%, mais la précision est plus baisse, d'un 78,57%. L'entité qui atteint le taux le plus élevé est celle de la date. Évidemment, les catégories qu'on n'a pas trouvé dans le corpus ou qui n'avaient pas de suffisantes occurrences n'ont pas eu aucun résultat suite à l'entraînement.

Mesures d'évaluation du modèle CRF pour la d'identification

| Accuracy | Precision | Recall | FB1 |
|---|---|---|---|
| 98,85 % | 78,57 % | 43,31 % | 55,84 % |
| Age | 60 % | 30 % | 40 % |
| Date | 90,91 % | 62,50 % | 74,07 % |
| Location | 76,92 % | 40 % | 52,63 % |
| Name | 70,59 % | 34,29 % | 46,15 % |

FIGURE 4.1 – Évaluation du modèle de désidentification CRF

## 4.1.2 Prédiction

L'idée initiale était de prédire l'amyloïdose cardiaque par le biais d'un jeu de données construit à partir des informations extraites de dossiers cliniques de patients issus des hôpitaux de la Fundación San Juan de Dios. À cause du retard dans l'envoi de données, nous avons décidé de continuer avec le même sujet mais en utilisant des données d'entrée différentes.

Comme l'on a déjà exposé, le corpus final est donc composé de messages de patients atteints de différents maladies cardiaques rares en format texte libre.

L'objectif est de prédire si le patient est atteint d'amyloïdose ou non.

Encore une fois, à cause du fait de la nature de l'amyloïdose cardiaque comme maladie rare, il a été très difficile de trouver des messages de patients atteints. De plus, comme nous avons décrit précédemment (voir 1.2.1), les amyloses cardiaques appartiennent en même temps à un type d'amyloïdose, selon la protéine générée. C'est pour cela que dans l'ensemble de messages il y a des patients qui sont atteints d'un type d'amylose, par exemple AA, mais qu'en plus mentionnent que leur organe atteint est le cœur ; ceux-ci sont considérés comme des patients d'amyloïdose cardiaque.

La première approche envisagée était de n'annoter que les symptômes, et de masquer les noms de maladies afin de ne perturber le modèle et d'éviter que le classifieur se fonde sur les occurrences du mot « amyloïdose » et des expressions dérivées du mot pour effectuer la classification. Cependant, après une observation plus en détail du corpus, nous nous sommes aperçus qu'il y a plus de maladies qui sont mentionnées dans les messages en plus de celle que fait l'objet de la prédiction, et que probablement, si on les remplaçait par des labels comme « <DISORDER> », on perdrait d'information qui puisse-être utile, puisque certaines maladies peuvent causer ou être associées à d'autres maladies.

D'autre part, lorsque nous avons commencé à annoter le corpus, nous avons remarqué qu'il y a d'autres informations cliniques importantes et pertinentes pour la tâche de prédiction en plus des symptômes. Nous avons donc décidé de concevoir un jeu d'étiquettes plus étendu afin que le modèle apprenne de toutes les informations cliniques qui puissent être présentes dans le message.

Le jeu d'étiquettes concernant les informations cliniques est compris par 36 éléments. Certains ont été pris à partir de l'étude réalisé pour la prédiction de l'amyloïdose cardiaque (Garg et al. 2016), comme arrêt cardiaque, hypertension, etc., et d'autres ont été rajoutés lors de la tâche d'annotation, dû au besoin de les intégrer étant donné leur forte présence tout au long du corpus. Parmi ces derniers nous pouvons mettre l'accent sur les médecins de spécialité, les médicaments, les organes atteints, les transplants, les difficultés

respiratoires, etc. Nous intégrons ci-dessous l'inventaire final d'étiquettes :

- Biopsy
- Breathlessness. Tout ce qui concerne des problèmes respiratoires : « *very out of breath* », « *shortness of breath* », « *struggling breathing* », « *difficult breathing* »...
- Cardiac arrest. Nous avons décidé de créer une classe spécifique pour l'arrêt cardiaque étant donné la pertinence de celui-ci comme symptôme de l'amyloïdose cardiaque. Cependant, nous remarquons que finalement on n'a trouvé que trois occurrences pour cette étiquette.
- Chemotherapy. Les traitements de l'amyloïdose, en particulier de l'amyloïdose AL, visent d'abord à éliminer les cellules qui fabriquent les anticorps monoclonaux par des protocoles de chimiotérapie, le plus souvent dérivés de ceux de myélome. Les plus utilisés sont le MDex et le VCD (Georgin-Lavialle et al. 2017).
- Dialysis. Lorsque les reins ne fonctionnent plus à cause d'une amyloïdose, les patients sont souvent traités par la dialyse.
- Digestive problems. Tout au long de la tâche d'annotation, nous nous sommes aperçus que beaucoup de patients décrivaient des difficultés avec la digestion et l'alimentation : « *digestive disorders* », « *chronic diarrhea* » ou « *trouble eating* ».
- Disorder. Tous les mentions de maladies : « *amyloidosis* », « *Hepatitis C* », « *Laukaemia* », « *HIV* », « *diabetes* », etc.
- Edema
- Eye conditions. Cette étiquette a aussi été créé lors de l'annotation à cause des occurrences des symptômes et problèmes concernant les yeux mentionnés par les utilisateurs. En fait, l'amylose à TTR touche principalement les nerfs (neuropathie), le cœur, mais aussi plus rarement les yeux (Georgin-Lavialle et al. 2017) : « *getting black eyes* », « *persistent dark circles around my eyes* », « *purple hue* », « *bruising around his eye* » ou « *stye on his bottom eye lid* ».
- Heart failure. Du fait que l'objectif était de prédire l'amyloïdose car-

diaque, les symptômes, traitements et examens concernant le cœur ont béneficié de classes particulières : « *severe congestive heart failure* », « *cardiac issues* », « *excessive heartbeat* », « *restrictive cardiomyopathy* », « *thickened heart walls* ». Toutes ces difficultés sont considérés des symptômes particuliers à l'amyloïdose cardiaque.

- Heart test. « *Cardiac MRI* », « *echocardiography* » ou « *EKG* ».
- Hypertension
- Hypertrophy. Comme l'on a déjà exposé précédemment, un des symptômes le plus important pour l'amyloïdose c'est la rigidité et l'augmentation de volume des organes, plus particulièrement du cœur : « *enlarged heart* », « *thickening of his heart* », « *heart was too muscular* », « *concentric left ventricular hypertrophy* », « *liver and spleen were bigger* », etc.
- Hypotension
- Inheritance. Étant donné que l'une des caractéristiques de l'amyloïdose peut être son caractère héréditaire, cette classe essaie de recueillir cette possibilité : « *family history with the disease* » ou « *inherit the disease* ».
- Measures : « *22.8mg/L* », « *5 mg/dL* », « *increased to 200* », « *67-70 mm Hg* » ou « *600mg* ».
- Medical practitioner. Les utilisateurs parlent souvent des médecins spécialistes qui les ont traité. De fait, il y a des nombreuses occurrences : « *cardiologist* », « *neurologist* », « *oncologist* », « *hematologist* », « *neprologist* »...
- Medication. Dans l'article *Mieux connaître Les Amyloses* (Georgin-Lavialle et al. 2017), ils citent plusieurs médicaments que sont employés actuellement pour traiter l'amyloïdose. Pour cette raison, nous avons ajouté cette étiquette afin de les ressembler ou d'intégrer d'autres médicaments pour aider le modèle à associer des médicaments à des maladies : « *Velcade* », « *CyBorD* », « *Taffamidis* » et « *Thalidomide* » entre autres.

- Mobility limitations. La plupart des patients décrivent leurs difficultés pour marcher, faire du sport, etc. : « *difficulty getting up* », « *difficulty of walking* », « *wheelchair* », « *not walk independently* », « *not being able to ride bike* »...
- Nausea. Il existe une forte présence des expressions concernant ce symptôme : « *vomiting* », « *chronic nausea* », « *dizziness* », etc.
- Nervous system. Du fait que le système nerveux est souvent atteint par l'amyloïdose, cette étiquette rassemble tous ces cas : « *neuropathy* », « *nerve pain in my head* », « *nephrotic syndrome* », « *neurological problem* ».
- Obesity. Même si l'obésité est l'un des symptômes ou indicateurs des maladies cardiaques, nous n'avons pas trouvé des nombreuses occurrences dans notre corpus.
- Organ affected. Cette étiquette recueille les organes atteints de chaque patient.
- Pacemaker. Comme les *pacemakers* sont souvent utilisés comme partie du traitement pour l'amyloïdose cardiaque, nous avons décidé de le considérer comme une classe pour l'annotation.
- Part body. Parties du corps mentionnés mais qui ne sont pas strictement l'organe atteint.
- Pressure. « *Blood pressure was on the low side* », « *barely pumping blood* », « *uncontrolled high blood pressure* »...
- Protein. L'origine de l'amyloïdose est la création des protéines malformées. Cette étiquette comportera un élément clé pour la prédiction de l'amyloïdose : « *amyloid* », « *k protein (Kappa)* », « *depositing abnormal proteins* », « *protein in my urine* », « *elevated protein* », « *new deposits of amyloid fibrils* », etc.
- Psychological problems. Les patients parlent parfois des conséquences psychologiques des maladies : « *anxiety* », « *depression* », « *sessions with a psychologist* », « *psychiatric support* », etc.
- Remission. Habituellement, on parle de rémission lorsqu'une affection

cède du terrain et que l'état du patient s'améliore, soit temporairement soit par complète. Par contre, cela ne signifie pas toujours que la maladie a été totalement éliminée. Par exemple : « *going into remission* », « *will reach remission* », « *still in remission* », « *CR* ».

- Sensitivity. « *Numbness* », « *body to burn* », « *numbing sensations* », « *lower body burns mostly* ».
- Symptom. Cette classe rassemble tout ce qui est considéré comme symptôme mais que n'appartient pas à une classe particulière : « *fevers* », « *pain in my liver area* », « *chest pain* », « *cholesterol* », « *mutated gene* », « *abdominal pain* », « *kidney failure* », etc.
- Test : « *blood test* », « *MRI* », « *bone narrow test* », « *Congo red rye* », « *endocospy* », « *ultrasoung* », « *genetic test* », « *urine test* »...
- Transplant. Le greffe de l'organe atteint qui produit les protéines anormales est souvent l'une des solutions à l'amyloïdose : « *stem cell transplant* », « *heart transplant* », « *bonne marrow transplannt* », « *liver transplant* », « *SCT* »...
- Treatment. Tout ce qui est considéré comme traitement mais qui n'est pas inclus dans aucune des dernières classes : « *anti-nausea medications* », « *thrombectomy* », « *immunotherapy* », « *valve replacement* », « *physiotherapy* ».
- Weakness. Très fréquemment, les patients expérimentent des faiblesses : « *fatigued* », « *very tired* », « *body was weakening* », « *tiredness* », « *debilitating* », « *sleeping much more* », « *strength had diminished* ».
- Weight. « *Losing weight* », « *lost 40 pounds* », « *lost an extraordinary amount of weight* », « *lose weight* »...

Du fait que les messages recueillis concernent des maladies cardiaques rares, nous avons décidé de créer des étiquettes plus précises et spécifiques à celles-ci, comme par exemple « Heart test » ou « Heart failure ».

D'autre part, étant donné que l'amyloïdose est une maladie qui touche principalement l'adulte environ 65 ans, nous avons décidé de faire des tranches supplémentaires pour l'attribut âge, puisque celui-ci deviendra probablement

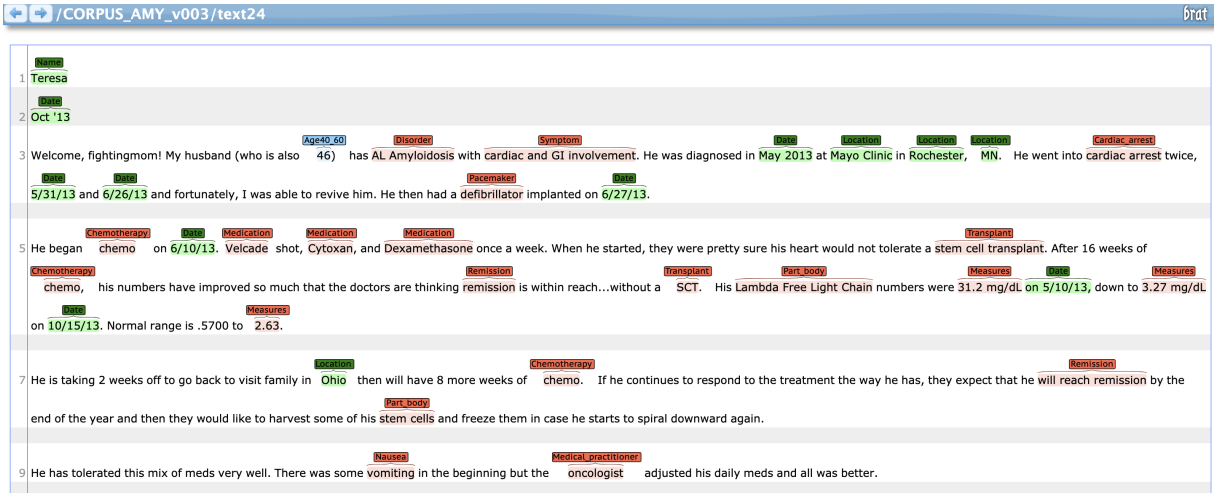



FIGURE 4.2 – Annotation d'un message dans l'outil BRAT

un facteur important pour le diagnostic de la maladie. Finalement, les tranches définies sont les suivantes :

- 0-15 ans
- 15-40 ans
- 40-60 ans
- 60-80 ans
- 80-100 ans

L'outil d'annotation utilisé était BRAT, un outil d'annotation basé sur le web pour l'annotation de textes. Nous intégrons un exemple d'annotation (figure 4.2) d'un des messages du corpus dans le serveur BRAT :

### Traitements des données

Les algorithmes d'apprentissage automatique fonctionnent sur des *features* numériques, en attendant comme *input* un *array* de deux dimensions où les lignes sont des instances et les colonnes sont des *features*. Si l'on veut exécuter des algorithmes d'apprentissage automatique sur des données textuels, il est donc nécessaire de transformer nos « documents » en représentations vectorielles de telle sorte que l'on puisse implémenter des algorithmes d'apprentissage automatique numérique. Ce processus est connu comme *feature extraction* ou plus simplement, *vectorization*.

Le fait de représenter des documents de façon numérique nous permet d'effectuer des analyses et de créer les instances avec lesquelles travaille l'apprentissage automatique.

L'encodage le plus simple de l'espace sémantique est le modèle *bag of words*, dont l'approche principale est que le signifié et la similarité sont encodées dans le vocabulaire. Ce modèle, bien que simple, est extrêmement effectif et représente le point de départ d'autres modèles plus complexes.

Afin de vectoriser un corpus avec l'approche *bag of words* (BOW), nous représentons chaque document du corpus comme un vecteur dont la longueur est celle du vocabulaire du corpus. Nous pouvons considérer quatre types d'encodage de vecteurs : fréquence, one-hot encoding, TF-IDF, et représentations distribuées.

Les vecteurs de fréquences comportent les modèles d'encodage les plus simples puisqu'ils consistent à remplir le vecteur avec la fréquence de chaque mot qui apparaît dans le document. Suivant ce schéma d'encodage, chaque document est représenté comme un *multiset* de tokens qui le composent et la valeur pour chaque position du mot dans le vecteur est son compteur. Cette représentation peut être soit le dénombrement lui-même (*integer*) soit l'encodage normalisé où chaque mot est pondéré selon le nombre total d'occurrences du mot dans le document.

*One-hot encoding* est une méthode d'encodage de vecteur booléen qui indique si le token existe dans le document (1), ou sinon (0). Autrement dit, chaque élément du vecteur *one hot* reflète la présence ou l'absence du token. L'encodage *one-hot* ramène un document à ses constituants.

Nous allons implémenter ces deux dernières approches pour la transformation des nos messages.

Nous avons choisi l'outil Weka pour mener à bien nos expériences de classification de messages en ce qui concerne l'amyloïdose. Weka reçoit en entrée un corpus où chaque document est représenté par un vecteur où les indices correspondent à un mot du vocabulaire total et la valeur est soit 1 si le token est présent dans le document en particulier, 0 sinon, c'est-à-dire, il suit l'approche

*one hot encoding*

Une fois qu'on a transformé chacun de nos messages du corpus en vecteurs, il faut créer le fichier .arff que l'on va donner en entrée à l'outil Weka pour entraîner les modèles. Le fichier en format .arff se compose des suivants sections :

- ARFF Header Section : cette section contient les déclarations de relation et d'attributs. Le nom de relation est défini dans la première ligne du fichier, dans notre cas, « classification ». Les déclarations d'attributs prennent la forme d'une séquence ordonnée qui définit le nom de l'attribut et le type de donnée. L'ordre dans lequel les attributs sont déclarés indique la position de la colonne dans la section de données du fichier, c'est-à-dire, si un attribut est le troisième à être déclaré, Weka attend que toutes les valeurs de cet attribut se trouvent dans la troisième colonne. Dans notre cas, nos attributs sont les mots du vocabulaire extrait du corpus. La dernière ligne de cette section contient les différentes classes à apprendre par le modèle.
  *@ATTRIBUTE 'abdomen' numeric*
  *@ATTRIBUTE 'abdominal' numeric*
  *@ATTRIBUTE 'abilities' numeric*
  *@ATTRIBUTE 'ability' numeric*
  *@ATTRIBUTE 'ablation' numeric*
  *...*
  *@ATTRIBUTE 's_label' LEBER_HEREDITARY_OPTIC_NEUROPATHY, MYOTONIC_DYSTROPHY, ALSTRÖM_SYNDROME, AMYLOIDOSIS_DIS, POSTURAL_ORTHOSTATIC_TACHYCARDIA, GAUCHER, PRIMARY_CARNITINE_DEFICIENCY, MITOCHONDRIAL*
- ARFF Data Section : cette partie du fichier contient la ligne de déclaration de « data » et les lignes des instances. Chaque ligne dans notre fichier .arff correspondra à un vecteur qui représente un message de notre corpus. À la fin de la ligne se trouve la classe à prédire.
  *0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,*

*AMYLOIDOSIS_DIS*

### Résultats

Nous avons décidé de mener à bien deux types de classifications : une première classification qui vise à prédire la maladie dont parle un utilisateur dans son message, et un deuxième type de classification qui envisage la classification entre amyloïdose et non amyloïdose.

**Classification multiclasse** Cette première classification comprend huit classes, celles de maladies à prédire : *amyloidosis*, *Alström syndrome*, *Gaucher disease*, *Leber hereditary optic neuropathy*, *mitochondrial disease*, *myotonic dystrophy*, *postural orthostatic tachycardia*, *primary carnitive deficiency*.

Les résultats sont assez décevants étant donné la seule classe pour laquelle le modèle fournit des vraies mesures est celle de l'amyloïdose. Le reste de maladies obtiennent 0 ou aucun pourcentage pour les mesures de précision, rappel ou F-mesure.

Nous intégrons une synthèse des pourcentages obtenus (4.3) pour la classe « Amyloidosis » pour les différents algorithmes implémentés.

Classification multiclasse

| Classe - Amyloïdose | Precision | Recall | F-mesure |
|---|---|---|---|
| SMO | 0,802 | 1 | 0,890 |
| Naive Bayes Multinomial | 0,788 | 1 | 0,881 |
| Random Forest | 0,788 | 1 | 0,881 |
| Decision Tree | 0,787 | 0,787 | 0,787 |
| Naive Bayes | 0,793 | 0,989 | 0,880 |
| Classification via regression | 0,808 | 0,944 | 0,870 |

FIGURE 4.3 – Mesures d'évaluation obtenues pour la classe « amyloidosis » des algorithmes d'apprentissage automatique implémentés

| Algorithm | Correctly classified instances | Incorrectly classifies instances | CLASS : AMYLOIDOSIS | | | CLASS : OTHER DISEASE | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F-mesure | Precision | Recall | F-mesure |
| SMO | 92 | 21 | 0,814 | 0,976 | 0,888 | 0,818 | 0,321 | 0,462 |
| Bayes Naive Multinomial | 84 | 29 | 0,769 | 0,941 | 0,847 | 0,444 | 0,143 | 0,216 |
| Random Forest | 85 | 28 | 0,752 | 1 | 0,859 | ? | 0 | ? |
| Random Tree | 74 | 39 | 0,795 | 0,729 | 0,761 | 0,343 | 0,429 | 0,381 |
| Naive Bayes | 78 | 35 | 0,798 | 0,788 | 0,793 | 0,379 | 0,393 | 0,386 |
| Classification via Regression | 74 | 39 | 0,767 | 0,776 | 0,772 | 0,296 | 0,286 | 0,291 |

FIGURE 4.4 – Tableau d'algorithmes d'apprentissage automatique pour la classification binaire (Amyloïdose - Autre maladie)

**Classification binaire** Du fait des incertains résultats obtenus lors du premier type de classification, nous avons décidé de mener à bien une deuxième classification, cette fois-ci binaire. L'objectif est de prédire si un message parle d'amyloïdose ou d'autre maladie, n'importe laquelle, afin de vérifier si cette classification plus contrainte nous permet d'obtenir de meilleurs *scores*.

Dans une première expérience, nous avons ré-utilisé le même corpus que l'on avait construit précédemment pour l'entraînement de l'algorithme multiclasse. La seule modification que nous avons réalisé est de rassembler tous les catégories de maladies qui ne sont pas « amyloïdose » par la catégorie « OTHER_DISEASE ». De cette façon, on transforme un problème de classification multiclasse en un problème de classification binaire.

Après avoir entraîné plusieurs modèles d'apprentissage automatique, nous intégrons un tableau (4.4) qui rassemble les mesures obtenues pour chacun.

Nous intégrons également les matrices de confusion pour ces trois algorithmes les plus performants.

Bien que les pourcentages obtenus pour la prédiction de l'amyloïdose atteignent tous un 0.7% en précision, rappel et F-mesure, la prédiction de la

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 67 | 18 | AMYLOIDOSIS |
| | NEGATIVE | 17 | 11 | OTHER DISEASE |

FIGURE 4.5 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) - Naive Bayes

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 80 | 5 | AMYLOIDOSIS |
| | NEGATIVE | 24 | 4 | OTHER DISEASE |

FIGURE 4.6 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) - Naive Bayes Multinomial

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 83 | 2 | AMYLOIDOSIS |
| | NEGATIVE | 19 | 9 | OTHER DISEASE |

FIGURE 4.7 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) - SMO

classe « OTHER_DISEASE » reste entre le 0.3% et le 0.4%. Cette différence est probablement due à la disproportion des classes dans le corpus : 89 cas d'amyloïdose face aux 24 cas d'autres maladies.

Le modèle qui fournit les taux les plus élevés est celui entraîné avec l'algorithme Sequential Minimal Optimization (SMO), un algorithme qui permet de résoudre rapidement le problème quadratique de Support Vector Machines. SVM est une méthode de classification binaire par apprentissage supervisé, introduit par Vapnik en 1995. Cette méthode repose sur l'existence d'un classifieur linéaire. Du fait qu'il s'agit d'un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonctions dites noyau (*kernel*) qui permettent une séparation optimale des données (Platt 1998).

Dans un deuxième temps, nous avons décidé d'exploiter les étiquettes cliniques du corpus annoté afin de vérifier si un message peut être catégorisé par le biais des informations cliniques présentes dans un message. Pour mener à bien cette expérience, les éléments des vecteurs ne représenteront pas des mots du vocabulaire extrait du corpus mais des étiquettes, plus particulièrement les 36 étiquettes cliniques que nous avions utilisé lors de la phase d'annotation.

Cette expérience repose également sur un problème de classification binaire, mais se basant sur les occurrences des étiquettes, c'est-à-dire, on ne compte pas seulement la présence ou non de l'étiquette mais aussi le nombre de fois que celle-ci apparaît dans le message.

Nous pouvons visualiser ci-dessous le vecteur qui correspond au premier message :

*2, 0, 0, 0, 0, 0, 6, 0, 0, 0, 1, 6, 1, 0, 1, 0, 3, 0, 0, 0, 0, 5, 0, 0, 3, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, AMYLOIDOSIS*

*« A 48-year-old woman with no significant past medical history presented to the hospital with anasarca and was found to have nephrotic range proteinuria. Kidney biopsy demonstrated amyloidosis involving the*

*blood vessels, interstitium, and glomeruli. Bone-marrow biopsy revealed clonal plasma cells accounting for 23% of the marrow cellularity. Diagnoses of systemic primary (AL) amyloidosis and smoldering multiple myeloma were made. Echocardiography was remarkable for moderate concentric left ventricular hypertrophy with a thickened interventricular septum, dilated atria, a plethoric inferior vena cava, and a pulmonary artery pressure (PAP) of 67-70 mm Hg. After 2 years of treatment with thalidomide and dexamethasone, her PAP normalized and her interventricular septum diameter approached normal ; nonetheless, autologous peripheral blood stem-cell transplantation was complicated by atrial fibrillation, hypotension, and volume overload. She remains in hematologic remission without evidence of progressive organ dysfunction while taking maintenance lenalidomide »*

Nous pouvons observer que le vecteur recueille 2 occurrences pour l'étiquette « biopsy » dans la position 0, 6 occurrences de « disorder » dans la position 6, etc.

Les résultats ressemblent à ceux obtenus lors de la première expérience de classification binaire où le modèle a appris à partir des tokens des messages. La prédiction de la classe « amyloïdose » réussit des meilleurs taux de précision et rappel que celle concernant les autres maladies.

Il faut remarquer que lors de cette expérience, nous privilégions la présence des étiquettes cliniques, c'est-à-dire, le modèle apprend à partir de celles-ci. Par contre, il a aucune idée des mots ou expressions qui contient le message. Évidemment, étant donné que notre jeu d'étiquettes se focalise sur les symptômes de l'amyloïdose, cette classe est celle à obtenir de meilleurs résultats. Cependant, il faut aussi souligner le fait que dans cette façon de vectorisation, nous enlevons la présence du mot « amyloydosis » ou des expressions similaires, donc nous pouvons affirmer que le modèle est capable de « reconnaître » l'amyloïdose sans se fonder sur la présence ou non du mot.

Nous intégrons également un tableau (figure 4.8) qui renferme les résultats pour les différents algorithmes testés.

Classification binaire - Étiquettes cliniques

| Algorithm | Correctly classified instances | Incorrectly classifies instances | CLASS : AMYLOIDOSIS | | | CLASS : OTHER DISEASE | | |
|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F-mesure | Precision | Recall | F-mesure |
| SMO | 89 | 24 | 0,814 | 0,788 | 0,881 | ? | 0 | ? |
| Naive Bayes Multinomial | 83 | 30 | 0,847 | 0,809 | 0,828 | 0,393 | 0,458 | 0,423 |
| Random Forest | 89 | 24 | 0,822 | 0,933 | 0,874 | 0,5 | 0,250 | 0,333 |
| Random Tree | 78 | 35 | 0,821 | 0,775 | 0,798 | 0,310 | 0,375 | 0,340 |
| Naive Bayes | 64 | 49 | 0,870 | 0,528 | 0,657 | 0,288 | 0,708 | 0,410 |
| Classification via Regression | 90 | 23 | 0,8 | 0,989 | 0,884 | 0,667 | 0,083 | 0,148 |

FIGURE 4.8 – Tableau d'algorithmes d'apprentissage automatique pour la classification binaire (Amyloïdose - Autre maladie) à partir des étiquettes cliniques

Random Forest - Confusion Matrix - Étiquettes cliniques

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 83 | 6 | AMYLOIDOSIS |
| | NEGATIVE | 18 | 6 | OTHER DISEASE |

FIGURE 4.9 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) à partir des étiquettes cliniques - Random Forest

Les algorithmes qui fournissent les mesures d'évaluation les plus élevées sont ceux de Random Forest, Naive Bayes Multinomial et la classification via régression. Contrairement à l'expérience antérieure, le SMO est l'algorithme avec des scores les plus bas.

Nous intégrons aussi les matrices de confusion pour ces trois algorithmes les plus performants.

Par ailleurs, même si l'algorithme SMO ne réussit pas si des bons « scores », il nous permet d'examiner les poids attribués aux différentes étiquettes. Encore une fois, il n'est pas étonnant que les attributs avec les poids les plus élevés

Naive Bayes Multinomial - Confusion Matrix - Étiquettes cliniques

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 72 | 17 | AMYLOIDOSIS |
| | NEGATIVE | 13 | 11 | OTHER DISEASE |

FIGURE 4.10 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) à partir des étiquettes cliniques - Naive Bayes Multinomial

Classification via Regression - Confusion Matrix - Étiquettes cliniques

| | | ACTUAL VALUES | | |
|---|---|---|---|---|
| | | POSITIVE | NEGATIVE | |
| PREDICTED VALUES | POSITIVE | 88 | 1 | AMYLOIDOSIS |
| | NEGATIVE | 22 | 2 | OTHER DISEASE |

FIGURE 4.11 – Matrice de confusion pour la classification binaire (Amyloïdose - Autre maladie) à partir des étiquettes cliniques - Classification via régression

correspondent aux symptômes de l'amyloïdose, particulièrement l'amyloïdose cardiaque : *cardiac arrest*, *heart test*, *breathlesness*. D'autre part, l'arbre de décision permet d'observer que les principales étiquettes qui permettent la classification entre amyloïdose et non-amyloïdose sont celles de *heart test*, *breathlessness* et *medical practitioner*.

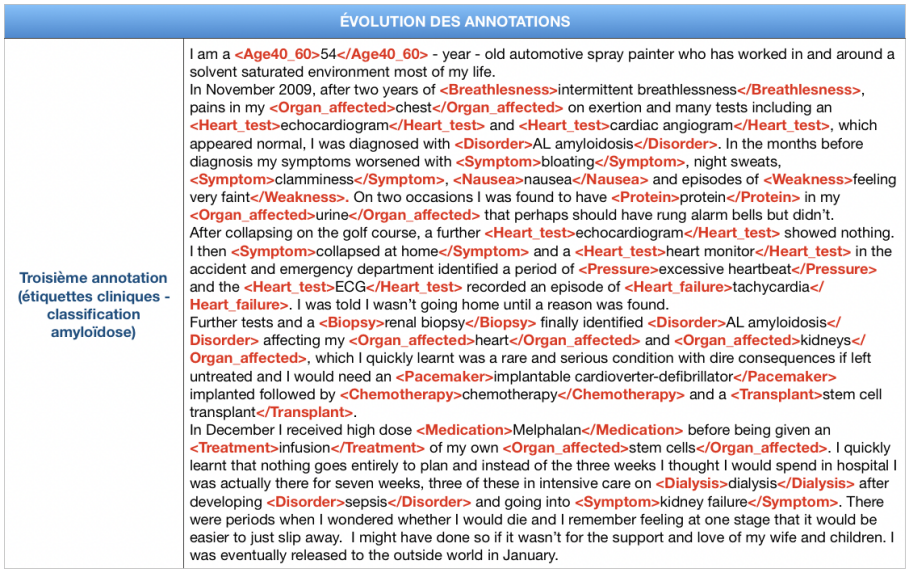
Finalement, nous intégrons un tableau qui recueille l'évolution des annotations d'un message du corpus lors de différentes tâches et expériences menées à bien :

| ÉVOLUTION DES ANNOTATIONS | |
|---|---|
| **Première annotation (étiquettes identifiantes et cliniques)** | I am a <Age40_60>54</Age40_60> - year - old automotive spray painter who has worked in and around a solvent saturated environment most of my life.<br><br>In <Date>November 2009</Date>, after two years of <Breathlesness>intermittent breathlessness</Breathlesness>, pains in my <Organ_affected>chest</Organ_affected> on exertion and many tests including an <Heart_test>echocardiogram</Heart_test> and <Heart_test>cardiac angiogram</Heart_test>, which appeared normal, I was diagnosed with <Disorder>AL amyloidosis</Disorder>. In the months before diagnosis my symptoms worsened with <Symptom>bloating</Symptom>, <Date>night sweats</Date>, <Symptom>clamminess</Symptom>, <Nausea>nausea</Nausea> and episodes of <Weakness>feeling very faint</Weakness>. On two occasions I was found to have <Protein>protein</Protein> in my <Organ_affected>urine</Organ_affected> that perhaps should have rung alarm bells but didn't.<br>After collapsing on the golf course, a further <Heart_test>echocardiogram</Heart_test> showed nothing. I then <Symptom>collapsed at home</Symptom> and a <Heart_test>heart monitor</Heart_test> in the accident and emergency department identified a period of <Pressure>excessive heartbeat</Pressure> and the <Heart_test>ECG</Heart_test> recorded an episode of <Heart_failure>tachycardia</Heart_failure>. I was told I wasn't going home until a reason was found.<br>Further tests and a <Biopsy>renal biopsy</Biopsy> finally identified <Disorder>AL amyloidosis</Disorder> affecting my <Organ_affected>heart</Organ_affected> and <Organ_affected>kidneys</Organ_affected>, which I quickly learnt was a rare and serious condition with dire consequences if left untreated and I would need an <Pacemaker>implantable cardioverter-defibrillator</Pacemaker> implanted followed by <Chemotherapy>chemotherapy</Chemotherapy> and a <Transplant>stem cell transplant</Transplant>.<br>In <Date>December</Date> I received high dose <Medication>Melphalan</Medication> before being given an <Treatment>infusion</Treatment> of my own <Organ_affected>stem cells</Organ_affected>. I quickly learnt that nothing goes entirely to plan and instead of the three weeks I thought I would spend in hospital I was actually there for seven weeks, three of these in intensive care on <Dialysis>dialysis</Dialysis> after developing <Disorder>sepsis</Disorder> and going into <Symptom>kidney failure</Symptom>. There were periods when I wondered whether I would die and I remember feeling at one stage that it would be easier to just slip away. I might have done so if it wasn't for the support and love of my wife and children. I was eventually released to the outside world in <Date>January</Date>. |

FIGURE 4.12 – Annotation Phase 1 - Étiquettes cliniques et identifiantes

| ÉVOLUTION DES ANNOTATIONS | |
|---|---|
| **Deuxième annotation (étiquettes identifications - tâche de désidentification)** | I am a <Age>54</Age> - year - old automotive spray painter who has worked in and around a solvent saturated environment most of my life.<br><br>In **<Date>**November 2009</Date>, after two years of intermittent breathlessness, pains in my chest on exertion and many tests including an echocardiogram and cardiac angiogram, which appeared normal, I was diagnosed with AL amyloidosis. In the months before diagnosis my symptoms worsened with bloating, night sweats, clamminess, nausea and episodes of feeling very faint. On two occasions I was found to have protein in my urine that perhaps should have rung alarm bells but didn't.<br>[...]<br>My wife found The National Amyloidosis Centre in <Location>London</Location> on the internet. We emailed them with the proposed treatment and received an immediate answer assuring us that if my diagnosis was definitely AL amyloidosis, the suggested treatment was the best option for me, emphasising that time was of the essence.<br><br>In <Date>November 2009</Date> I received my first lot of chemotherapy followed by stem cell stimulating injections.<br>[...] |

FIGURE 4.13 – Annotation Phase 2 - Désidentification - Étiquettes identifiantes



| ÉVOLUTION DES ANNOTATIONS | |
|---|---|
| Troisième annotation (étiquettes cliniques - classification amyloïdose) | I am a <Age40_60>54</Age40_60> - year - old automotive spray painter who has worked in and around a solvent saturated environment most of my life.<br>In November 2009, after two years of <Breathlesness>intermittent breathlessness</Breathlesness>, pains in my <Organ_affected>chest</Organ_affected> on exertion and many tests including an <Heart_test>echocardiogram</Heart_test> and <Heart_test>cardiac angiogram</Heart_test>, which appeared normal, I was diagnosed with <Disorder>AL amyloidosis</Disorder>. In the months before diagnosis my symptoms worsened with <Symptom>bloating</Symptom>, night sweats, <Symptom>clamminess</Symptom>, <Nausea>nausea</Nausea> and episodes of <Weakness>feeling very faint</Weakness>. On two occasions I was found to have <Protein>protein</Protein> in my <Organ_affected>urine</Organ_affected> that perhaps should have rung alarm bells but didn't.<br>After collapsing on the golf course, a further <Heart_test>echocardiogram</Heart_test> showed nothing. I then <Symptom>collapsed at home</Symptom> and a <Heart_test>heart monitor</Heart_test> in the accident and emergency department identified a period of <Pressure>excessive heartbeat</Pressure> and the <Heart_test>ECG</Heart_test> recorded an episode of <Heart_failure>tachycardia</Heart_failure>. I was told I wasn't going home until a reason was found.<br>Further tests and a <Biopsy>renal biopsy</Biopsy> finally identified <Disorder>AL amyloidosis</Disorder> affecting my <Organ_affected>heart</Organ_affected> and <Organ_affected>kidneys</Organ_affected>, which I quickly learnt was a rare and serious condition with dire consequences if left untreated and I would need an <Pacemaker>implantable cardioverter-defibrillator</Pacemaker> implanted followed by <Chemotherapy>chemotherapy</Chemotherapy> and a <Transplant>stem cell transplant</Transplant>.<br>In December I received high dose <Medication>Melphalan</Medication> before being given an <Treatment>infusion</Treatment> of my own <Organ_affected>stem cells</Organ_affected>. I quickly learnt that nothing goes entirely to plan and instead of the three weeks I thought I would spend in hospital I was actually there for seven weeks, three of these in intensive care on <Dialysis>dialysis</Dialysis> after developing <Disorder>sepsis</Disorder> and going into <Symptom>kidney failure</Symptom>. There were periods when I wondered whether I would die and I remember feeling at one stage that it would be easier to just slip away. I might have done so if it wasn't for the support and love of my wife and children. I was eventually released to the outside world in January. |

FIGURE 4.14 – Annotation Phase 3 - Classification amyloïdose - Étiquettes cliniques

## 4.2 Discussions

### 4.2.1 Désidentification

De toute évidence, nous avons implémenté les deux méthodes principales pour la désidentification dont parlent Meystre *et al.* (MEYSTRE et al. 2010) dans leur bilan : le *pattern matching* et le *machine learning*. Le fait d'avoir décidé de développer ces deux approches nous a permis de mener à bien l'application de deux méthodologies différentes ainsi que de réfléchir à leurs particularités, avantages et inconvénients.

En conséquence, après le développement de traitements concernant les deux méthodes, nous convenons que tous les deux ont besoin dans une certaine mesure de leur adaptation au corpus de travail en question.

En premier lieu, l'approche dite orientée connaissances se caractérise par être très spécifique aux corpus à partir desquels les traitements ont été conçus. Cela veut dire, les règles grammaticales et syntaxiques créés par exemple pour détecter les initiales, les dictionnaires de noms et prénoms, la liste de noms de maladies ou les expressions régulières pour capturer les dates sont particuliers

à la langue. En plus, le fait de filtrer les noms de maladies restreint nos traitements au domaine de la santé. À cause de cela, tout le travail mis en place pour la désidentification du corpus espagnol n'est pas réutilisable pour l'autre corpus de messages en anglais, ni pour un autre corpus portant sur un sujet différent, étant donné qu'il a été conçu envisageant le domaine de la santé.

Deuxièmement, l'approche orientée données semble avoir le même type d'inconvénients en termes de réutilisation des méthodes. Dans ce cas, le travail qui doit être réalisé avant la création du modèle de reconnaissance d'entités est également coûteux. Si dans la première méthodologie nous devions rassembler des ressources comme des dictionnaires ou des listes et construire des règles et des patrons, cette fois-ci, cette approche nécessite d'un large corpus annoté d'où l'algorithme puisse apprendre. Par ailleurs, une fois le modèle créé, ceci ne fonctionne qu'avec des corpus vraisemblables et, en plus, dans la même langue que les données fournies en entrée à l'algorithme.

Alors, les deux méthodes développées ne sont pas interchangeables à cause essentiellement de la langue, mais aussi des aspects culturels. Le modèle CRF ne fonctionnerait pas sur des textes cliniques en espagnol, il faudrait annoter un autre corpus et entraîner un modèle à nouveau. De la même manière, le système reposant sur des ressources linguistiques ne peut pas être implémenté sur des messages écrits en anglais, puisque ni les noms ni les dates seront reconnus par les dictionnaires et les règles construites précédemment. Par exemple, les patrons construits pour les numéros d'identité et les numéros de téléphone sont certainement liés non seulement à la langue mais aussi à la culture. Il faudrait une fois de plus construire des ressources linguistiques et des patrons particuliers à l'anglais.

Ainsi, il n'est pas frappant que toute la littérature que nous avons consultée met l'accent sur l'importance du caractère généralisable des solutions conçues pour la désidentification. Les patrons, règles et dictionnaires sont souvent manuellement confectionnés, avec un coût élevé en termes de temps et des experts du domaine, et sont peu généralisables. En revanche, ils sont rapide et facilement modifiables lorsqu'il faut augmenter la performance, et nécessitent peu

ou aucunes données d'entraînement annotées. Cependant, les avantages des méthodes de *machine learning* entraînent l'apprentissage automatique et la reconnaissance de patrons, et ils peuvent être progressivement adaptés à des nouveaux types de documents ou des domaines sans besoin d'augmenter la complexité. Le principal inconvénient est que, contrairement aux méthodes précédentes, elles nécessitent des larges corpus annotés.

Finalement, dans le but de comparer ces deux approches, on envisage comme futur travail de développer les deux méthodologies mais dans les langues contraires. Cela veut dire, annoter le corpus de textes cliniques en espagnol afin d'évaluer et comparer la performance de deux méthodes, et construire un système de désidentification à partir des dictionnaires et règles étant le but aussi d'évaluer quelle approche réussit mieux.

En conclusion, il faut souligner que le fait d'avoir conçu et implémenté les deux méthodologies nous a permis de connaître plus en détail le développement de ces approches, en plus que d'avoir une idée plus spécifique des avantages et inconvénients de chacune. C'est ainsi que nous concluons que la meilleure option serait de concevoir des solutions qui reposent sur les deux méthodologies, ce qu'on connaît comme des solutions hybrides, et dont nous avons parlé lors de la description des principales méthodes implémentées pour la désidentification de données.

### 4.2.2 Prédiction

Après la réalisation de toutes ces expériences décrites tout au long du projet, nous nous sommes posé la suivante question : quelles seraient les améliorations envisageables pour poursuivre ce travail et le rendre utilisable ?

Comme on pouvait s'y attendre, la classe « amyloïdose », qui a bénéficié de davantage de travail que les autres, en termes de nombre d'étiquettes accordées pour l'annotation et de nombre d'observations dans le corpus, obtient d'excellents résultats. Toutefois, la performance des autres classes n'atteint pas telle réussite. On espère que, dans une version future, le fait d'accorder autant d'attention aux autres classes permettrait d'améliorer largement les résultats.

Dans notre travail, nous avons décidé au début de constituer le corpus pour la classification de l'amyloïdose cardiaque avec des observations soit des patients atteints de cette maladie soit d'une autre maladie cardiaque rare. La raison pour laquelle nous avons déterminé de composer ainsi le corpus est dû à la nature de l'amyloïdose cardiaque comme maladie cardiaque rare. Cependant, il faut reconnaître qu'après la réalisation de ce travail, nous nous sommes aperçu que la principale difficulté du dépistage de l'amyloïdose cardiaque réside dans le fait que ses symptômes correspondent en même temps à d'autres maladies ou insuffisances cardiaques. C'est ainsi que très souvent le diagnostic reste au niveau de la maladie cardiaque sans aller jusqu'à la détection de l'amyloïdose cardiaques.

Par conséquence, un futur travail pour améliorer le modèle de prédiction de l'amyloïdose cardiaque consisterait à créer un nouveau corpus composé par des patients d'amyloïdose cardiaque et d'autres atteints seulement d'une maladie ou insuffisance cardiaque de sorte que le modèle puisse apprendre à les distinguer.

D'autre part, en ce qui concerne la distribution d'observations de notre corpus de messages, il est indéniable que notre corpus a des dimensions réduites et, qu'en plus, les instances de la classe amyloïdose sont majoritaires, donc on risque que le modèle ait sur-appris. Une grande amélioration sans doute serait d'agrandir le corpus. D'une part, pour incrémenter le nombre total de messages, d'autre part, pour ajuster les proportions de classes.

En outre, l'annotation automatique des entités cliniques pourrait être aussi utile pour générer automatiquement des tableaux cliniques de patients à partir de, par exemple, soit des messages de patients soit des notes de médecins ou d'infirmières. Ces annotations serviraient à extraire les informations les plus importantes de ces textes, comme les symptômes, les maladies, les résultats de tests, des médicaments prescrits, etc. Ce travail pourrait être utile pour résumer de façon schématique l'information la plus pertinente d'un patient, débarrassant en même temps de ce travail au personnel de santé.

Finalement, nous considérons que le travail mis en place peut représenter

un point de départ pour des futures expériences qui pourraient s'avérer vraiment utiles en rapport avec le dépistage précoce de l'amyloïdose cardiaque. La tâche d'annotation des informations cliniques pourrait être aussi employée pour entraîner un algorithme de CRF pour la reconnaissance d'entités comme des symptômes, organes atteints, âge, et d'autres étiquettes que nous avons envisagées et qui peuvent être pertinentes pour la détection de maladies. Il faudrait aussi réfléchir de nouveau à notre jeu d'étiquettes. Si jamais on décide de modifier le cible de la classification afin de créer un modèle qui prédite l'amyloïdose parmi des patients atteints d'insuffisance cardiaque peut-être qu'il faudrait ajuster notre jeu d'étiquettes pour qu'il vise de façon plus contrainte cette problématique. Par exemple, il faudrait ajouter des étiquettes plus particulières pour les différents tests cardiaques ou les distincts problèmes qui peut souffrir le cœur. Grâce a ces annotations, il serait possible ensuite de classifier les patients pour essayer de détecter des patients qui risquent d'avoir amyloïdose cardiaque. De cette façon, notre modèle fonctionnerait comme un système de prévention pour les médecins. Lorsqu'il trouve un possible cas d'amyloïdose il préviendrait les médecins afin que ceux réalisent les tests nécessaires pour vérifier ce diagnostic. Ce système représenterait peut-être un support essentiel pour l'équipe de médecins.

# Section 5

# Conclusion

Ce travail naît du projet convenu entre l'entreprise pharmaceutique Pfizer, la Fundación San Juan de Dios et Sopra Steria, avec l'objectif d'améliorer la détection précoce de l'amyloïdose cardiaque. Le fait de pouvoir détecter cette maladie de façon anticipé entraînerait des grandes bénéfices, aussi bien en ce qui concerne le dépistage précoce de l'amyloïdose cardiaque et le traitement des malades atteints, qu'en termes de dépenses de services sanitaires.

D'une part, la détection précoce de maladies permettrait de traiter les symptômes en avance en fin de réduire l'impact de celles-ci. De plus, la détection des symptômes de l'amyloïdose cardiaque entraînerait une réduction du nombre de visites des patients qui se rendent à l'hôpital fréquemment puisqu'ils n'arrivent pas à être correctement diagnostiqués. De fait, le temps entre les premiers symptômes et le diagnostic de l'amyloïdose est d'environ un an et les patients voient en moyenne cinq médecins avant que le diagnostic soit fait (Georgin-Lavialle et al. 2017). Il est donc essentiel que le diagnostic soit réalisé le plus rapidement possible de façon que l'aggravation progressive des organes atteints puisse être évitée, ou au moins ralentie.

D'autre part, le fait de bien diagnostiquer un patient dès sa première consultation entraînerait des épargnes, autant économiques comme en termes de services cliniques.

Ce projet a permis de prendre conscience de l'importance de l'intelligence artificielle dans le domaine de la santé. L'application de ce domaine au secteur

de la santé a abouti à des grandes avances comme le développement de traitements, l'amélioration de la prise en charge de malades, l'aide aux patients, et plus particulièrement, la détection précoce de maladies. Cette dernière application est le principal objectif de la recherche menée lors de ce projet. Le dépistage précoce comporte plusieurs bénéfices pour le patient, comme refréner les symptômes, pallier l'impact de la maladie et même améliorer les conditions de vie des patients atteints.

Bien que nous n'avons pas eu le temps d'exploiter les vraies données du projet et exposer les résultats, le travail réalisé représente un préambule des tâches que l'on est en train de développer dans le cadre du projet réel à Sopra Steria. Nous avons mené à terme les deux principales tâches dans lesquelles repose ce projet :

- Pseudonymisation des données. Nous voudrions remarquer que le processus mené à terme est bien une pseudonymisation. Il s'agit pas d'anonymisation puisqu'étant donné que l'objectif final est de réaliser des tests de dépistage aux patients diagnostiqués par le modèle de prédiction avec amyloïdose cardiaque pour vérifier le diagnostic, il est indispensable de pouvoir arriver jusqu'au patient à travers les données identifiantes. La tâche consiste donc à masquer les données identifiantes des patients pour traiter les données et entraîner le modèle, en plus de pouvoir permettre dans un futur aux médecins de réaliser le diagnostic des patients déclarés comme atteints d'amyloïdose cardiaque par le modèle.
- Analyse et observations des symptômes et maladies associées à l'amyloïdose cardiaque. Le but initial du projet était de réussir à perfectionner le dépistage de l'amyloïdose cardiaque et de distinguer les cas où il s'agit d'autres symptômes ou maladies des cas où ces douleurs et maladies correspondent aux symptômes de l'amyloïdose cardiaque.

Malgré les différentes expériences que l'on a réalisé avec le jeu de données construit, il faut reconnaître que les résultats ne sont pas assez satisfaisants. La raison de cet échec est probablement la dimension du corpus. Il serait nécessaire de récupérer plus des observations, particulièrement des messages concer-

nant d'autres maladies, afin de compenser les deux classes. Si l'on réussissait à rassembler plus des messages de patients décrivant leurs symptômes, nous pourrions incrémenter les taux obtenus pour la reconnaissance de l'amyloïdose, même si celles-ci ne sont pas si décevantes.

D'autre part, bien que nous n'avons pas eu le temps de montrer les expériences menées avec les vraies données du projet, ce travail a représenté un point de départ pour la réalisation de ceci. Ce travail nous a permis d'acquérir un sens plus complet et détaillé de plusieurs questions largement pertinentes pour la réalisation du projet réel : les différents types d'amyloïdoses, en quoi consistent, quels sont les possibles symptômes reliés aux amyloses, les traitements qui existent actuellement, etc. Tous ces connaissances étaient sans aucune doute essentielles pour pouvoir mener à bien le projet de prédiction de l'amyloïdose cardiaque. Si nous aspirons à pouvoir diagnostiquer une maladie de façon précoce à travers des algorithmes d'apprentissage automatique, il faut forcément comprendre les données dont nous disposons, en plus d'avoir une idée vraiment détaillée de la maladie et tout ce qu'elle implique.

Les amyloïdoses, étant difficiles à diagnostiquer par les médecins, nous observons dans nos expériences que les prédictions sont aussi complexes à réaliser pour une machine, en particulier pour un apprentissage statistique dans la mesure où il faudrait savoir quels sont les éléments essentiels du diagnostic pour pouvoir fournir à la machine ces éléments. Actuellement, il reste difficile de prédire les amyloïdoses, mais les expériences menées peuvent servir comme aide aux médecins pour attirer leur attention sur des cas probables d'amyloïdose. Autrement dit, la machine met en évidence des possibilités, et les médecins vérifient et valident ou invalident ces possibilités.

# Bibliographie


[Min58] Marvin L MINSKY. « Some methods of artificial intelligence and heuristic programming ». In : *Proc. Symposium on the Mechanization of Thought Processes, Teddington*. 1958.

[Nil80] Nils J. NILSSON. *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., 1980.

[Rou95] Maurice ROULEAU. « La langue médicale : une langue de spécialité á emprunter le temps d'une traduction ». In : *Traduction, terminologie, rédaction (TTR)* 8 (1995), p. 29-49.

[Pla98] John PLATT. *Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines*. Rapp. tech. Microsoft Research, 1998.

[Est99] Jefatura del ESTADO. *Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal*. Rapp. tech. Agencia Estatal Boletín Oficial del Estado, 1999. URL : https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750.

[MWD00] Fiszman M., Chapman WW. et Aronsky D. « Automatic detection of acute bacterial pneumonia from chest X-ray reports ». In : *J Am Med Inform Assoc* (2000).

[Kha+01] Javed KHAN et al. « Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks ». In : *Nat Med* (2001).

[Cun+02] Hamish CUNNINGHAM et al. « GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications ». In : juil. 2002.

[TBK02] Ricky TAIRA, Alex BUI et Hooshang KANGARLOO. « Identification of patient name references within medical documents using semantic selectional restrictions ». In : *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* (fév. 2002), p. 757-61.

[Ber03] JJ. BERMAN. « Concept-match medical data scrubbing. How pathology text can be used in research ». In : *Arch Pathol Lab Med* (2003), p. 680-6.

[Ara+06] Eiji ARAMAKI et al. « Automatic deidentification by using sentence features and label consistency ». In : *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data* (jan. 2006).

[Bec+06] BA. BECKWITH et al. « Development and evaluation of an open source software tool for deidentification of pathology reports ». In : *BMC Med Inform Decis Mak* (2006), p. 12.

[Guo+06] Yikun GUO et al. « Identifying Personal Health Information Using Support Vector Machines ». In : *i2b2 Workshop on Chanllenges in Natural Language Processing for Clinical Data* (2006).

[SFK06] György SZARVAS, Richárd FARKAS et András KOCSOR. « A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms ». In : oct. 2006, p. 267-278. DOI : 10.1007/11893318_27.

[Mur+07] Shawn N. MURPHY et al. « Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside ». In : *AMIA 2007 Symposium Proceedings* (2007).

[SFB07] György Szarvas, Richárd Farkas et Róbert Busa-Fekete. « Research Paper : State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. » In : *JAMIA* 14 (juin 2007), p. 574-580. doi : 10.1197/jamia.M2441.

[GX08] James Gardner et Li Xiong. « HIDE : An Integrated System for Health Information DE-identification ». In : juil. 2008, p. 254-259. isbn : 978-0-7695-3165-6. doi : 10.1109/CBMS.2008.129.

[Nea+08] Ishna Neamatullah et al. « Automated de-identification of free-text medical records ». In : *BMC Medical Informatics and Decision Making* (juil. 2008). url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2526997/.

[Ada+09] David Adams et al. *L'amylose héréditaire*. Rapp. tech. Association Française contre l'amylose, 2009. url : http://www.amylose.asso.fr/amylose_hereditaire.php.

[BGA10] E. Bombardieri, F. Gianmmarile et C. Aktolun. « 131I/123I-metaiodobenzylguanidine (MIBG), scintigraphy : procedure guidelines for tumour imaging ». In : *European Journal of Nuclear Medicine and Molecular Imaging* (2010).

[Mey+10] Stephane M. Meystre et al. « Automatic de-identification of textual documents in the electornic health record : a review of recent research ». In : *BMC Medical Research Methodology* (août 2010). url : https://www.ncbi.nlm.nih.gov/pubmed/20678228.

[ASS11] Ujma Ansari, Dipesh Sharma et Sunita Soni. « Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction ». In : *International Journal of Computer Applications* 17 (mar. 2011), p. 43-48. doi : 10.5120/2237-2860.

[Elu+11] Sofía Garrido Elustondo et al. « Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction ». In : *Elsevier* 44 (juil. 2011).

[DA12] Chaitrali DANGARE et Sulbha APTE. « Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques ». In : *International Journal of Computer Applications* 47 (juin 2012), p. 44-48. DOI : 10.5120/7228-0076.

[KL12] Baker R. KELTY et Rice LAWRENCE. « The Amyloidoses : Clinical Features, Diagnosis and Treatment ». In : *Methodist Debakay Cardiovasc Journal* (2012).

[Gro13] Cyril GROUIN. « Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique ». Thèse de doct. Université Pierre et Marie Curie Paris VI, 2013.

[Nou+13] Damien NOUVEL et al. « Fouille de régles d'annotation pour la reconnaissance d'entités nommées ». In : *TAL* 54 (jan. 2013), p. 13-41.

[Dam14] Thibaud DAMY. « Actualité sur l'amylose cardiaque, pourquoi faut-il y penser en 2014 ? » In : *E-Cordiam : Recommendations Cœur, Diabète, Métabolisme* (déc. 2014).

[DST14] DHEEBA, Albert SINGH et S. TAMIL. « Computer-aided detection of breast Cancer on mammograms : a swarm inntelligence optimized wavelet neural network approach ». In : *J Biomed Inform* (2014).

[Par14] Data Protection Working PARTY. *Article 29 Opinion 05 2014 on Anonymisation Techniques*. Rapp. tech. Data Protection Working Party, mai 2014. URL : https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf.

[Cro15] William CROWN. « Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions ». In : *Value in Health* 18 (jan. 2015). DOI : 10.1016/j.jval.2014.12.005.

[Hat15] Mohamed HATMI. « Reconnaissance des entités nommées dans des documents multimodaux ». Thèse de doct. Université de Nantes, 2015.

[HAB15] TJ. HIRSCHAUER, H. ADELI et JA. BUFORD. « Computer-Aided diagnosis of Parkinson's Disease Using Enhanced Probabilistic Neural Network ». In : *J Med Syst* (2015).

[Kay+15] Mehmet KAYAALP et al. « Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation ». In : *AMIA Annual Symposium Proceedings* (2015).

[lho15] Agence technique de l'information sur L'HOSPITALISATION. *Classification statistique International des maladies et des problèmes de santé. Bulletin officiel Nº 2015/9bis*. Rapp. tech. Ministère des affaires sociales et de la santé, 2015.

[Mir+15] MIRTSKHULAVA et al. « Artificial Neural Network Model in Stroke diagnosis, modeling and simulation ». In : *IEEE* (2015), p. 29-49.

[Con16] Parlément européen et le CONSEIL DE L'UNION EUROPÉENNE. *Règlement (UE) 2016/679 du Parlament Européen et du Conseil du 27 avril 2016 relatif á la protection des personnes physiques á l'égard du traitement des données á caractére personnel et á la libre circulation de ces données, et abrogeant la directive 95/46/CE (réglement général sur la protection des données)*. Rapp. tech. Parlement Européen et le Conseil de l'Union Européenne, 2016.

[Gar+16] Ravi GARG et al. « A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records ». In : (sept. 2016).

[TI16] Rodolphe THIEBAUT et INSERM. « Bid data en santé : Des défis techniques, humains et éthiques à relever ». In : *Inserm* (juil. 2016).

[ASA17] N. Afzal, S. Sohn et S. Abram. « Mining peripheral arterial disease cases from narrative clinical notes using natural language processing ». In : *J Vasc Surg* (2017).

[CDF17] VM. Castro, D. Dligach et S. Finan. « Large-scale identification of patients with cerebral aneurysms using natural language processing ». In : *Neurology* (2017).

[DLS17] Franck Dernoncourt, Ji Young Lee et Peter Szolovits. « NeuroNER : an easy-to-use program for named-entity recognition based on neural networks ». In : (mai 2017). URL : https://arxiv.org/abs/1705.05487.

[Geo+17] Sophie Georgin-Lavialle et al. *Mieux connaître Les Amyloses*. Rapp. tech. Association Française contre l'amylose, 2017.

[Jia+17] Fei Jiang et al. « Artificial intelligence in healthcare : past, present and future ». In : *Stroke and Vascular Neurology* (2017).

[WS17] Jenna Wiens et Erica Shenoy. « Machine Learning for Healthcare : On the Verge of a Major Shift in Healthcare Epidemiology ». In : *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 66 (août 2017). DOI : 10.1093/cid/cix731.

[Cha+18] Jean Charlet et al. « Intelligence artificielle et santé ». In : *Inserm* (juil. 2018).

[dIM18] Société française d'Hématologie, Norbert Ifrah et Marc Maynadié. « Hématologie ». In : Elsevier Masson, juin 2018. Chap. 10.

[MG18] Techical Committe IT Service Management et IT Governance. *ISO/IEC 20000-1 :2011 Information technology - Service Management*. Rapp. tech. International Standard Organization, 2018.

[Paí18] El País. « Los datos médicos, un tesoro que ya comienza a explotarse ». In : *El País* (juil. 2018).

[Rui+18] Enrique RUIZ-MORI et al. « Cardiac Amyloidosis : a case report ». In : *SciELO* 18.4 (déc. 2018).

[nom19] Commission de contrôle des informations NOMINATIVES. *Anonymisation ou pseudoanonymisation*. Rapp. tech. Commission de contrôle des informations nominatives, 2019. URL : https://www.ccin.mc/fr/fiches-pratiques/anonymisation-ou-pseudonymisation.